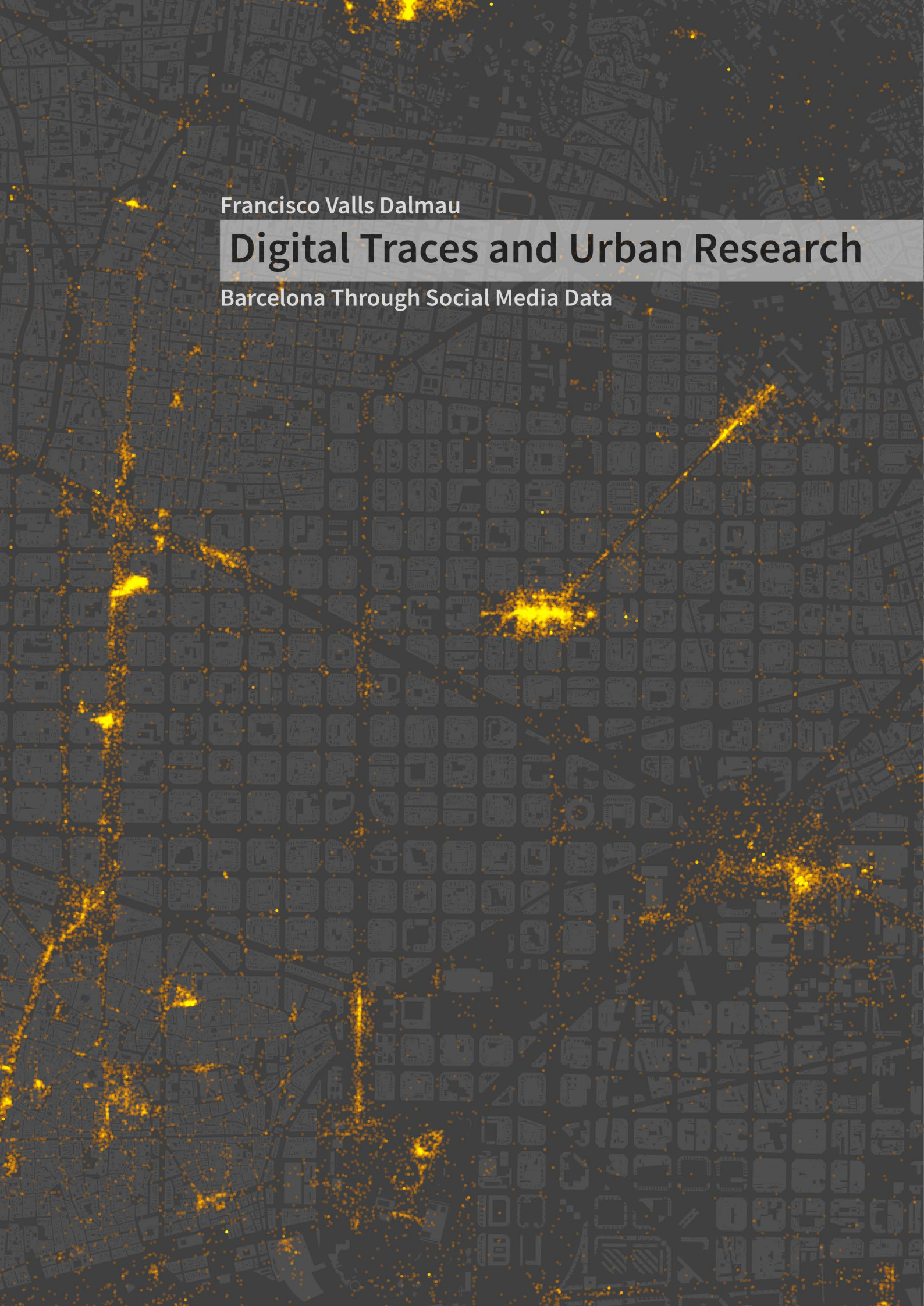


Francisco Valls Dalmau

Digital Traces and Urban Research

Barcelona Through Social Media Data

Digital Traces and Urban Research:
Barcelona Through Social Media Data

Author

Francisco Valls Dalmau

Director

Josep Roca Cladera

Universitat Politècnica de Catalunya

Department of Architectural Technology

Doctorate in Urban and Architectural Management and Valuation

March 2019

To my wife Cristina,
for her patience and support.
You are the best!

Acknowledgments

My special thank belongs to Dr. Ernest Redondo Domínguez for his support and insightful remarks on my work. His encouragement was extremely important for me to get my work up to this point. I would also like to thank Dr. Pilar Garcia Almirall for introducing me to GIS technology and encouraging me to pursue this avenue of research. I am also grateful for the generosity of the referees for their time reading this dissertation and their comments on the aspects that should be addressed.

This research was supported by the Non-Oriented Fundamental Research Project EDU2012-37247/EDUC of the VI National Plan for Scientific Research, Development and Technological Innovation 2008-2011, Government of Spain, titled "E-learning 3.0 in the teaching of architecture. Case studies of educational research for the foreseeable future".

Abstract

Most of the world's population now resides in urban areas, and it is expected that almost all of the planet's growth will be concentrated in them for the next 30 years, making the improvement of the quality of life in the cities one of the big challenges of this century. To that end, it is crucial to have information on how people use the spaces in the city, and allows urban planning to successfully respond to their needs.

This dissertation proposes using data shared voluntarily by the millions of users that make up social network's communities as a valuable tool for the study of the complexity of the city, because of its capacity of providing an unprecedented volume of urban information, with geographic, temporal, semantic and multimedia components.

However, the volume and variety of data raises important challenges regarding its retrieval, manipulation, analysis and representation, requiring the adoption of the best practices in data science, using a multi-faceted approach in the field of urban studies with a strong emphasis in the reproducibility of the developed methodologies.

This research focuses in the case of study of the city of Barcelona, using the public data collected from Panoramio, Flickr, Twitter and Instagram. After a literature review, the methods to access the different services are discussed, along with their available data and limitations. Next, the retrieved data is analyzed at different spatial and temporal scales.

The first approximation to data focuses on the origins of users who took geotagged pictures of Barcelona, geocoding the hometowns that appear in their Flickr public profiles, allowing the identification of the regions, countries and cities with the largest influx of visitors, and relating the results with multiple indicators at a global scale.

The next scale of analysis discusses the city as a whole, developing methodologies for the representation of the spatial distribution of the collected locations, avoiding the artifacts produced by overplotting. To this end, locations are aggregated in regular tessellations, whose size is determined empirically from their spatial distribution. Two spatial statistics techniques (Moran's I and Getis-Ord's G*) are used to visualize the local spatial autocorrelation of the areas with exceptionally high or low densities, under a statistical significance framework. Finally, the kernel density estimation is introduced as a non-parametric alternative.

The third level of detail follows the official administrative division of Barcelona in 73 neighborhoods and 12 districts, which obeys to historical, morphological and functional criteria. Micromaps are introduced as a representation technique capable of providing a geographical context to commonly used statistical graphics,

along with a methodology to produce these micromaps automatically. This technique is compared to annotated scatterplots to relate picture intensity with different urban indicators at a neighborhood scale.

The hypothesis of spatial homogeneity is abandoned at the most detailed scale, focusing the analysis on the street network. Two techniques to assign events to road segments in the street graph are presented (direct by shortest distance or by proxy through the postal addresses), as well as the generalization of the kernel density estimation from the Euclidean space to a network topology.

Beyond the spatial domain, the interactions of three temporal cycles are further analyzed using the timestamps available in the picture metadata: daytime/nighttime (daily cycle), work/leisure (weekly cycle) and seasonal (yearly cycle).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Alice: "Would you tell me, please, which way I ought to go from here?"
> The Cheshire Cat: "That depends a good deal on where you want to get to."
>
> Alice in Wonderland

1.1 Motivations

1.1.1 Background

Since mid-20th century, the number of people living in urban areas has steadily increased in absolute and relative terms [1]. With the majority of the world's population living in urban areas, and projections indicating that over the next 30 years most of the population growth will occur in cities, the necessity to improve the quality urban life and its sustainability becomes increasingly important, especially considering that with only 2% of the earth's surface, cities consume 78% of the world's energy[1].

Cities are among the most complex artifacts produced by humanity [2], and are constantly evolving and adapting, driven by the interactions between its inhabitants, acting with different levels of coordination, sometimes intentionally but oftentimes unknowingly.

[1] According to data from UN Habitat, available at *http://unhabitat.org/urban-themes/climate-change/* at the time of writing.

Urban planning and urban design shape the complexity of the city, through building codes, housing policies or urban regulations; planners and designers establish the guidelines of what they believe the city should develop towards, but what the city will ultimately become depends on a number of factors: social, economic and cultural.

To plan a better future, urban planners and designers must interpret the past of the city and understand the nature of the mechanisms that drive its processes. Therefore, a deeper understanding of the city is crucial, especially considering the contrasting slow pace of urban transformation compared to the needs of a quickly changing society.

At the same time, to match the growing complexity of urban phenomena, the approaches to understand the city have had to update accordingly. This dissertation focuses on a data-driven approach [3], using Barcelona as its case of study in a multi-faceted analysis of data gathered from social networks, as an emerging avenue of research to investigate this complexity.

1.1.2 Research Development Process

The interest in understanding the urban environment from the perspective of its users began with the participation in the PATRAC[2] research project, in particular with the evaluation of the physical accessibility of the street network of the historic center of Tossa de Mar [4, 5], and later with the study of the perception of the urban environment from a pedestrian's perspective using cadastral data [6].

As the research focus gravitated towards a deeper understanding of pedestrian movement patterns, the research focused on using "desire lines" —trails that can appear spontaneously outside designated pathways—, to reveal the preferences of pedestrians, with the objective of producing a agent-based model capable of explaining and predicting this behavior.

Unfortunately, this approach was not ultimately fruitful because of the difficulty of finding a sufficient number of examples to fit the model against and, more importantly, because of the limited cases in which it was applicable, as it was not appropriate on paved surfaces, or low-footfall areas. Therefore, despite some advances in modeling pedestrian behavior [7], the data gathering strategy shifted to extracting tracking data in randomized controlled experiments, following the behavior of test subjects in virtual environments [8].

During the development of this experimental design, some issues began to arise, in part because the anticipated difficulty in recruiting a significant number of test subjects, but also regarding the design of a suitable control mechanism of

[2] "Patrimonio Accesible: I+D+i para una Cultura sin Barreras".

the subjects' avatars capable of successfully mimicking the real world experience. These issues were corroborated by the literature review, where few references were found on this approach.

In this context, in summer 2016 a side project unexpectedly provided the data that until that moment had proven elusive, when the author was able to download data from the now defunct photo sharing service Panoramio, unbeknownst that the procedure was relatively easy because the service API was very simple to use, with few restrictions and with a very straightforward data structure. During the following months, the retrieval process was refined, and data from other services with more complex APIs and data structures were also successfully retrieved.

This approach provided a very rich dataset, consisting on millions of locations, with associated temporal (timestamps), semantic (descriptions or tweets) and multimedia (links to image or video content) data, as well as other metadata (e.g. user profiles, devices). These retrieved data allowed answering much more complex research questions, conducting more detailed analyses at larger scales, taking advantage of a natural experiment with thousands of anonymous participants.

1.1.3 Case of Study

The selected case of study was the city of Barcelona (Catalonia), which according to Mastercard Global Destination Cities Index 2017 Report was the 12th most visited city around the world [9], while Euromonitor [10] ranks Barcelona as the 25th in its city destinations ranking[3] and 6th in the Airbnb top destination cities[4]. Furthermore, it was the third most photographed city according to Sightsmap[5] and the sixth city with the highest ratio of pictures of visitors versus its official population [11], based on Flickr usage. This popularity suggested that it would be a thoroughly visited and photographed destination, and therefore it was expected to provide a large amount of data for analysis.

Another reason was diversity of the urban fabric of Barcelona, where the historical evolution of the city has shaped a complex built environment [12], with different neighborhoods with their own distinct character and functional mix, providing opportunities to test the developed methodologies in a variety of urban settings.

The last reason was the author's familiarity with the city. This personal knowledge allowed validating the results obtained as the research progressed, or in some occasions surprise the author when results did not match expectations, generating new research questions. The developed methodology was later tentatively applied

[3]Euromonitor estimates a 5.7% year-over-year growth from 2014 to 2015.

[4]Airbnb offered 23,000 listings in Barcelona during 2016, according to Euromonitor.

[5]Sightsmap uses data from Panoramio and is available at *http://www.sightsmap.com/* at the time of writing.

to another 19 cities, and the initial findings suggested that it was robust and reproducible.

1.2 Research Propositions

1.2.1 Hypothesis

The main premise of the dissertation is that social media data has the capacity of providing new data-driven avenues in urban research, as a complement to existing and well-established approaches. However, the unique challenges of these data[6] —volume, velocity, variety and veracity— require the introduction of new retrieval, wrangling, analysis and visualization workflows based on the principles of data science, adapted to the field of urban research.

Because of the complex and multifaceted nature of the city, it is necessary to develop multiple approaches to convert these data into information. This dissertation discusses a variety approaches tailored for the study different aspects of public life at different scales, topologies and dimensions, to transform raw data into knowledge in a principled way.

1.2.2 Aims and Objectives

The aim of this research is to develop a deeper understanding of the potential of social media data analysis as an emerging data-driven methodology for the study of the city, that can complement more classic approaches in urban research. To achieve this aim, the following objectives were set:

1. To identify novel data-driven approaches using social media data in the field of urban research, and in other disciplines where data collection to empirically understand human or social behavior has traditionally been a challenge.
2. To develop methodologies to collect data from increasingly available sources of user-generated content shared on social networks, in particular those capable of providing spatial, temporal, textual and/or multimedia information.
3. To implement the best practices of reproducible research into urban analysis, and develop new methodologies capable of processing very large and complex urban data sets, favoring the use of open source tools.

[6] Sometimes summarized as "the four V's of Big Data", as explained in the IBM Big Data & Analytics Hub infographic, available at *http://www.ibmbigdatahub.com/infographic/four-vs-big-data* at the time of writing.

4. To define a set of representation techniques capable of summarizing the multifaceted complexity of urban phenomena in multiple spatial and temporal scales, suitable for exploratory analysis and the effective communication of results.

1.3 Thesis Overview

1.3.1 Outline

After an introductory chapter discussing the importance of data in urban research (chapter 2), the next two chapters (chapters 3 and 4) describe the data retrieval and handling methodologies developed. The next chapters explore different facets of Barcelona from these data, in three spatial scales —macro (chapter 5), meso (chapter 6) and micro (chapter 7)—, followed by two less conventional approaches —focusing on network-constrained analysis (chapter 8) and temporal cycles (chapter 9)—, and the last chapter wraps up the conclusions. The chapters are mostly self-contained, with their own introduction, literature review and conclusion where applicable.

Citations and references to literature are provided according to the format defined by the Institute of Electrical and Electronics Engineers (IEEE) transactions, journals and conferences. However, since the research was developed within an emergent field of study, a number of references are from web resources instead of formal publications and are provided as footnotes with their corresponding links.

To address the issue of web citation decay [13], as an estimated 13% of Internet references in scholarly articles were found to be inactive 27 months after publication [14], most of the links provided (valid at the time of writing) should be recoverable in the foreseeable future through web archiving services, using archive.is[7], the Internet Archive Wayback Machine[8] or WebCite[9] as proxies.

1.3.2 Chapter 2: The Urban Data Revolution

Chapter 2 briefly reviews the methodological approaches in the study of urban phenomena, distinguishing between quantitative and qualitative strategies.

[7]The archive.is (formerly known as archive.today) on-demand archiving service is available at *http://archive.is/* at the time of writing.

[8]The Internet Archive Wayback Machine digital library is available at *http://web.archive.org/* at the time of writing.

[9]The WebCite on-demand archiving service is available at *http://www.webcitation.org/* at the time of writing.

The importance of modeling is recognized as the necessary synthesis of these approaches, stressing the dependency of models on data and highlighting two examples of models in the fields of architecture (pedestrian behavior) and urban studies (urban transformation processes). The new research paradigm based on the abundance of data and supported by data science is discussed next, showcasing examples of data-driven research in urban studies and other domains, followed by a discussion of the effective representation of these spatial and temporal data in multiple scales. Finally, the limitations of tracking technologies to capture human movement are examined, followed by a discussion on how the digital traces of the online world, analogous to the physical traces in the real word, can be a valuable source of data to study urban phenomena.

1.3.3 Chapter 3: Collecting Digital Traces

Chapter 3 discusses the retrieval methodology from the sources studied in the research. After a brief introduction explaining the fundamental differences between the classic (static) websites and the current (dynamic) Web 2.0, an overview of the different services compares the four selected services (Panoramio, Flickr, Twitter and Instagram) discussing their popularity, the features offered programmatically through their web interfaces, and the differences in quantity and quality of the collected data. Next, each service is separately discussed in more detail using a common outline: first a service overview, followed by a description of its features and limitations, an in-depth review of the available data, and a map of the point locations retrieved. The chapter wraps up with a comparison of the different spatial distributions in a selection of landmarks within Barcelona.

1.3.4 Chapter 4: Determining User Origins

Chapter 4 begins discussing the methodology developed to extract the hometowns from the public profiles of Flickr users who took a geolocated pictures of Barcelona, as well as the geocoding process to convert this information into geographical coordinates and administrative units. The second half of the chapter discusses the definition of the geographic units and aggregation levels that will be used in chapter 5, classified into seven concentric regions around Barcelona (municipal, metropolitan, regional, autonomic, national, supranational, and international).

1.3.5 Chapter 5: A Global Perspective

Chapter 5 focuses on the global scale, turning the classic spatial analysis approach inside-out and changing the research question from “where” into “*from* where”.

Beginning with the challenges of visualizing world-scale data, the unavoidable trade-offs of projecting a geoid into a flat plane are discussed, and the treemap method for displaying hierarchical data is introduced. Next, the aggregated picture and user counts are analyzed and mapped according to their cities and countries of origin (as defined in chapter 4) using suitable visualization techniques. Finally, the influence of each country's population and the size of its economy on the results as explanatory variables are also explored, finishing with an estimation of their relative economic contribution to the wealth of the city, according to the gross domestic product per capita of its visitors.

1.3.6 Chapter 6: The City Scale

Chapter 6 narrows the scope down onto the urban scale, and discusses the issues regarding the analysis and representation of very large spatial data sets consisting of point *events*, exploring three different approaches to mitigate these issues. The first approach consists in the collection of events into a fine grid (rasterization), and discusses the cell size selection criteria, the transformation of the cell counts, and the mapping of values to a color scale. The second approach aims to avoid some biases introduced by rasterization using a rigorous spatial statistics framework, computing two local measures of spatial autocorrelation —cluster and outlier detection using local Moran's I and hot and cold spot analysis using local Getis-Ord's G*— of events collected in a hexagonal lattice, and discusses the optimal tile size and the conceptualization of spatial relationships. The third and final approach aims to circumvent the discretization step through a smoothing estimation of intensity using the kernel density estimation, and discusses the optimal kernel function and bandwidth selection criteria, and the appropriate transformation and classification methods.

1.3.7 Chapter 7: The Neighborhood Scale

Chapter 7 focuses on the neighborhoods —the smallest administrative units in the city of Barcelona— as suitable aggregation units instead of the regular tessellations discussed in chapter 6. After discussing the limitations of choropleth maps, linked micromaps are introduced as a powerful alternative that enhances classic visualization techniques with the spatial reasoning of maps, presenting the methodology developed to produce an optimized neighborhood geometry using topology-aware simplification and small polygon exaggeration. This approach is used to explore bivariate relationships (picture counts per area, per capita and per road length), time series data, and the temporal cycles further discussed in chapter 9, compared to conventional visualization techniques.

1.3.8 Chapter 8: A Network Approach

Chapter 8 begins avoiding the assumption of spatial homogeneity considered in chapter 6, and focuses on the analysis of the event density constrained to the street network. After examining the heterogeneity of the urban fabric of Barcelona by mapping the distribution of the postal addresses and road network densities and the variation of the distances from the building entrances to their road segment axes, two complementary and computationally intensive methodologies are discussed to obtain event densities along the street network. The first methodology matches each event to the nearest road segment, either directly (shortest distance) or through the nearest street address (acting as an anchor), and the resulting event counts are divided by the corresponding segment length. In contrast, the second methodology uses a generalization of the kernel density estimation in a linear network instead of a two-dimensional plane. The results of both methodologies applied to the collected data are mapped, and the suitable data transformation and visualization methods are discussed.

1.3.9 Chapter 9: The Temporal Dimension

Chapter 9 leaves the spatial domain and focuses on the distribution of the collected temporal data. After introducing the complexity of handling temporal data, the required transformation and filtering of the retrieved data is discussed. Two complementary visualization and analysis techniques based on the aggregation of data into temporal bins are used to track the evolution of the retrieved events: time series decomposition to extract the trend and seasonal components, and calendar heatmaps to discover more subtle temporal patterns. Finally, three overlapping cycles of the life of the city are explored in detail through the collected data: the daily cycle (night and day), the weekly cycle (work and leisure) and the yearly (seasonal) cycle, concluding with a representation of the matrices of pairwise cycle interactions using heatmaps.

1.3.10 Conclusions

The final chapter presents the conclusions of the dissertation discussing the four subgoals defined in the introduction: researching using urban data, the retrieval methodologies, handling the complexity of urban data, and the most appropriate visualization strategies. The chapter wraps up with a discussion of the future research directions and applications.

Chapter 2

The Urban Data Revolution

> "Errors using inadequate data are much less than those using no data at all"
>
> Charles Babbage

2.1 The Complexity of the City

2.1.1 The Study of the City from a Scholarly Point of View

In his influential "The Two Cultures" [15], Charles Percy Snow distinguishes two cultures in the western society, one corresponding to *science* and another corresponding to *humanities*. Juval Portugali [16] recognizes that this division can also be observed in the study of the city (sometimes as qualitative and quantitative approaches), but includes a third category, corresponding to recent developments in complexity theory.

Despite having existed for thousands of years, the scholarly study of cities has primarily taken place from the 20th century onward, although the roots of the knowledge area that is known as "urban studies" can be traced to the 19th century, with the Adam Smith's inspired economic geography of Johann Heinrich von Thünen[1] [17], contrasting with the informal observation of Paris of Charles Baudelaire as a Flâneur [18] or the aesthetic approach to the design of the urban space of Camillo Sitte [19].

[1] Location theory was further developed by Alfred Weber (Über den Standort der Industrien, 1929), Walter Christaller (Die zentralen Orte in Süddeutschland, 1933) and William Alonso (Location and Land Use, 1964).

The first part of the century, before World War II, was marked by a functional view of the city, and the formulation of central place theories [20] in parallel to the conceptualization of the attractiveness of city centers as a gravitational law [21]. The design of the layout of the new cities —or in some cases the expansion of existing ones— was influenced by these concepts, such as the project of *la ville radieuse* (the radiant city) by *Le Courbusier*[2] theorized in *la Charte d'Athènes* (the Athens Charter) on 1933. At the same time, despite the prevalent reductionist approach towards urban phenomena, some voices focused on the effects of city living on its inhabitants [22].

After World War II, the pendulum swings away from this functional approach, as its issues of a segregated and specialized city began to be more widely criticized through the works of Lewis Mumford [23], Jane Jacobs [24] and Christopher Alexander [25]. At the same time, Kevin A. Lynch began his study on how observers produce their own mental maps of the city [26] from five elements (paths, edges, districts, nodes and landmarks). The quantitative approach did not disappear [27] and the functional approach was pursued in the fields of economics on location theory by August Lösch [28] and William Alonso [29].

2.1.2 Contemporary Approaches to City Complexity

As urbanization processes increased [30], planning focused on achieving a more efficient use of resources [31], making cities more compact [32] and densely populated [33], increasing the number of interactions [34], and therefore requiring improved transportation [35].

Research interest shifted to studying the complexity of the evolution of the morphology of the built environment [36, 37] and its street network [38], and the organization of the city [39] through the interactions of its inhabitants [40, 41], as well as regional dynamics [42], using assemblage thinking to study relationships instead of things [43].

Urban modeling, understood as building of mathematical models of cities and regions [44] began in the early 1960s. One of the earliest models that described an urban processes as an emergent process was the Schelling Segregation Model (also known as "tipping model") [45], which modeled the observed segregation in neighborhoods [46]. This model[3] showed that even starting from an integrated initial condition, and despite the agents only having a mild segregation preference, the patterns could become heavily segregated (Fig. 2.1).

[2] Neé Charles-Édouard Jeanneret.

[3] An example NetLogo model illustrating this process is available for download at *http://ccl.northwestern.edu/netlogo/models/Segregation* at the time of writing.

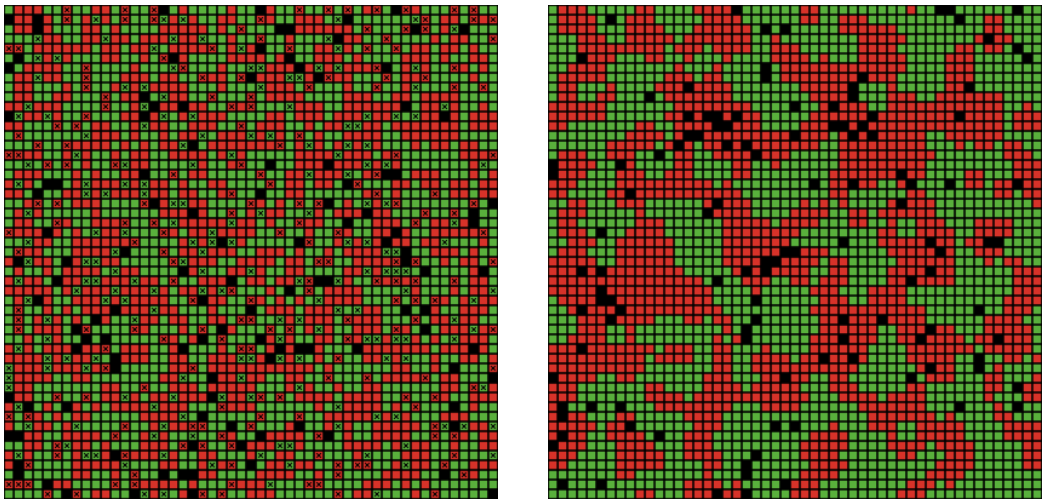


Figure 2.1: Segregation example model in NetLogo developed by Wilensky (1997), where the initial condition is not segregated, and the simulated inhabitants only have a 30% similarity preference, with the initial condition (left) and final segregated result (right).

The urban complexity also influenced urban and architectural design [47]. In particular, the multidisciplinary applied research methodology [48] of danish architect and urban designer Jan Gehl on the use [49] and perception [50] of public spaces and city life [51], opened the door for more human-centered design [52], with easier to navigate spaces [53] because of a greater understanding of the formal elements in the city [54].

This approach pioneered by Gehl, based on the observation and documentation of urban phenomena, can be complemented by the unprecedented amount of data available today. According to the Cities of Data project[4] (Fig. 2.2), the interest on urban data research is accelerating [55] because of the impulse of urban science institutions[5] that involve multidisciplinary teams such as Sidewalk Labs[6], an Alphabet[7] subsidiary focused on urban innovation.

[4]The Cities of Data website is available at *http://www.citiesofdata.org/* at the time of writing.
[5]A detailed inventory of academic institutions, laboratories, and organizations in urban science and informatics research is available at *http://www.citiesofdata.org/foundations-of-urban-science/proj-inventory/* at the time of writing.
[6]Sidewalk Labs website is available at *http://www.sidewalklabs.com/* at the time of writing.
[7]A multinational conglomerate and parent of Google.

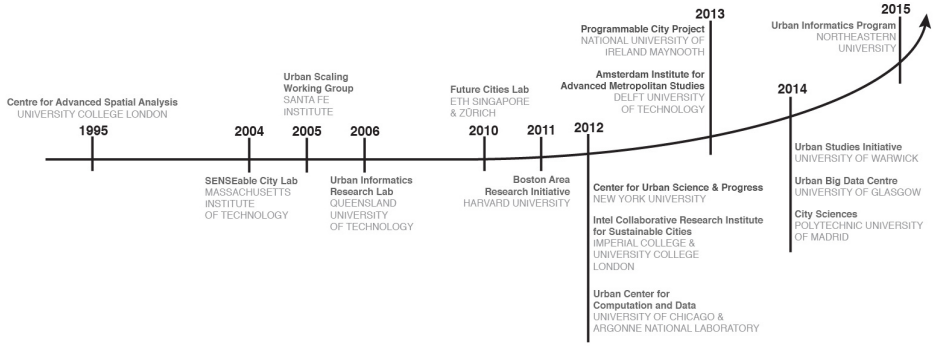


Figure 2.2: Timeline of new urban science institutions. Source: Cities of Data website.

2.2 Models, Data, Architecture and Urban Space

2.2.1 Why Model?

According to the Internet Encyclopedia of Philosophy[8], a model is a representation of some object, behavior, or system that one wants to understand. Models[9] are useful beyond their capacity to predict future outcomes, and Epstein proposes 16 reasons other than prediction to build a model [56], highlighting their capacity to explain phenomena explicitly by defining their underlying mechanisms[10]. Thus, models allow discussion between stakeholders and possess important educational values, especially if they are easily understandable, becoming "illuminating abstractions".

Oftentimes, we do not even realize we are using a model[11] because we are using an implicit one, where according to Epstein "the assumptions are hidden, their internal consistency is untested, their logical consequences are unknown, and their relation to data is unknown".

Even when we are using an explicit model, we might not be aware we are doing so; for example, the widely used regression analysis is also a model —a statistical

[8]The Internet Encyclopedia of Philosophy is available at *http://www.iep.utm.edu/models/* at the time of writing.

[9]Models in science are thoroughly discussed in the Stanford Encyclopedia of Philosophy available at *http://plato.stanford.edu/entries/models-science/* at the time of writing.

[10]However, Epstein also criticizes the common analogy of comparing brains and computers and the information processing metaphor, in an essay available at *http://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer* at the time of writing.

[11]We arguably *reason* about the world using models, although not always formal or explicit ones.

one—, where we assume that the relationships between variables are linear[12] [57].

Models are useful tools to study complex phenomena, and are increasingly used in the social sciences to understand the mechanisms that explain the observations, and in some cases to make predictions. In particular, the fields of architecture and urban design can benefit from models of pedestrian behavior or urban dynamics (further discussed in subsections 2.2.2 and 2.2.3 respectively).

2.2.2 Modeling Pedestrian Behavior

Pedestrian behavior models [58] can be considered as a specialized class of models that simulate the activities or some aspect of the conduct of a person or group in an urban or architectural[13] environment.

One of the most developed areas of interest on pedestrian behavior, because of its implications on safety[14], is egress behavior from a crowded space in an emergency situation, and specialized building evacuation models for specific situations have been developed [59], and documented in the comprehensive Pedestrian and Evacuation Dynamics series [60, 61, 62, 63, 64]. Other mature areas on pedestrian behavior research are the models of retail [65], simulation [66] and visual analysis [67].

Crowd modeling and simulation technologies assume some factors that influence the simulated behavior and can be broadly divided [68] into three categories —physical, social and psychological—, which require real-world experimental data[15] to be able to quantify their influence, and can be classified according to three modeling approaches regarding their complexity, which in turn affects the maximum number of simulated elements (scalability), the simulation speed, and the application area [68]:

Flow-based approaches model the crowd movement as the continuous flow of a fluid instead of individual elements, allowing the simulation of very large and dense crowd sizes at a reasonable computational cost, sacrificing the increased performance at the expense of accuracy, and are therefore suitable for real-time applications.

Entity-based approaches model elements as a set of homogeneous entities. These entities are treated as a particle system governed by forces[16] that

[12] And in the case of statistical significance testing, that the variables are normally distributed.

[13] Or in some cases, a natural or quasi-natural environment.

[14] And also because of the possibility to be analyzed using tools developed for transportation engineering.

[15] Which is very difficult and costly to obtain or replicate.

[16] For example Helbing's social force model.

simulate the influence of different factors (local and/or global) that affect the particle movement.

Agent-based approaches model each individual as an autonomous agent with rules that determine its decisions in the simulated world. The agents can have different degrees of rule complexity and cognition, and generally operate on local information about the environment surrounding the agent. Agent-based models can be used to study movement geography [45] at macro, meso and micro scales of study.

More recently, a more nuanced classification has been proposed according to the distinct methodologies of modeling pedestrian movements [69], according to whether they model the interactions at the microscopic (individual level) or operate at the macroscopic (system) level.

2.2.3 Modeling Urban Processes

To explain the overwhelming complexity of modern cities, urban models describe city processes and therefore reach a larger spatial and temporal scales than pedestrian models[17], as they simulate phenomena in a metropolitan or even territorial scale, and the actions of the simulated individuals can span decades or centuries (and therefore the simulated behavior is necessarily less detailed).

Models operating on the meso-scale can involve elements of both pedestrian and urban modeling in the street network [70], simulating their spatial awareness and preferences, or modeling the intelligibility of the street layout [71] from the pedestrian point of view.

Contemporary approaches to urban modeling [72, 73, 74] rely in computer-based simulations, and urban models have been developed on many key areas of the study of the city: land use patterns [75, 76], land cover change [77, 78], sprawl [79, 80], zoning [81], and spatial perception [82].

Models of urban systems have also been developed to simulate the interactions of networks of cities on a territorial scale [83] and even global[18] models in a four-century time scale [84]. Beyond studies of contemporary phenomena, the simulation of the Anasazi people rapid decline and final abandonment of the Long House Valley in the 14th century [85] has been explained in archaeology using a multiagent computational model of monoagriculturalist agents.

Specialized software has been developed to aid the creation of urban models on

[17] A similar relationship can arguably be found between architecture and urbanism.
[18] Using data available for United States and Europe.

GIS platforms such as the Agent Analyst extension[19] [86] for ArcGIS (ESRI), the Land Change Modeler[20] for TerrSet (Clark Labs), and the GIS extension[21] for the open-source NetLogo[22] [87] agent-based software development platform.

2.2.4 Models and Data

For any model to be able to make predictions[23], it is required to be fed enough data to fit the model against, with the standard approach being the partition of data into a *training set* to calibrate the model parameters and a *test set* to asses the model predictive performance.

Moreover, the parameters of the models need to be rigorously calibrated[24] [44, 56] using real-world data, and their results should in turn guide data collection in a feedback loop. Therefore, data and models can be considered inextricably linked.

Unfortunately, data in the social sciences has been sparse and usually restricted to WEIRD (Western, Educated, Industrialized, Rich and Democratic) subjects[25], until very recently with the availability of large data sets mined from Web 2.0 content [26], where users can be considered participants of a natural experiment.

Under the assumption that it is possible to infer behavior from traces left in the environment, we can analogously[27] use traces in the digital world to build, test and refine models. The research of this thesis focuses on how these data can be collected, transformed and used to analyze urban processes, hopefully becoming the foundation to support or refute theories while suggesting new avenues of research.

[19] The documentation for Agent Analyst is available at *http://resources.arcgis.com/en/help/agent-analyst/* at the time of writing.

[20] The Land Change Modeler product page is available at *https://clarklabs.org/terrset/land-change-modeler/* at the time of writing.

[21] The documentation for the GIS extension for NetLogo (developed by Eric Russell) is available at *http://ccl.northwestern.edu/netlogo/docs/gis.html* at the time of writing.

[22] NetLogo uses the Java Virtual Machine but it is written primarily in Scala, and its syntax is loosely based in Logo and Lisp.

[23] Although some models are only explanatory but not predictive.

[24] Or at least the results of the simulations should *resemble* the observed phenomena.

[25] Typically US undergraduates in artificial laboratory conditions, as discussed by Robert Colvile in the essay available at *https://aeon.co/essays/american-undergrads-are-too-weird-to-stand-for-all-humanity* at the time of writing.

[26] Although these data is also inevitably biased, as discussed in chapter 3.

[27] No pun intended.

2.2.5 Urban Data Sources

Most of these data is locked into the digital vaults of corporations, where where they configure the basis of their business intelligence strategy, and are therefore proprietary. However, from the study of the city standpoint, there are three valuable data sources worth mentioning which provide unprecedented amounts of machine-readable data:

Open data or more specifically open-data government initiatives, understood as the counterpart of similar open movements —of which the open-source movement is probably the most well known—, and exemplified in the initiatives of the governments of USA[28], UK[29] or EU[30], as well as regional (as in the case of Catalonia[31]) and local (as in the case of Barcelona[32]) governments, which allow free public access to most of their privacy-insensitive data.

User-generated content shared in Web 2.0 platforms where users are both producers and consumers of content[33]. User-generated content is not limited to social media alone but also includes blogs, wikis, forums, business reviews, product recommendations, or video and photo sharing.

Internet of things (IoT) as the network of Internet-enabled devices, capable among other things of collecting data through their sensors and exchanging it through the Internet infrastructure.

While until recently most of urban research relied on census [88] or cadastral data when available [6, 89], new sources [90] are beginning to gain traction [91], coming from companies that are willing to share some of their data such as Uber[34], or through data compiled from public-facing websites of companies offering services, such as the Inside Airbnb[35] initiative, which provides data from Airbnb as a form of hacktivism to make their data open and available to the public. Furthermore, the availability of geographically located pictures taken by volun-

[28]The United States of America's open data portal is available at *http://www.data.gov/* at the time of writing.
[29]The United Kingdom's open data portal is available at *http://data.gov.uk/* at the time of writing.
[30]The European Union's open data portal is available at *http://data.europa.eu/* at the time of writing.
[31]The Catalan Government's open data portal is available at *http://governobert.gencat.cat/* at the time of writing.
[32]The Barcelona's open data portal is available at *http://opendata-ajuntament.barcelona.cat/* at the time of writing.
[33]TIME Magazine appropriately named "You" as the Person of the Year in 2006.
[34]The Uber Movement website is available at *http://movement.uber.com/* at the time of writing.
[35]The Inside Airbnb data is available at *http://insideairbnb.com/* at the time of writing.

teers, from photo sharing sites but also from official initiatives such as Geograph[36] —a project that since 2005 aims to obtain a picture for each of the 331,957 squares measuring 100 ha in UK[37] and Ireland[38]—, provide valuable data for urban research.

2.3 Data-Driven Research

2.3.1 The Unreasonable Effectiveness of Data

The subsection title is borrowed from the paper of the same name [92], which is itself a play on the title of an influential Eugene P. Wigner paper published in 1960 and titled "The unreasonable effectiveness of mathematics in the natural sciences" [93]. It is also the title of Peter Norvig's lecture[39] of September 23, 2010 in the University of British Columbia Department of Computer Science, which is brilliantly summarized in its subtitle: "How Billions of Trivial Data Points can Lead to Understanding".

In this lecture, Norvig argues that beside the traditional scientific method to understand the world —where the scientist observes the world, tries to abstract its observations into an idea, which is then formulated into a theory[40]—, it is possible a different approach[41], where data can be used to feed computer models to extract understanding, provided the size of data is sufficiently large.

This paradigm shift in the way of conducting science [94] is summarized by Miller and Goodchild [95, page 449] as follows:

> The ease of collecting, storing, and processing digital data may be leading to what some are calling the fourth paradigm of science, following the millennia-old traditional of empirical science describing natural phenomena, the centuries-old tradition of theoretical science using models and generalization, and the decades-old traditional of computational science simulating complex systems.

This approach can provide unprecedented research opportunities for domains of knowledge that have traditionally lacked access to data to support their theories,

[36] The Geograph Britain and Ireland project is available at *http://www.geograph.org.uk/* at the time of writing.

[37] Ordnance Survey National Grid.

[38] Irish national grid reference system.

[39] A video of Peter Norvig's lecture is available on YouTube at *https://youtu.be/yvDCzhbjYWs* at the time of writing.

[40] Which, using Popperian terms, should be possible to falsify.

[41] It is arguably a quantitative distinction, as the scientific method is (or should be) always be based on data.

especially in the social sciences, and in particular has the potential to be a useful tool in the field of urban studies, either to provide answers to old questions or, more importantly, to suggest new ones.

2.3.2 The Age of "Big Data"

According to some sources[42], 90% of the World's data has been created in the last two years, and 2.5 exabytes[43] of data are being produced daily[44], a figure that is expected to grow to 44 zettabytes[45] per day by 2020. As an example, 400 hours of video were uploaded to YouTube every minute in 2015[46], an extraordinary growth for a service whose first video was uploaded on 23 April 2005.

Most of these data [96] is denominated under the umbrella term of "Big Data"[47] [97], a much debated topic without a clear consensus regarding its definition and which seems to have the potential of becoming quickly obsolete[48]. While most of the definitions of Big Data focus on its volume —hence the term *big*—, Miller and Goodchild [95] remark three significant challenges when dealing with Big Data, identified as "the three Vs":

Volume as the amount of data that can be collected and stored by the most powerful computational systems, which usually entails the impossibility to keep the entire data sets in memory.

Velocity as the speed at which data can be used to extract knowledge (queried, analyzed and visualized), which users expect to be near real-time, and can quickly become obsolete, particularly when dealing with the continuous arrival of time-stamped real-time data.

Variety as the result of data being collected from increasingly heterogeneous sources obeying to multiple necessities, and generally structured very differently or simply stored unstructured, arising the necessity to be validated and processed to be integrated into a coherent analysis workflow.

---

[42] According to Science Daily in the post "Big Data, for better or worse: 90% of world's data generated over last two years" published on May 22, 2013, available at *https://www.sciencedaily.com/releases/2013/05/130522085217.htm* at the time of writing.

[43] One exabyte = $1{,}000^6$ bytes.

[44] With 2016 data, according to the GigaOm post "Artificial Intelligence: It's Not Man vs. Machine. It's Man And Machine" posted on October 5, 2016, available at *https://gigaom.com/2016/10/05/artificial-intelligence-its-not-man-vs-machine-its-man-and-machine/* at the time of writing.

[45] One zettabyte = 1000^7 bytes.

[46] According to data gathered by Statista, available at *https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/* at the time of writing.

[47] The term is usually written with both words capitalized.

[48] As is arguably happening at the time of writing with the term "Smart City".

Some authors add a fourth "V", including the term *veracity* to emphasize that the quality of data must be assessed, especially documenting its errors, omissions and biases [98], avoiding the propagation of quality issues in its validity that can influence the decision making process.

This abundance of data is fueling what some authors denominate "the fourth industrial revolution" [99], providing the raw material to train the sophisticated artificial intelligence algorithms [100] that seem capable to automate tasks and solve problems that today seem unthinkable[49].

2.3.3 Data Science Workflow

It is crucial to adopt an adequate workflow to handle this abundance of data and transform it into knowledge, otherwise we will be data-rich but information-poor. According to Wickham and Grolemund, "Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge." [101, page ix].

These authors suggest an approach to data science where the complete workflow takes place inside a literate programming [102] environment, to produce reproducible results [103], using the best practices [104] in scientific research.

This tried and tested approach guided the research methodology for this thesis, and its workflow is summarized in a diagram[50] [101, page ix] that include four distinct phases (Fig. 2.3):

Import is the first step of the analysis, when source data must be translated into data structures —generally tabular[51] or tree-like[52]— and data types capable of supporting the analysis process. This raw data must be converted from multiple (and sometimes poorly documented) formats stored in files, databases or web services (discussed in chapter 3).

Tidy is then next step after import, and is geared towards making the data consistent, where according to Wickham [105] each variable is a column, each observation is a row, and each type of observational unit is a table. This process generally requires some kind of data reshaping and correct

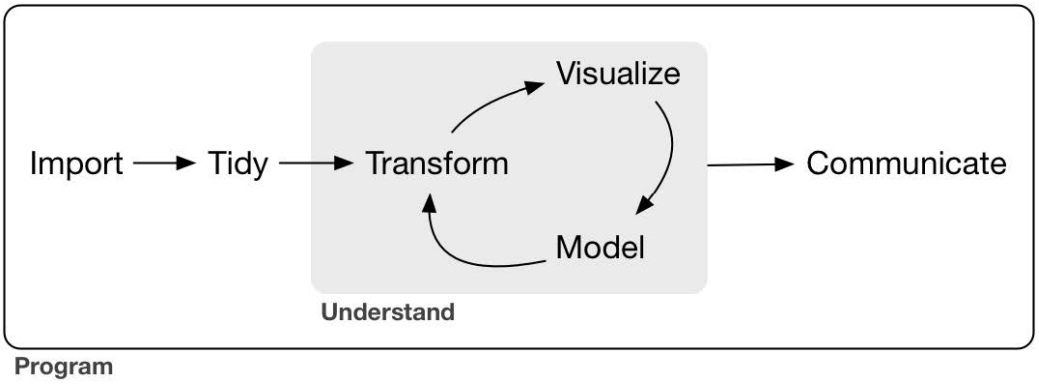
[49] Such as beating a Go master, according to the Scientific American post "How the Computer Beat the Go Master", published on March 19, 2016, available at *(https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/* at the time of writing.

[50] The diagram appears in the online version of the book "R for Data Science" (Hadley Wickham and Garrett Grolemund) available at *http://r4ds.had.co.nz/* at the time of writing.

[51] For example a data.frame, data.table or tibble object in the R language, or a DataFrame object (pandas library) or numpy array in Python.

[52] For example nested lists in R and Python.

Figure 2.3: Hadley Wickham and Garrett Grolemund model of the workflow in a typical data science project. Source: R for Data Science (http://r4ds.had.co.nz/), online edition of the book of the same name.



handling of missing data[53] [106].

Understand is an iterative process where two complementary knowledge generation approaches are possible: visualization and modeling; visualization can reveal unexpected patterns in the data, and raise new questions or challenge a working hypothesis, but depends on human interpretation, while models, as computational or mathematical tools, do not have these issues, but require a more precise formulation of the problem.

Communicate is the last (but not least) stage of data science, where the resulting models and visualization are disseminated to the community. Effective communication of the findings of a research is therefore an essential step in the scientific progress.

Models are in general regarded as tools for hypothesis confirmation, while visualization is considered a data exploration tool geared towards hypothesis generation. However, these conventional roles are not clear-cut and can be reversed [107], with the difference lying in whether observations are used *one* (confirmation) or *many* (exploration) times [101], regardless of the approach used.

2.3.4 Data-Driven Approaches

These novel data-driven approaches can support advances in domains of knowledge which until recently were eminently speculative, using data science to enhance the traditional scholarly approaches, either using exploratory data analysis as a hypothesis generation mechanism, or as a foundation to rigorously prove or refute hypotheses.

[53] Which on occasions can be considered part of the transformation stage.

In fact, the availability of data collected by journal indexing databases such as Scopus[54] or Web of Science[55] have allowed the study of the very process of scientific production, through the relationships of published papers —more than 50 million in 2009 [108]— and their authors, answering questions about the relationship between number of references and paper length [109], the citing behavior of authors [110, 111, 112], the consequences of rebuttals [113], the existence of "scientific memes" [114], and the evolution of the scientific impact of authors throughout their career [115].

This approach surprisingly resonates in domains such as philology [116], where the digitization of large corpus of literature done by Project Gutenberg[56] or Google Books[57] is allowing the application of artificial intelligence technology [117] to the analysis of literary works[58]. These analysis are not achievable in a practical amount of time for a reasonable-sized team of experts, as for example in the content analysis of 150 years of British periodicals to identify patterns of cultural change [118].

Furthermore, the analysis of thousands of stories has allowed to extract the basic shapes of their emotional arcs [119], reducing the stories to just six basic patterns. Another study has been capable of following the evolution of stories [120] through successive retellings over time, analyzing 427 versions of the Little Red Riding Hood tale and a corpus of 900 chain letters.

Some other research relies on indirect measures through data collected for other purposes, such as using seafood prices to analyze the ecologic impact of dissolved oxygen [121], using satellite-recorded nighttime lights to estimate the GDP, modeling the criminal records as a social network to predict gunshot violence [122], or using Twitter data to predict heart disease mortality [123]. However, these approaches has also attracted criticism[59], such as Google Flu Trends[60], which used search queries to estimate the influenza activity [124].

Of these indirect measures, social media seems to have the most potential [125, 126], and can be used for large scale studies which would be otherwise impossible

[54] The Elsevier's Scopus database was available at *http://www.scopus.com/* at the time of writing.

[55] The Clarivate's Web of Science database is available at *http://clarivate.com/products/web-of-science/* at the time of writing.

[56] The Project Gutenberg website is available at *http://www.gutenberg.org/* at the time of writing.

[57] Google Books is self described as "the world's most comprehensive index of full-text books", and is available at *http://books.google.com/* at the time of writing.

[58] According to Leonid Taycher, software engineer at Google, there were 129,864,880 books in August 5, 2010, based on the information posted in the blog entry available at *http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html* at the time of writing.

[59] See the 2014 Tim Harford article titled "Big data: are we making a big mistake?", available at *http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html* at the time of writing.

[60] The service is discontinued but historic estimates are still available at *http://www.google.org/flutrends/about/* at the time of writing.

because of the large required sample sizes and time spans covered. These sources are also capable to collect semantically rich data with additional metadata (time-stamps, location, relationships between participants) and has been used to analyze complex phenomena, such as the 2011–12 Russian protests [127].

Data is can also be used for artistic purposes, as in the mesmerizing images —which resemble the dioramas of Japanese artist Sohei Nishino[61]— created by Google engineer Alexander Mordvintsev using the DeepDream[62] software—, operating a neural network trained for computer vision in reverse, and also in the videos[63] produced from Google Street View imagery using the Hyperlapse[64] software. Similarly, a machine vision algorithm trained with images of Jackson Pollock's drip paintings was capable of differentiate between original and non-original artworks in 93% of the cases [128].

2.3.5 Urban Data Applications

The use of quantitative data to understand urban phenomena is not new, as governments (from local to global) have been collecting statistical data for over a century [129], and data collection efforts for the census[65] can be traced to the Babilonians (3800 BC) and the Chinese Han Dynasty (2 AD), while the term Cadastre (as "Capitastrum") appears for the first time in 1669 in a Gilles Ménage (Aegidius Menagius) text [130].

The availability of worldwide datasets has allowed the discovery of scaling laws in several city metrics (e.g. crime, gross domestic product, income, patents) [2, 131] and formulate the underlying law that explains the phenomena as the efficiency of the energy spent on moving people, goods, and information across urban networks [132].

Furthermore, the growing availability of data benefits from using Geographic Information Systems (GIS) technologies [133], to integrate geospatial and temporal data [134], prepare data for simulation ingestion (e.g. network topology

[61] The exhibition review in The Guardian "Unreal cities: Sohei Nishino's magical photographic maps of London, Tokyo and utopia" of February 24, 2010 is available at *http://www.theguardian.com/artanddesign/2011/feb/24/sohei-nishino-diorama-maps* at the time of writing.

[62] DeepDream sample code is available at *http://github.com/google/deepdream* at the time of writing.

[63] Produced by Teehan+Lax and available on Vimeo at *https://vimeo.com/636538730* at the time of writing.

[64] Project hosted on GitHub at *https://github.com/TeehanLax/Hyperlapse.js* at the time of writing.

[65] In the Bible census is mentioned in the Book of Numbers (counting the Israelite population during their flight from Egypt) and the travels of Mary and Joseph to Bethlehem to be enumerated in the census.

construction, least cost pathway creation) and analyze post-simulation data using geostatistical methods.

Some of the earliest experiences of analyzing the city through data [3] took place from 2004 at the MIT SENSEable City Lab[66] —under the direction of the architect and technologist Claudio Ratti—, which were granted access to data from service providers such as telecommunication companies [135, 136], in close collaboration with governments.

The application of big data to the city [137], introduces new difficulties because the huge data sets involved, which complicates handling and visualization of data, but also because source data is invariably unstructured [138].

These data-driven approaches are also beginning to help cities make informed decisions, providing policymakers new tools to check the "pulse" of the city. One of the most popular tools are dashboards [139] that allow stakeholders to monitor aspects of the city in real-time[67], making cities smarter [140], an approach essayed by many cities in the UK [141], but also world heritage sites like Venice [142].

Furthermore, the semantic information contained in user-generated content can be used for sentiment analysis and to search for keywords, and has been used by the author to define a functional program for a public space, using a similar experience as a reference, using a text mining approach [143].

This approach also opens unprecedented new avenues of data-driven experimental design in urban research; in a recent online experiment [144], volunteers were asked to observe pairs of geotagged street level images and compare their perceived relative safety, social class and uniqueness. This allowed the production of maps of their urban perception and also to analyze the correlation of the results against the rates of violent crime. More recently, a similar study used Google Street View imagery and machine learning to track changes in neighborhoods and their correlation to safety [145].

Beyond the urban scale, new data sources open an avenue to study travel patterns on a global scale [11], tracking the behavior of users of photo sharing communities, to understand the impact of tourism on world-class cities (a similar approach for the case of Barcelona is discussed in chapter 5).

[66] The MIT SENSEable City Lab research group website is available at *http://senseable.mit.edu/* at the time of writing.

[67] For example the BarcelonaNow dashboard (part of the DECODE Project) available at *http://bcnnow.decodeproject.eu/* at the time of writing.

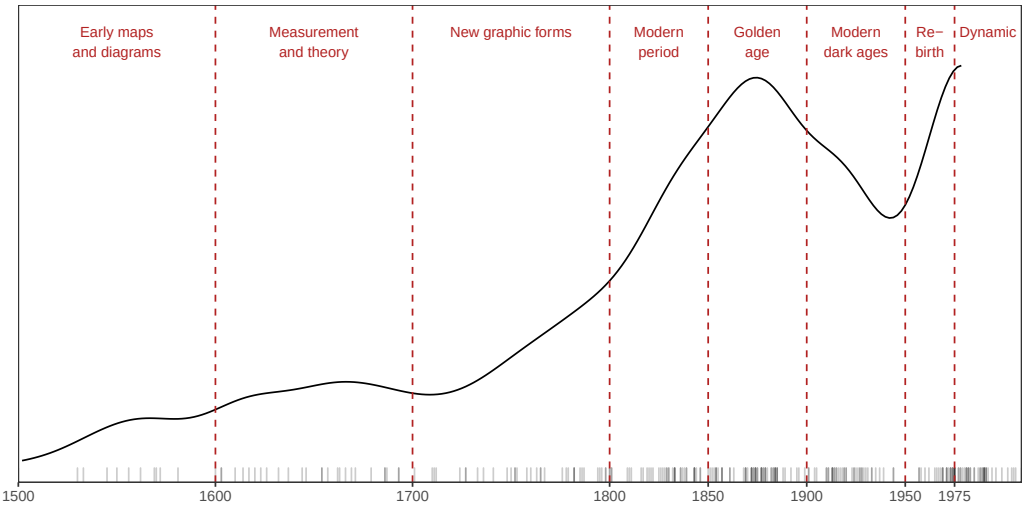


Figure 2.4: Milestones and Epochs in the history of data visualization according to Michael Friendly. Milestones are shown as a rug plot accompanied by a density estimate. Source: Own work, adapted from Friendly (2008) using XML data from the corresponding Milestones Project website.

2.4 The Geography of Data

2.4.1 Effective Visualization

John W. Tukey highlighted the importance of good data visualization with the sentence "The simple graph has brought more information to the data analyst's mind than any other device." [146, page 457]. Visualization helps understanding and, more importantly, suggest research questions and allows effective communication [147].

The history of data visualization [148] can be divided into epochs (Fig. 2.4), corresponding to milestones in the development of visual thinking and visual representation, cataloged[68] by Michael Friendly and Daniel J. Denis[149]. According to this classification, we are currently in a new golden age of visualization, where the challenges are the representation of high-dimension and high-volume data interactively [150], as in the popular Gapminder[69] and Our World in Data[70] websites.

[68]The illustrated chronology of innovations is available at *http://www.datavis.ca/milestones/* at the time of writing.

[69]This Gapminder foundation website is available at *http://www.gapminder.org/* at the time of writing.

[70]This Max Roser web publication is available at *http://ourworldindata.org/* at the time of writing.

Beginning with the modern groundbreaking approach of Bertin from the late 1960s [151, 152], followed by the systematic study to the most effective strategy to graphically communicate statistical data developed by Cleveland [153, 154, 155, 156, 157, 158, 159, 160, 161], and the study of the aesthetic and conceptual properties of graphics as a vehicle to deliver meaningful information through the work of Tufte [162, 163, 164, 165, 166], the field of data visualization offers a rich catalog of techniques [167] to synthesize complex datasets into meaningful graphical vehicles to convey information —and in some cases, disinformation [168]—.

More recently, with the capacity of computers to store, manage and display large amounts of data, a more procedural approach to representing data has emerged with the work of Friendly [169] and Wilkinson [170], further developed by Wickham [171], with the focus shifting to programmatically defining the mapping of values and concepts to graphical elements, which are then generatively produced through software, such as the JavaScript library D3.js[71], the R package ggplot2 [172] or the Seaborn[72] Python library.

2.4.2 Visualizing Cartographic Data

Effective visual representation allows the extraction of information from complex data sets. This is specially critical when data includes a location component, as in the case of cartographic data, where representation is constrained to the spatial domain

The representation of geographic data [173, 174, 175, 176] has specific challenges because it deals with different kinds of information [177] depending on the nature of the data, making the visualization of historical [178] and conceptual [179] information challenging, because many potential pitfalls —intentional or unintentional— must be avoided [180].

Despite these limitations, modern approaches use data visualization as a form of storytelling [181], and because of their familiarity in our life experience [182], maps can be powerful communication tools to tell stories[73], as evidenced by the ESRI Story Maps[74] website. Likewise, abstract maps presented in data space instead of geographic space are a powerful vehicle to deliver information, as for

[71]The Data Driven Documents (D3) JavaScript library is available at *http://d3js.org/* at the time of writing.

[72]The matplotlib-based Seaborn Python data visualization library is available at *http://seaborn.pydata.org/* at the time of writing.

[73]Beyond the influence of Robert Louis Stevenson's "Treasure Island" novel in our collective imagination.

[74]ESRI Story Maps can be found at *http://storymaps.arcgis.com/* at the time of writing.

example in the Newsmap[75] news aggregator service.

Beyond the historical utility of maps as visualization tools [183, 184], they can be appreciated as an art form for their aesthetic properties [185, 186, 187], even when disconnected from any representation of the physical world [179, 188].

2.4.3 New Cartographies

The availability of many new sources of data has allowed a new data-driven geography [95], transitioning from the traditional data-scarce scenario to a new data-rich environment, allowing the development of new cartographies. This approach is still in its infancy, and as maps of old said *"hic sunt dracones"*[76] in uncharted territories[77], its practitioners are still learning as they are doing.

Because of the shifting nature of data and its complexity, the natural medium of expression for the results of these experience is the web, as it offers unmatched interactivity to produce dynamic maps that can be easily distributed worldwide. However, some temporal phenomena is sometimes presented in non-interactive online video format.

Some initiatives are focused on exploring a multifaceted approach to data generated by cities such as San Francisco[78], London[79] or Barcelona[80] itself, while other cartographers disseminate their results through their personal websites, as is the case of James Cheshire[81], Eric Fischer[82], Anita Graser[83], John Nelson[84], William Rankin[85], Herwig Scherabon[86], or Andy Woodruff[87], some of which have been

[75] The Newsmap service is available at *http://newsmap.jp/* at the time of writing (Flash required).
[76] "Here be dragons"
[77] As seen in "The Enchanting Sea Monsters on Medieval Maps" exhibition in the Smithsonian Institution, available at *http://www.smithsonianmag.com/science-nature/the-enchanting-sea-monsters-on-medieval-maps-1805646/* at the time of writing.
[78] Maps are aggregated by the sfgeo project, available at *http://sfgeo.org/* at the time of writing.
[79] The Mapping London project is available at *http://mappinglondon.co.uk/* at the time of writing.
[80] The 300.000 Km/s team website is available at *http://300000kms.net/* at the time of writing.
[81] James Cheshire website is available at *http://spatial.ly/* at the time of writing.
[82] The GitHub profile of this data artist and developer is available at *http://github.com/ericfischer* at the time of writing.
[83] Anita Graser research on temporal data visualization (focused trajectories) is available at *http://anitagraser.com/* at the time of writing.
[84] John Nelson's Adventures In Mapping website is available at *http://adventuresinmapping.com/* at the time of writing.
[85] William Rankin's Radical Cartography website is available at *http://www.radicalcartography.net/* at the time of writing.
[86] The personal page of this Austrian designer is available at *http://scherabon.com* at the time of writing.
[87] Andy Woodruff website is available at *http://andywoodruff.com/* at the time of writing.

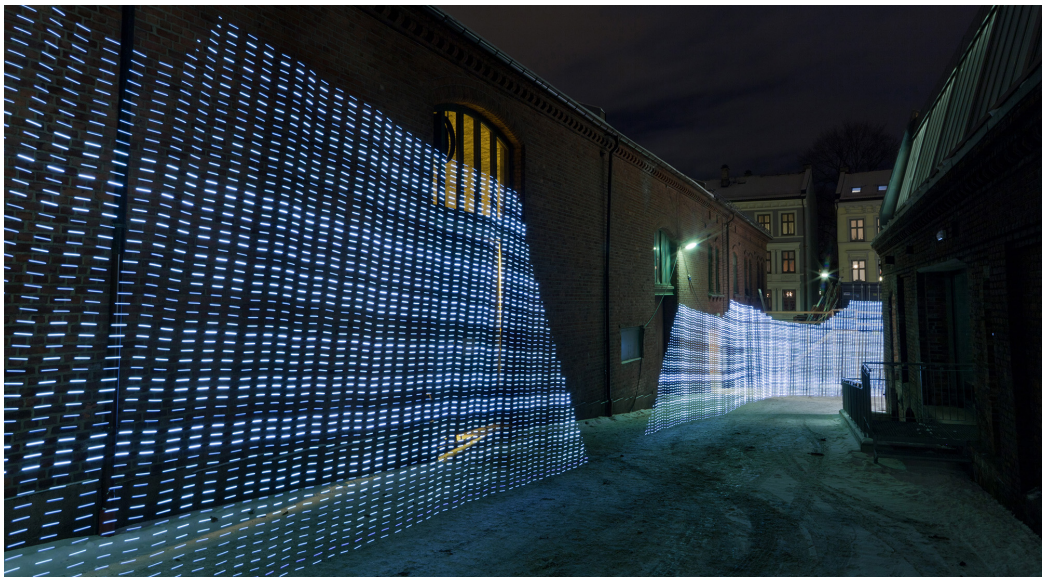


Figure 2.5: Image from "Immaterials: Light painting WiFi", revealing wireless network intensity in the streets of Oslo through long exposure photography. Source: Timo Arnall personal website (www.elasticspace.com).

featured in the Strange Maps blog[88].

Despite the prevalence of online mapping, some initiatives have used the urban environment itself as the support medium. That was the case of "Immaterials: Light painting WiFi", a project[89] by Timo Arnall[90], that used long-exposure photography of a a rod with 80 attached lights around Grünerløkka area of Oslo, Norway[91]. The lights responded to the Received Signal Strength Indicator (RSSI) of the WiFi networks, revealing a snapshot of the cross-section of the invisible wireless network field strength[92] (Fig. 2.5).

[88]The Frank Jacobs' Strange Maps blog is available at *http://bigthink.com/blogs/strange-maps* at the time of writing.

[89]It was exhibited at Arts Santa Mònica (Barcelona, October 14, 2011 – March 4, 2012) in the "Invisible Fields: Geographies of Radio Waves" exhibition curated by José Luis de Vicente and Honor Harger, catalog available at *http://www.lighthouse.org.uk/programme/invisible-fields* at the time of writing.

[90]Timo Arnall's website is available at *http://www.elasticspace.com/* at the time of writing.

[91]A Flickr album of the process is available at *https://www.flickr.com/photos/timo/albums/72157626020532597* at the time of writing.

[92]Also available in video format at Vimeo at *http://vimeo.com/20412632* and YouTube at *http://youtu.be/cxdjfOkPu-E* at the time of writing.

2.4.4 Mapping the City

The modern cartographic approach at the urban scale began with the physical representation of public spaces both inside and outside buildings (Fig. 2.6), recognizing the importance of the ground floor as the central element of the urban space [6] as well as the transition between public and private space [189].

This cartographic representation can be traced to the Map of Rome[93] by the architect and surveyor Giambattista Nolli (Fig. 2.6a) in a series of twelve engraved copper plates, which influenced the plan of Barcelona by Miquel Garriga i Roca (Fig. 2.6b), known as *els Quarterons Garriga i Roca*[94], consisting in 119 plans of the mid-nineteenth century Barcelona.

In some sense, the Google Indoor Maps[95] service could be considered a contemporary version of this approach (Fig. 2.6c), adapted to mobile devices with pervasive connections, where data allows taking this concept beyond the ground floor and at a global scale. Furthermore, improvements in 3D technology are enabling life-like representation of the built environment [190, 191] which should benefit from data stored in future 3D and 4D cadastres [192, 193, 194].

Beyond physical representation, one of the earliest applications of data-driven spatial analysis is the John Snow[96] analysis on the 1854 Broad Street cholera outbreak [195] in the Soho district of London (Fig. 2.7), which was based on the compilation of tabular data (Fig. 2.7a) which was spatially aggregated to reveal clusters around a water pump (Fig. 2.7b).

However, the complexity of the cities of today requires mapping urban phenomena from multiple points of view as a research method in multiple scales [196], sometimes requiring approaches beyond mapping, including other non-cartographic visualization strategies [197].

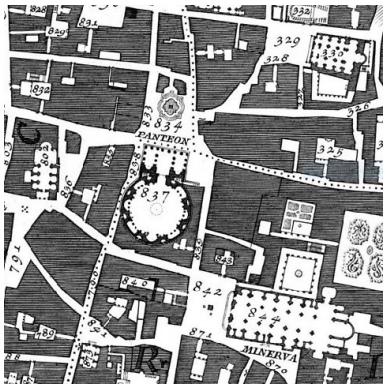
Effective communication needs adequate visualization strategies, specially when the size of data is large and has a spatial component (Fig. 2.8). The Austrian

[93] An interactive online version of the map can be explored at The Nolli Map website, a project by the University of Oregon available at *http://nolli.uoregon.edu/* at the time of writing (flash required).

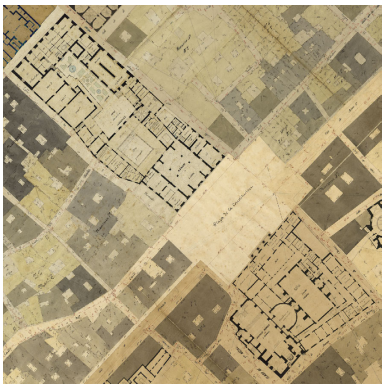
[94] At the time of writing, the map can be explored interactively at *http://darreramirada.ajuntament.barcelona.cat* and the cartographic material is downloadable at the 1:1250 scale in the CartoBCN website at *http://w20.bcn.cat/cartobcn/*

[95] More information on Google Indoor Maps is available at *http://www.google.com/maps/about/partners/indoormaps/* at the time of writing.

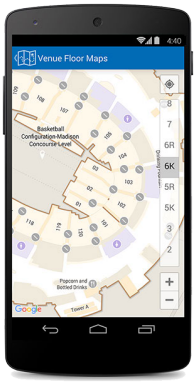
[96] The John Snow Archive and Research Companion at the Michigan State University is available at *http://matrix.msu.edu/~johnsnow/* at the time of writing.



(a) Giambattista Nolli Map of 18th century Rome (1748)



(b) Quarterons Garriga Roca of 19th century Barcelona (circa 1860)

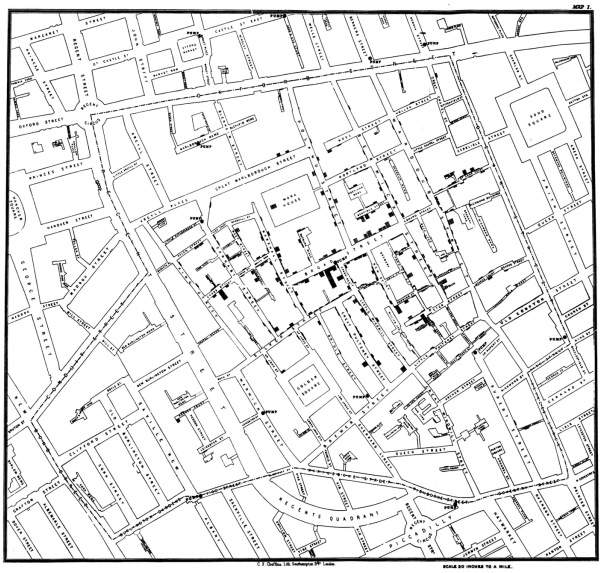


(c) Google Indoor Maps (2018)

Figure 2.6: Maps of public space inside and outside buildings in the 18th century (Rome), 19th century (Barcelona), and 21st century (Madison Square Garden, NY). Sources: The Nolli Map Website at the University of Oregon (Rome), CartoBCN (Barcelona), Google Indoor Maps website (New York City).

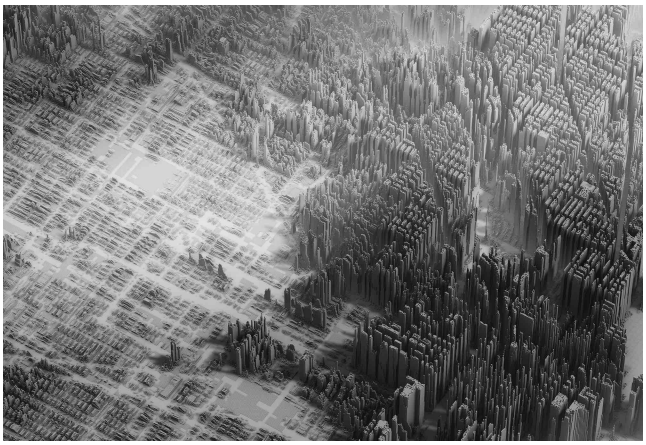
Sub-Districts.	Popula-tion in 1851.	Deaths from Cholera in the week end-ing 5th August.	Water Supply.				
			Southwark & Vauxhall.	Lambeth.	Pump-wells.	River Thames, ditches, etc.	Unascertained.
St. Saviour, Southwark	19,709	26	24	—	—	2	—
St. Olave, Southwark	8,015	19	15	—	—	2	2
St. John, Horsleydown	11,360	18	17	—	—	1	—
St. James, Bermondsey	18,899	29	23	—	—	6	—
St. Mary Magdalen	13,934	20	19	—	—	1	—
Leather Market	15,295	23	23	—	—	—	—
Rotherhithe	17,805	26	17	—	—	9	—
Battersea	10,560	13	10	—	1	2	—
Wandsworth	9,611	2	—	—	—	2	—
Putney	5,280	1	—	—	1	—	—
Camberwell	17,742	19	19	—	—	—	—
Peckham	19,444	4	4	—	—	—	—
Christchurch, Southwk.	16,022	3	2	1	—	—	—
Kent Road	18,126	8	7	1	—	—	—
Borough Road	15,863	21	20	1	—	—	—
London Road	17,836	9	5	4	—	—	—
Trinity, Newington	20,922	14	14	—	—	—	—
St. Peter, Walworth	29,861	20	20	—	—	—	—
St. Mary, Newington	14,033	5	5	—	—	—	—
Waterloo Road (1st)	14,088	5	5	—	—	—	—
Waterloo Road (2nd)	18,348	5	5	—	—	—	—
Lambeth Church (1st)	18,409	5	2	1	—	1	1
Lambeth Church (2nd)	26,784	10	7	2	—	—	1
Kennington (1st)	24,261	11	9	1	1	—	—
Kennington (2nd)	18,848	3	3	—	—	—	—
Brixton	14,610	1	—	1	—	—	—
Clapham	16,290	5	4	—	1	—	—
St. George, Camberwell	15,849	9	7	2	—	—	—
Norwood	3,977	—	—	—	—	—	—
Streatham	9,023	—	—	—	—	—	—
Dulwich	1,632	—	—	—	—	—	—
Sydenham	4,501	—	—	—	—	—	—
	486,936	334	286	14	4	26	4

(a) Table relating deaths and water supply source

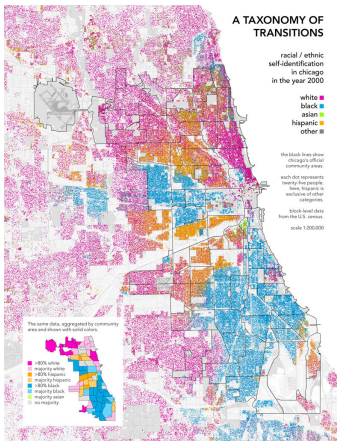


(b) Map showing clusters of cholera cases around the pump at the intersection of Broad Street and Cambridge Street

Figure 2.7: John Snow analysis on the 1854 Broad Street cholera outbreak in the Soho district of London, England. Sources: Google Books (table) and The John Snow Archive and Research Companion at the Michigan State University (map).



(a) Detail of the Chicago panel in the Income Inequality exhibition (Scherabon, 2016)



(b) Self-identification map of Chicago (Rankin, 2009)

Figure 2.8: Maps derived from US Census Bureau data to highlight social issues. Sources: Herwig Scherabon website (Los Angles) and William Rankin website (Chicago).

designer[97] Herwig Scherabon produced a series of large prints[98] of Los Angeles and Chicago, to reveal the spatial pattern of income inequality[99] in those cities, visualizing the median household income in a fine grid as the extruded height of its cells (Fig. 2.8a), using data from the US Census Bureau[100]. With the same source, the historian and cartographer William Rankin used a dot mapping technique[101] to show the clusters and fuzzy boundaries of racial / ethnic self-identification of Chicago, compared to the 1920 delimitation of Chicago's official community areas (Fig. 2.8b).

Another approach, which can be traced back to Kevin Lynch's study on how users understand their surroundings [26], tries to extend the concept of mental maps of the city using modern technology [198]. In this line of research, social media can be used as a tool for indirectly mapping specialized clusters in the city. In the case of New York City, Twitter and Foursquare data has been used to

[97] According to his website, "he studied architecture but made a u-turn towards graphic design".

[98] The work won the 2016 Information is Beautiful Award in the Student Category, according to the official website available at *http://www.informationisbeautifulawards.com/news/196-2016-winners-special-awards* at the time of writing.

[99] The Income Inequality exhibition is showcased at *http://www.scherabon.com/income-inequality/* at the time of writing.

[100] The United States Census Bureau provides downloadable data through its website at *http://www.census.gov/* at the time of writing.

[101] The Chicago map "A Taxonomy of Transitions" is available at *http://www.radicalcartography.net/chicagodots.html* at the time of writing.

determine digital neighborhoods [199], and geo-tagged pictures from Flickr has been analyzed to visualize the spatial distribution of residents, domestic tourists and foreign tourists [11].

On the temporal dimension, other valuable research focuses in the publication of historic maps, scanning and georeferencing old cartographic documents and making them available online. The Old Maps Online[102] project hosts at the time of writing 400,000 maps from contributing archives and libraries. A notable example is Charles Booth's London website[103], which allows viewing 41 digitized notebooks and the London poverty maps from the Inquiry into Life and Labour in London (1886-1903).

Finally, beyond using maps to understand the city, maps can be used as tools to promote public participation through online mapping [200], where users become active producers and not only receptors.

2.4.5 Beyond the Urban Scale

Beyond our understanding of cities, data can also provide insights in broader —metropolitan, continental or global— scales. In fact, some issues that make data unsuitable to provide answers at the urban scale are mitigated when the number of samples increases (as the signal to noise ratio improves) or the geographic extent is larger (because the precision of the locations becomes less important).

One of the most original research granted by the availability of data from social networks that illustrates the possibilities of this approach is the study of the spatial distribution of sports fans (Fig. 2.9). The Atlantic produced a map[104] showing the favorite National Football League (NFL) team per county (Fig. 2.9a), using data from the Facebook Data Science team[105]. To produce the map, the number of "likes" received by each official team page was counted per county, and each county was assigned the team that received the plurality[106] of votes from its Facebook users.

In the same vein, even with greater detail, The New York times published two maps

[102] Old Maps Online is available at *http://www.oldmapsonline.org/* at the time of writing.

[103] The Charles Booth's London is hosted by The London School of Economics and Political Science available at *http://booth.lse.ac.uk* at the time of writing.

[104] The Atlantic published the feature "The Geography of NFL Fandom" on September 5, 2014, available at *http://www.theatlantic.com/technology/archive/2014/09/the-geography-of-nfl-fandom/379729/* at the time of writing.

[105] The original research on NFL fans on Facebook is available at *http://www.facebook.com/notes/facebook-data-science/nfl-fans-on-facebook/10151298370823859/* at the time of writing.

[106] "Relative majority" in the United Kingdom.

where data was aggregated into zip codes instead of counties[107], using aggregated data from from Facebook's advertising platform, and applying an algorithm to smooth the noise in the data and fill gaps when data was not available[108]. The results show the distribution of the Major League Baseball (MLB) team preferences[109] (Fig. 2.9b) and the corresponding National Basketball Association (NBA) distribution of supporters[110] (Fig. 2.9c).

Other mapping projects "augment" traditional maps (Fig. 2.10), and do not represent observable information on features of the physical world [201]. One of such projects is Radio Garden[111], which allow exploring, discovering and experiencing the sonic landscape of the world through the locations of live radio stations located on the globe (Fig. 2.10a).

However, the most well-known example is the augmented reality (AR) game Pokémon GO[112], based on the popular Nintendo video game series and developed by Niantic[113], which has spawned a myriad of mapping projects to visualize the locations of the most sought-after creatures. While most of these maps are quickly taken down by the owners of the game on the grounds of intellectual property infringement, some of these maps avoid legal trouble revealing only the locations of "Gyms" and "PokéStops" (Fig. 2.10b).

2.4.6 Including the Temporal Dimension

The massive data sets of satellite imagery have been traditionally impractical to store and manipulate on desktop computers. However, since the area of interest is generally a small subset of data, advances in transmission speeds have allowed a client-server infrastructure where the server stores and manages data while the client requests and receives only a small portion. This approach is used by

[107] The original research on Baseball fans on Facebook is available at *http://www.facebook.com/notes/facebook-data-team/baseball-on-facebook/10150142265858859* at the time of writing.

[108] The process is detailed in the post "Answering Readers' Questions About the Baseball Maps" by Tom Giratikanon and Kevin Quealy published on April 25, 2014 in The New York Times, available online at *http://www.nytimes.com/2014/04/26/upshot/answering-readers-questions-about-the-baseball-maps.html* at the time of writing

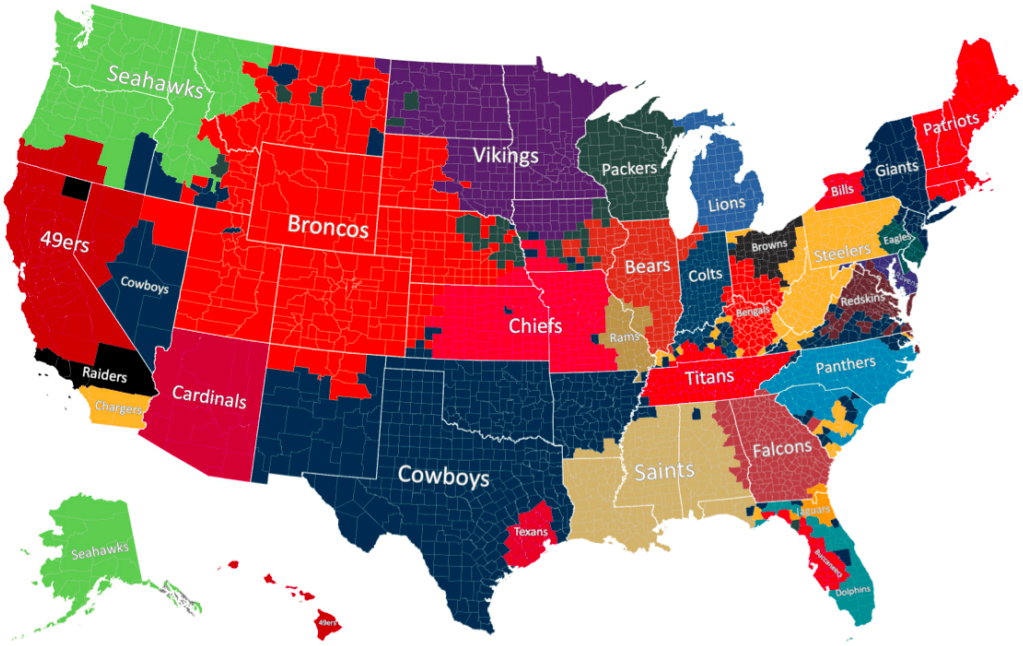
[109] The New York Times published the feature "Up Close on Baseball's Borders" on April 24, 2014, available at *http://www.nytimes.com/interactive/2014/04/23/upshot/24-upshot-baseball.html* at the time of writing.

[110] The New York Times published the feature "Which Team Do You Cheer For? An N.B.A. Fan Map" on October 19, 2014, available at *http://www.nytimes.com/interactive/2014/05/12/upshot/12-upshot-nba-basketball.html* at the time of writing.

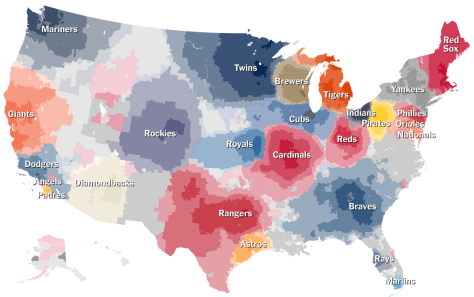
[111] The Radio Garden project is available at *http://radio.garden/* at the time of writing.

[112] The Pokémon GO website is available at *http://www.pokemongo.com/* at the time of writing.

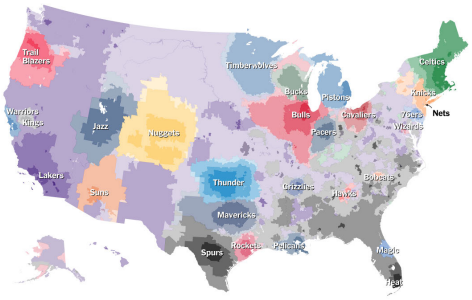
[113] The Niantic website is available at *http://www.nianticlabs.com/* at the time of writing.



(a) Map of NFL fans across the USA per county, according to Facebook data. Source: "The Geography of NFL Fandom" by Robinson Meyer published online in The Atlantic on September 5, 2014.

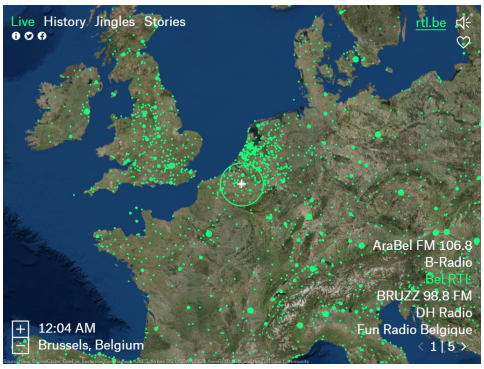


(b) Map of Baseball (MLB) team preferences in the USA per zip code, according to Facebook data. Source: "Up Close on Baseball's Borders" by Tom Giratikanon, Josh Katz, David Leonhardt and Kevin Quealy published online in The Upshot section of The New York Times on April 24, 2014.

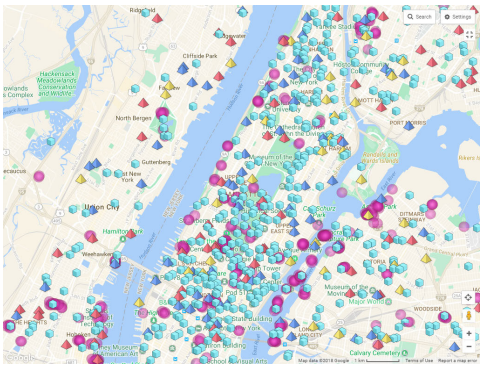


(c) Map of Basketball (NBA) team preferences in the USA per zip code, according to Facebook data. Source: "Which Team Do You Cheer For? An N.B.A. Fan Map" by Tom Giratikanon, Josh Katz, David Leonhardt and Kevin Quealy published online in The Upshot section of The New York Times on October 19, 2014.

Figure 2.9: Maps of NFL (per county, top) and MLB and NBA (per zip code, bottom) team preferences across the USA, according to Facebook data.



(a) Radio Garden Project web page showing the locations of available live radio stations in an interactive globe. Source: Screenshot of radio.garden taken on February 11, 2018.



(b) Pokémon GO map of Manhattan (NY) showing game element locations. Source: Screenshot of www.pokemongomap.info taken on February 12, 2018.

Figure 2.10: Screenshots of two websites that use maps to provide augmented capabilities

Google Maps[114], Bing Maps[115]or OpenStreetMap[116], or in the case of GIS users by the Web Map Service (WMS) protocol.

While the served satellite imagery is generally the most up-to-date, the Google Earth Engine Timelapse[117], powered by technology developed by the Carnegie Mellon University CREATE Lab[118], allows exploring 33 cloud-free mosaics corresponding to every year of the 1984–2016 period, using data from 5 different satellites[119].

This service uses a standard user interface that allows panning and zooming to anywhere on Earth, but includes a slider that allows selecting any year within the covered time period (Fig. 2.11), or view the temporal snapshots as a time-accelerated animation. These features allow visualizing phenomena such as the

[114]The Google Maps website is available at *http://maps.google.com/* at the time of writing.

[115]The Microsoft Bing Maps website is available at *http://www.bing.com/maps* at the time of writing.

[116]The OpenStreetMap community effort is available at *http://www.openstreetmap.org/* at the time of writing.

[117]The Google Earth Engine Timelapse multi-petabyte catalog of satellite imagery website is avaliable at *http://earthengine.google.com/timelapse/* at the time of writing.

[118]The Community Robotics, Education and Technology Empowerment Lab (CREATE Lab) website is available at *http://cmucreatelab.org/* at the time of writing.

[119]A YouTube playlist with some selected highlights is available at *http://www.youtube.com/playlist?list=PLWw80tqUZ5J_T8EKLKEWYd_NcFPiq9zTN* at the time of writing.

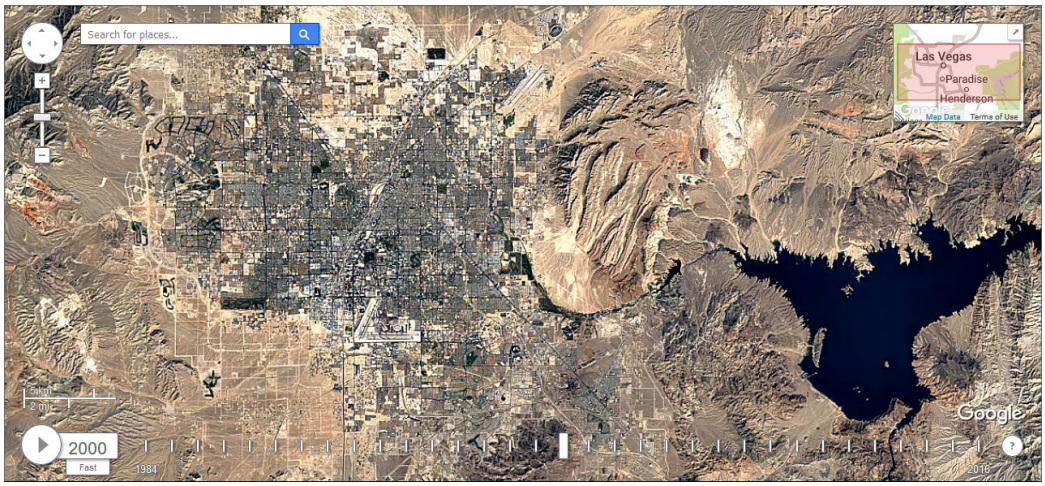


Figure 2.11: Timelapse feature of the Google Earth Engine that allows exploring worldwide satellite imagery in the 1984–2016 period, centered on Las Vegas (Nevada), where the increased urbanization and the shrinking of Lake Mead can be appreciated. Source: Screenshot of earthengine.google.com taken on February 10, 2018.

urban sprawl and deforestation[120] in space and time.

Using the same horizontal timeline user interface, the World Population History project[121] visualizes population as one dot per one million people for each year, from year 1 CE to the projected population in year 2050[122], in addition to historical milestones in human history placed in the timeline, classified into five different themes[123] (Fig. 2.12).

The non-interactive video format is also used to visualize events along time, from historical scales spanning centuries —to map of 14,238 events corresponding to geotagged Wikipedia articles that contain time information[124]— to a single day, as in the case of the video of the illegal download locations of scientific papers

[120] Time Magazine featured these data over nearly three decades of satellite photography in the special feature "TIME and Space" by Jeffrey Kluger, available at *http://world.time.com/timelapse/* at the time of writing.

[121] The World Population History project website is available at *http://worldpopulationhistory.org/* at the time of writing.

[122] A video of the animated data is available on Vimeo at *http://vimeo.com/130468614* at the time of writing.

[123] The featured themes are 1) food and agriculture, 2) health, 3) people and society, 4) environment, and 5) science and technology.

[124] The video is available on YouTube at *http://youtu.be/Ee8A8xLsG0w* and on Vimeo at *http://vimeo.com/19088241* at the time of writing.

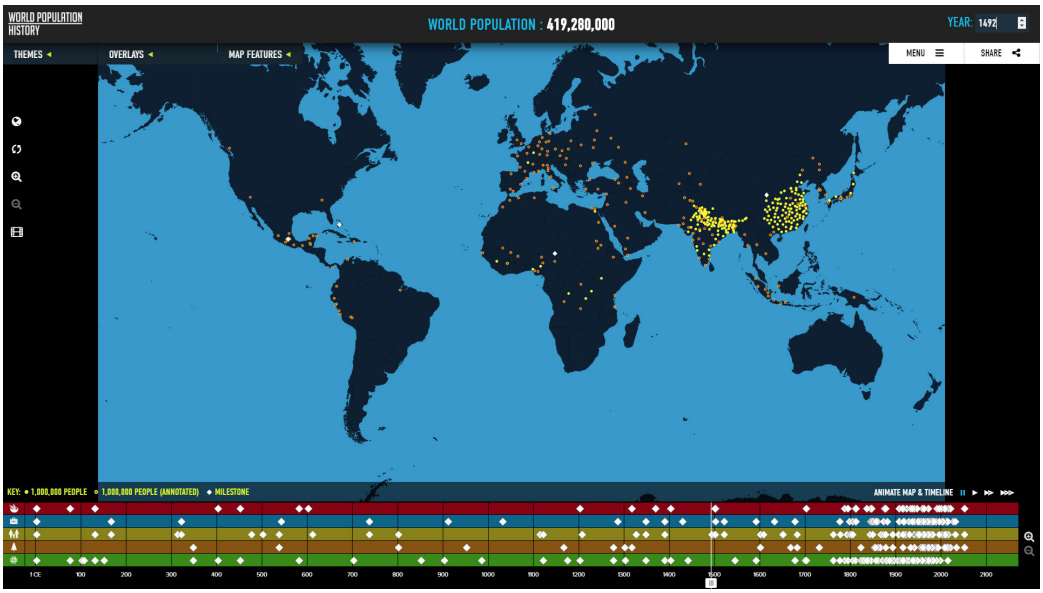


Figure 2.12: Map of the World Population History project that allows exploring the evolution of human population in the 1–2050 (projected) period from historical, environmental, social and political perspectives. Source: Screenshot taken on February 11, 2018 of the worldpopulationhistory.org website.

across the world[125] [111].

2.5 Tracking User Behavior

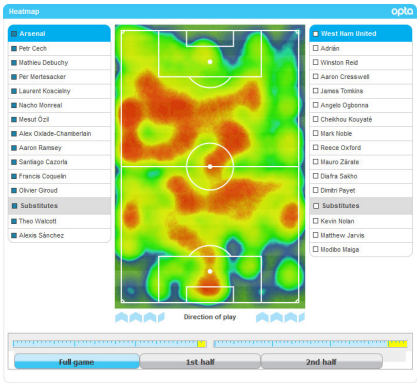
2.5.1 Location Tracking Technologies and their Limitations

The strategy[126] of recording human activity on public spaces to support planning decisions [202] can be traced to Pushkarev [203] and Whyte [204]. In particular the press of the time noticed the efforts of William H. "Holly" Whyte to track human activity in New York City in the 1970s [205], using video recordings to quantify the use of users of the public space, trying to answer questions regarding who?, what?, where?, when? and how long?.

Advances in tracking technology [206] are now capable of conducting most of these activities [207] without human intervention using computer vision [208] and machine learning [209] technologies. With the advent of cheap and connected

[125]The video of the Sci-Hub activity on February 5, 2016 is available at *http://www.sciencemag.org/news/2016/04/whos-downloading-pirated-papers-everyone* at the time of writing.

[126]This approach was inspired by the work of Kevin A. Lynch in the 1960s.



(a) Heatmap of player locations in a soccer match (Arsenal Vs West Ham United). Source: Screenshot of the web page "Football Matchday Widget Showcase" retrieved on February 14, 2018 from *http://www.optasports.com/football-matchday-widget-showcase.aspx*.

(b) Screenshot of an interactive map showing Kobe Bryant's scored and missed shots during all his basketball career, produced using data compiled by NBA (stats.nba.com). Source: "Every shot Kobe Bryant ever took. All 30,699 of them" by Joe Fox, Ryan Menezes and Armand Emamdjomeh published online in Los Angeles Times on April 14, 2016.

Figure 2.13: Visualization of tracking data in the field of sports analytics, for soccer (left) and basketball (right).

video cameras, conducting these studies during longer periods of time or in real-time [210] and in larger areas is becoming technically and economically feasible, using dedicated software [211].

Optical tracking technologies have to account for perceptive transformation form the camera point of view [212], occlusions from architectural elements, vehicles or other individuals, and changes in lighting [213]. Equivalent examples in sports (Fig. 2.13) such as soccer (Fig. 2.13a) and basketball[127] (Fig. 2.13b) do not suffer from these issues, as the space of the playing field is well defined, the number of tracked individuals is small and easily identified by their jersey numbers, and the duration of the match is relatively short.

Some approaches to improve tracking include using multiple cameras [214], or in the case of large groups, using markers to conduct experiments [215], or tracking group flow [216] instead of individuals. Privacy concerts can be mitigated using specialized hardware such as Kinect devices [217] or infrared cameras [218].

Data extracted from tacking is used to obtain information on the behavior (e.g. walking speed, position preferences, interpersonal distances) of multiple user

[127] "Every shot Kobe Bryant ever took. All 30,699 of them" by Joe Fox, Ryan Menezes and Armand Emamdjomeh was published online in Los Angeles Times on April 14, 2016 and is available at *http://graphics.latimes.com/kobe-every-shot-ever/* at the time of writing.

profiles (e.g. age, gender, mobility, group size) [219], and to improve the prediction accuracy of flow-based [220] or agent-based [221] pedestrian movement models, or in the design of architectural spaces [222], in particular in the retail [223] and tourism [224] sectors.

Another approach is using device-assisted tracking technologies, which have traditionally used to track wildlife [225]. Advances in GPS technology [226] have allowed tracking volunteers using standalone GPS devices in historic cities [227], and long term tracking of members of volunteer families [228] or elderly people [226]. With the integration of GPS in mobile phones this approach has become more feasible [229], in particular with shared data from sports tracking applications [230, 231], despite accuracy concerns [232].

As the use of cellphone is nowadays almost pervasive in our society, tracking phone usage activity can allow a broader picture of the flows of the population [233]. This is only possible in collaboration with the network operators, which provide anonymized data for analysis to some research groups to study specific cities as in the cases of Tallinn [234], Milan [235, 236], Rome [237], Barcelona [238] or Singapore [239], and in some cases entire countries such as the UK [135] or the USA [240].

Beyond mobile phones, public transport tickets or credit card transactions has the potential to allow a better understanding of mobility patterns [241], transitioning from survey-based analysis [242] to using data from ticket validations [243].

On the other end of the spectrum, trading breadth for depth, recent research focuses on the study of the visual perception of urban [244] and architectural [245] spaces through gaze analysis, to identify the regions of interest and help wayfinding [246], and experimental designs in virtual spaces [247] or web games [248].

2.5.2 Mobile Technologies and Digital Footprints

On January 9, 2007 —less than four decades after the introduction of the first 8-bit microprocessor[128]—, Steve Jobs introduced[129] the iPhone at the Moscone Center in San Francisco as the blend of three different products with the catchphrase "this changes everything"[130]:

[128]The Intel 8008 was introduced in 1972.

[129]The announcement transcript is available at *http://thenextweb.com/apple/2015/09/09/genius-annotated-with-genius/* at the time of writing.

[130]But "what we didn't know was that the everything was *us*", according to Nagel and Reiner in their short essay "Embedded beings: how we blended our minds with our devices" available at *http://aeon.co/ideas/embedded-beings-how-we-blended-our-minds-with-our-devices* at the time of writing.

1. A widescreen iPod with touch controls
2. A revolutionary mobile phone
3. A breakthrough Internet communications device

While not the first smartphone[131], the adoption of this class of devices has since been unprecedented[132], increasing in computing power while decreasing energy consumption because of Moore's law[133] [249], and progressively incorporating technologies that have displaced other devices with specialized functionality, such as standalone GPS devices, media players or digital cameras. Remarkably, on 2017 half of all pictures uploaded to Flickr were taken with smartphones, and Apple devices were the most popular, ahead than Canon and Nikon[134].

Today, almost everybody[135] carries in his or her pocket a powerful battery-powered multimedia computer with a user-friendly touch-based interface, coupled with a high quality digital camera[136] and a fairly accurate GPS, connected to the Internet through a pervasive high-speed wireless network, with access to an almost unlimited cloud storage capacity, and with a massive library of reasonably priced and easy to install[137] applications.

Services under the umbrella term of "Web 2.0", where users became both producers and consumers of content, were maturing since the beginning of the century; in particular, social media was strongly growing in popularity around the time the iPhone was announced, and began shifting part of its user base to mobile devices, which offered some features not available on desktop platforms, such as taking and sharing pictures and video content, using geolocation services and providing permanent untethered connections, in a form factor users had a closer and more personal[138] relationship with.

---

[131] During his presentation, Jobs mentioned the following devices: Motorola Q, BlackBerry, Palm Treo and Nokia E62.

[132] Worldwide smartphone user penetration is one third of the population in 2017, and projected to be 37% in 2020, according to the statista portal in the page "Smartphone user penetration as percentage of total global population from 2014 to 2020" available at *http://www.statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/* at the time of writing.

[133] Moore's law predicts that the transistor count of integrated circuits doubles roughly every two years.

[134] According to the December 7, 20017 post on the Flickr blog "Top Devices of 2017 on Flickr" available at *http://blog.flickr.net/en/2017/12/07/top-devices-of-2017/* at the time of writing.

[135] Spain's smartphone user penetration in 2017 was more than two thirds of the population (66.8%), according to Newzoo in the page "Top Countries by Smartphone Penetration & Users" available at *http://newzoo.com/insights/rankings/top-50-countries-by-smartphone-penetration-and-users/* at the time of writing.

[136] Smartphone cameras have almost displaced pocket digital cameras, and are approaching the quality of consumer single-lens reflex (SLR) devices.

[137] And automatically updated over the Internet.

[138] And arguably, even intimate.

2.5.3 Physical Traces

Intentional or unintentional physical traces left in objects or the environment can tell us about the behavior of those that produced them (Fig. 2.14). Although these marks can be almost unnoticeable individually, the repeated actions of a single subject or the collective actions of a group engaging in similar conducts can make those patterns visible[139] (Fig. 2.14a).

In the natural world, predators and scavengers (including human hunters) have learnt to exploit this as a source of information, and follow the trail of their prey while disguising their own tracks to avoid being spotted, but at the same time intentionally marking the environment with messages for their competitors.

Using the scientific method, some disciplines have traditionally relied on indirect sources of information to build a comprehensive body of knowledge in their disciplines; the field of paleontology has used fossils (markings, geological context, comparative anatomy) to study the evolution of organisms —as long as these organisms leave any geological trace—, while archaeology has been able to document the human biological and cultural evolution through the discovery of human remains and artifacts, and astronomers have studied the cosmic microwave background (CMB) to understand the universe.

Even in fiction, the explanation of a mystery form seemingly irrelevant clues using deduction has a became a classic plot device, from the 19th century stories of featuring Auguste Dupin (Edgar Allan Poe) or Sherlock Holmes (Sir Arthur Conan Doyle), through contemporary TV shows such as the franchises based on "CSI: Crime Scene Investigation"[140].

Morover, the Japanese appreciation of the aesthetic values and cultural significance of wear and tear in everyday objects manifests itself in the ancient art of *kintsugi*, where broken pottery objects are repaired with lacquer and elevated to art form (Fig. 2.14b).

At the core of all these very different examples described we can find a common process at work: a set of data in the form of the specific marks observed, and the interpretation of these data that allows the observer to make a deduction about the behavior of the agent that produced them, who does not need to participate or even be aware of this process beyond the initial action.

[139] As exemplified in the introduction of the 2011 dystopian science fiction novel "Wool" by Hugh Howey: "That always amazed him: how centuries of bare palms and shuffling feet could wear down solid steel. One molecule at a time, he supposed."

[140] Originally aired on CBS network 2000-2015.

(a) Handrail in the Barcelona School or Architecture, in the landing between two flights of stairs. Generations of students grabbing the horizontal section to make a 180 degree turn have chipped part of the white paint away, exposing the metal.

(b) Cracked pottery bowl repaired with lacquer and gold using the Japanese art of kintsugi. Source: Image uploaded by Wikipedia user Haragayato, under CC BY-SA 4.0.

Figure 2.14: Physical traces on objects revealing clues about the history of objects and their environment.

2.5.4 Digital Traces

We are living in an increasingly connected world, where an important part of our activity happens online, insomuch that the term *meatspace* has been coined in contrast with *cyberspace*. Hitherto, this paradigm shift has already had some effects[141] on human behavior such as the "Fear of missing out" (FoMO) where people experience anxiety when disconnected, and today the Internet[142] has become[143] *the* source of information, entertainment, communication, shopping and archival:

Information when we want to find out how a word is spelled, plan a vacation trip, choose a restaurant, find out the outcome of a soccer match, know the latest gossip, or find a new job, our first instinct is to perform a quick Internet check, replacing dictionaries, travel and restaurant guides or newspapers.

Entertainment when we need entertainment we increasingly turn to the Internet to watch movies or listen to music (legally or illegally), play online games or read a book. Even new forms of entertainment not possible before the Internet such as user-generated videos in YouTube are becoming to gain traction.

[141] Some of the consequences of new technologies on modern society are explored in the Black Mirror British TV series.

[142] Exemplified by its spelling with a capitalized initial.

[143] Among other things, some of which we cannot even imagine at present.

Communication when we want to announce the birth of a child or the death of a loved one, joke with friends, find a new partner or express our opinion, we turn to online communications.

Shopping when we want to buy anything from houses, cars or jewelry to toys, clothes or groceries, we check the prices on the Internet and increasingly shop diving into the vast catalogs of online shops.

Archival When we want to store the product of our work or the memories of our life, we store it online vaults.

Akin to their physical counterparts, these online activities leave *digital traces*. As users become increasingly aware of this situation, they turn to constructing their own curated *online identity* [250], become increasingly concerned about their *digital death* [251], or begin using *ephemeral* communication services like SnapChat [252]. However, while some of these online activities are kept private, an increasingly number of online activities are meant to be shared, either with a restricted circle of people, or with the entire online community: participating in public forums, reviewing and rating products or services, enrolling in collaborative projects, maintaining a public-facing profile in a service or sharing content.

In parallel, corporations collect data at an unprecedented level [253] —generally agreed upon on the terms of service (TOS), to "enhance" the user experience—, as in the case of browsing habits (to increase the accuracy of search results), locations (to monitor the road network status in real-time) or viewing history (to improve suggestions based on the tastes of the user).

2.5.5 Desire Lines and Walking Preferences

A "desire path" or "desire line"[144] is a trail that can appear spontaneously outside the constructed pathways when pedestrians' foot-fall erode the unpaved surface (Fig. 2.15). The popularity of the path (measured as how much traffic it receives) can be observed indirectly from its degree of erosion or its width, depending on the mechanical properties of the ground surface. However, the formation of trails cannot be explained by a single mechanism but through the interaction of multiple factors, which can be classified into two main groups: the configuration of the environment and the behavior of the users.

Sometimes desire lines can be interpreted as a failure of the intended urban design, because users collective decide that the proposed path system is inadequate for their needs and leave the designated pathways. This contrast is specially striking

[144] Also known as "game trail", "social trail", "herd path", "cow path", "goat track", "pig trail" or "bootleg trail" according to the Wikipedia article available at *http://en.wikipedia.org/wiki/Desire_path* retrieved on December 21, 2016.

(a) Desire path formed by pedestrians taking a shortcut

(b) Paved walkway following a former desire path

Figure 2.15: Examples of desire paths in Diagonal Avenue (Barcelona). Pictures taken by the author on January 9, 2015.

in cities designed according to the principles of the International Style, such as Brasilia (Brasil), where an intricate network of informal pathways is overlaid onto the rigidly planned road structure (Fig. 2.16).

This "bottom-up" conflict with the "top-down" approach favored by architects can also be found in architectural design[145] [254], where the users are forced to adapt the designs of architects to suit their requirements. However, other disciplines have learnt to harness the *wisdom of the crowds*, such as the open source movement [255] to collaboratively produce software that in some cases rivals commercial products[146] [256]. Furthermore, in the field of usability assessment, it has been proposed to extend the notion of desire lines, studying the marks that reveal the user's intent as well as the inadequacy of the designs [257].

2.5.6 Digital Desire Lines

Similarly to how desire lines can indirectly reveal the pedestrian's walking preferences, the online behavior can be used to provide insights into the behavior of individuals on the Internet. This research tends to focus on social network communities [258] because of the abundance of data they generate [259], revealing patterns that were theorized [260] but not supported by data, and have been used to analyze the preferences in the selection of news outlets [261, 262], the

[145] Discussed in the 1997 six-episode TV series by the BBC "How Buildings Learn".

[146] A similar model can also be found in the Wikipedia project, available at *http://www.wikipedia.org/* at the time of writing.

Figure 2.16: Aerial view of Brasilia (Brasil), showing the planned infrastructure along with an informal network of trails. Source: Google Maps (2013).

propagation of rumors [263] or even the formation of love[147] [264].

On many occasions, these digital footprints have a location component that can be used to reveal spatial patterns, as in the case of Google web searches[148] related to the August 21, 2017 total eclipse over continental USA (Fig. 2.17), showing highly a concentrated interest along the path of totality[149], with up to ten times more web searches compared to the rest of the country [265].

Another research area is the spatial analysis of textual content, that allows linguistic studies at regional levels [266]. The unprecedented amount of data from Twitter allows researching trends and variation in slang that were impossible a few decades ago [267], for example the regional variation of the male colloquial vocatives[150] (Fig. 2.18) in the USA[151], such as "dude" (Fig. 2.18a) and "bro" (Fig. 2.18b).

[147] A six-part series of blog entries about relationships were posted by the Facebook Data Science team during February 2014, available at *http://www.facebook.com/data/posts/10152217010993415* at the time of writing.

[148] Data can be retrieved from Google Trends at *http://trends.google.com/trends/explore?geo=US&q=SolarEclipse* at the time of writing.

[149] The path of the solar eclipse can be followed at *http://wapo.st/eclipseroadtrip* at the time of writing.

[150] An interactive version was published on December 23, 2014 in Quartz titled "The dude map: How Americans refer to their bros" by Nikhil Sonnad available at *http://qz.com/316906/the-dude-map-how-american-men-refer-to-their-bros/* at the time of writing.

[151] Using a 8.9 billion word corpus of 890 million geocoded Tweets collected from across the contiguous United States between October 11, 2013 and November 22, 2014 by Diansheng Guo, according to *https://sites.google.com/site/wordmapperinfo/* available at the time of writing.

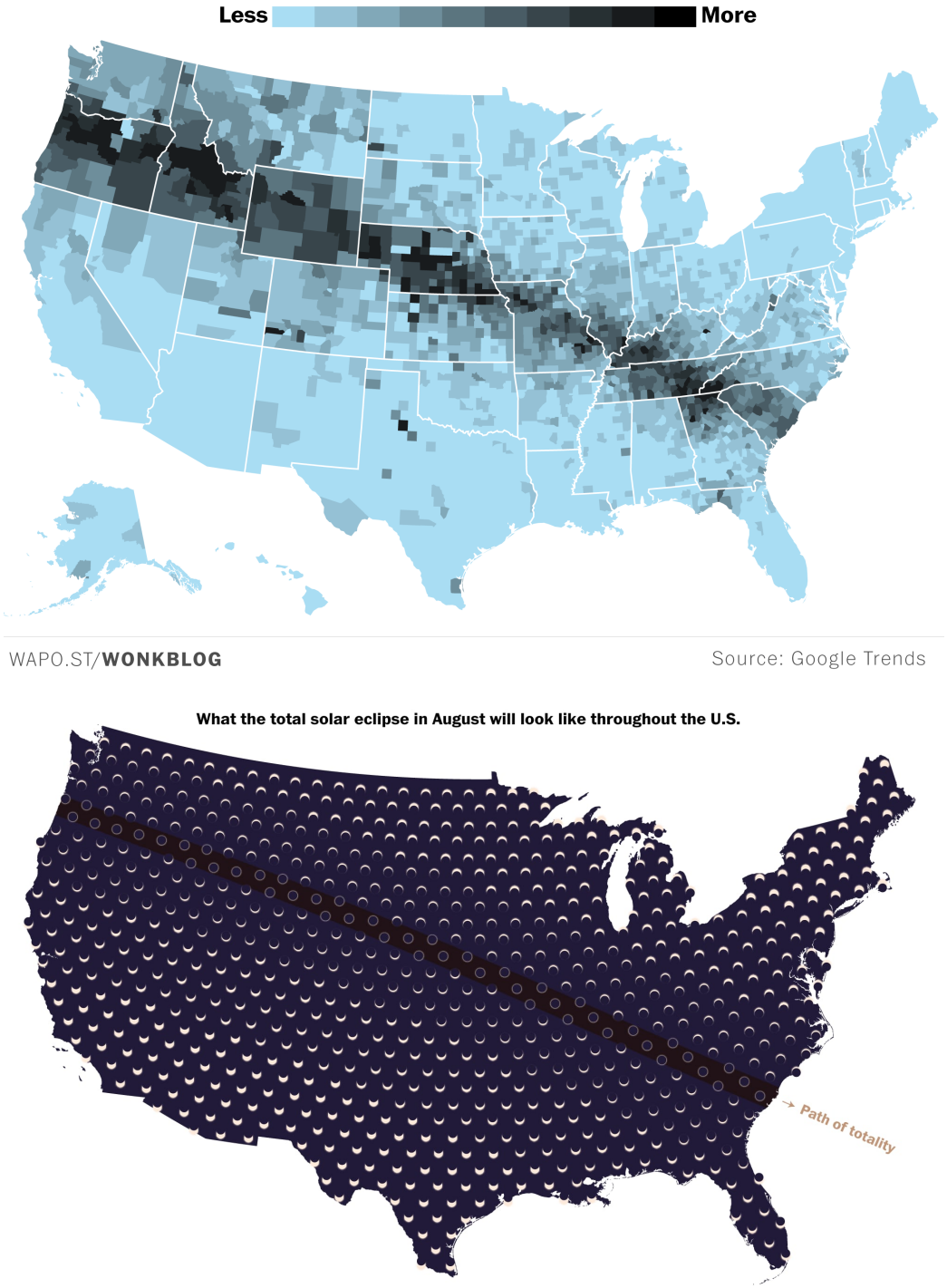
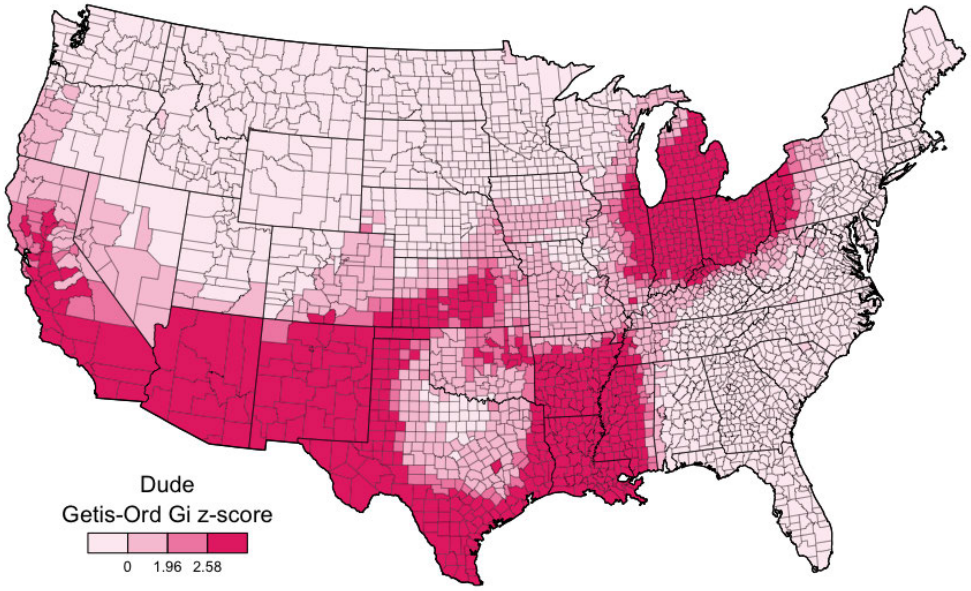
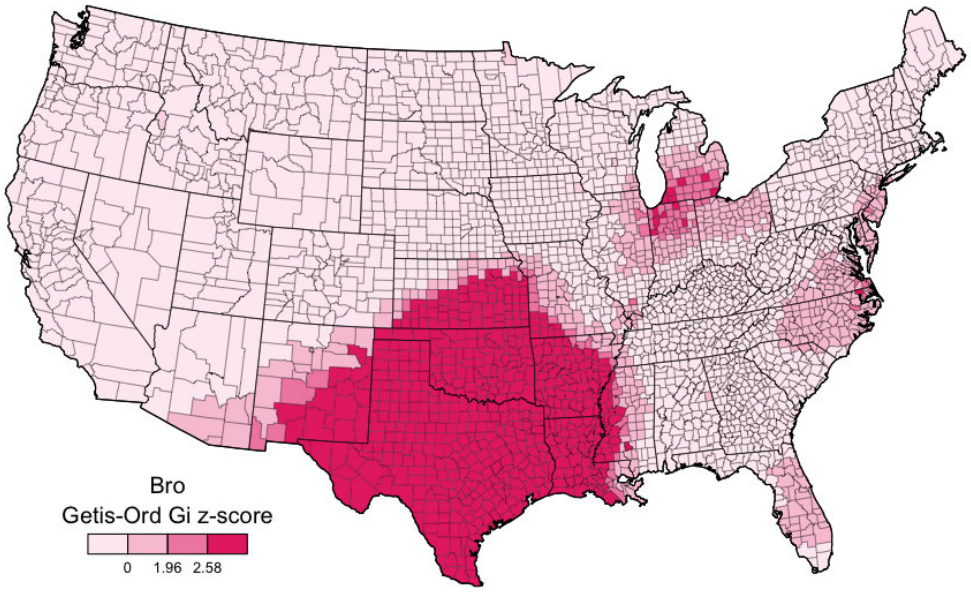


Figure 2.17: Web searches related to the August 21, 2017 total eclipse over continental USA, with the areas of higher interest following the path of totality. Source: "The path of the solar eclipse is already altering real-world behavior" by Christopher Ingraham published online in The Washington Post on August 1, 2017.



(a) Higher relative usage of ”dude”



(b) Higher relative usage of ”bro”

Figure 2.18: Regional hotspots of the male colloquial vocatives “dude” and “bro” according to geolocated Twitter word usage in US counties. Source: Jack Grieve research blog entry posted on December 28, 2014 using data collected by Diansheng Guo between October 2013 and November 2014.

The research of this dissertation focuses on assessing the suitability of using these novel approaches [91] to provide insight on some aspects of urban phenomena using Volunteered Geographical Information (VGI), using citizens taking part in a natural experiment as "sensors" [175] to study their preferences. In some sense, the results are expected to be biased by the preferences of the users, but these biases are themselves the focus of the research. Some of the results focusing on semantic data have already been published separately [143].

Chapter 3

Collecting Digital Traces

> "You know my methods, Watson. There was not one of them which I did not apply to the inquiry. And it ended by my discovering traces, but very different ones from those which I had expected."
>
> Sherlock Holmes in "The Adventure of the Crooked Man"

3.1 Searching and Collecting Online Content

3.1.1 Static content

Internet content was originally static, and websites were developed as a set of pages using the Hypertext Markup Language (HTML) specifications, either developing the code directly in a text editor[1], or using more user-friendly Graphical User Interface (GUI) front-ends to develop the pages visually[2]. Further developments were the introduction of Cascading Style Sheets (CSS) and JavaScript support in the late 1990s, which enabled richer designs and interactive content.

Originally, users accessed the pages through a Uniform Resource Locator (URL) stored in their computers as a bookmark[3], but as the number of web sites grew it

[1] Examples of popular text editors are EMACS or Vim in Unix systems.

[2] Examples of popular WYSIWYG (what you see is what you get) HTML editor software are Adobe DreamWeaver or Microsoft FrontPage (part of the Microsoft Office suite until 2003).

[3] Or directly typed in the browser address bar.

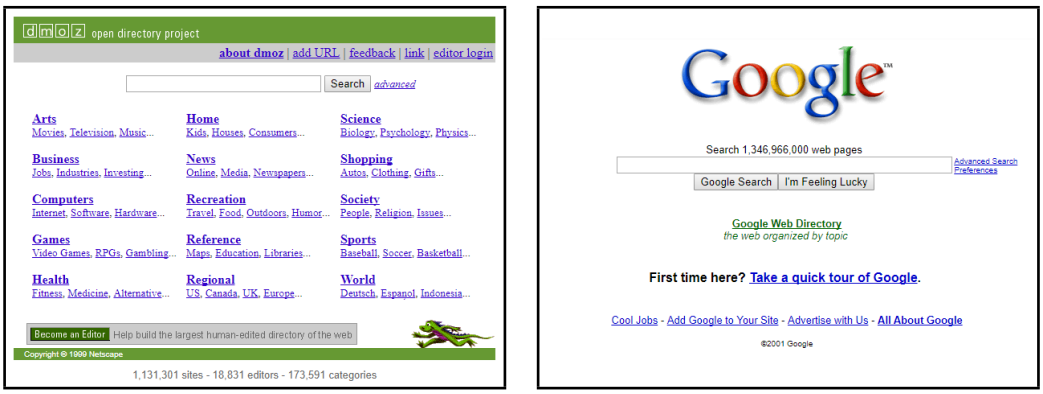


Figure 3.1: DMOZ (open directory project) screenshot from November 4, 1999 (left) and Google Search screenshot from June 1, 2001 (right). Source: The Internet Archive Wayback Machine.

became necessary to maintain directories such as DMOZ[4] (Fig. 3.1) that contained community-curated indices of web pages. However, at the beginning of the 21th century, the explosion and dynamism of Internet content production made this human-indexing impossible and it was therefore necessary to use software to crawl web sites and index their content automatically, with the appearance of emerging search engines such as Google[5] (Fig. 3.1) using its (at the time) brand new PageRank algorithm [268].

This automatic crawling and indexing was possible because the content was stored in HTML, which as a text-based format with a defined structure was relatively easy to parse without human intervention, and also because the content of these static websites did not change dramatically overnight, as they grew incrementally by adding, removing or modifying a few pages between crawls.

However, the rise of the Web 2.0 and dynamically generated web pages (which separated content from presentation, and generated HTML code on the fly upon request) introduced new challenges that required a paradigm shift.

[4]The DMOZ (Open Directory Project) archived website as it appeared in 1999 can be found at The Internet Archive Wayback Machine at *http://web.archive.org/web/19991104225904/http://www.dmoz.org/* at the time of writing.

[5]Google archived website as it appeared in 2001 can be found at The Internet Archive Wayback Machine at *http://web.archive.org/web/20010601003104/http://www.google.com/* at the time of writing.

3.1.2 Web Services

Introduced in 2000 by Roy Fielding [269], Representational State Transfer (REST) web services allow requesting resources from a remote computer system using a Uniform Resource Identifier (URI) and get a response in a defined format.

This approach is suitable to modern online applications which are continuously updated with updated content, such as a photo gallery or the comments in a blog entry, and also allows the service to tailor its content to each specific user profile.

However, since the returned content (web page or data) is only generated in the server when requested, it is not stored in advance (although it can be cached). Therefore, a mechanism had to be developed to allow exchanging data between web services, and at the time of writing most systems provide a Web Application Programming Interface (API) to query their database and/or request data.

3.1.3 Web APIs

Web APIs[6] allow a more flexible approach to the exchange of data at the expense of a more technically complex retrieval process, compared to a simple download of tabular data in Comma-Separated Values (CSV) format, readable in any spreadsheet application.

One difference is that only the requested data are returned, avoiding costly transfers over connections with slow bandwidth, which is crucial when querying the gigantic databases behind some web services. This implies that the request must be processed in the server side, before returning the response to the client.

Therefore, in contrast with the data sets prevalent in urban data (e.g. cadastral, demographic, land use, mobility), which are usually a snapshot of the recorded phenomena at a given time, most web APIs provide access to data that is continuously updated (added, modified or removed).

Another crucial difference is that the formats used are more flexible and extensible than a simple flat table, allowing nested structures of varying length, similar to some common data structures in most widely used programming languages (e.g. ragged arrays and key-value pairs), that need to be parsed, validated and checked for integrity to extract the stored information. The following are among the most common (and useful) formats returned in response to an API request:

- Hypertext Markup Language (HTML).
- JavaScript Object Notation (JSON).
- Extensible Markup Language (XML).

[6] An comprehensive web API catalog is maintained on the ProgrammableWeb website, available at *http://www.programmableweb.com/* at the time of writing.

Table 3.1: Summary of the main features of the APIs accessed in the research process for the present dissertation (in chronological order of exploration)

Source API	Authentication	Model	Response
Panoramio	No	Free (closed on 2016)	JSON
Flickr	Yes (token)	Free (with restrictions)	JSON, XML
Instagram	Yes (OAuth)	Restricted access	JSON
Twitter (Streaming)	Yes (OAuth)	Freemium	JSON
Twitter (REST)	Yes (OAuth)	Freemium	XML
Google Geocoding	Only paid tier	Freemium	JSON, XML

However, the flexibility and extensibility of these formats is a double-edged sword, and the structure of the returned response can can sometimes change without notice or its documentation can be incomplete or not up-to-date, especially in the case of still in-development APIs, relying on the human-readable nature of these text-based formats that make the returned data somewhat[7] self-documenting.

Finally, it must be noted that a growing amount of web services require authentication for billing purposes and/or to prevent abuse[8], and can offer the following tiers for the use their services:

- Free without restrictions (generally for community-based projects).
- Free with restrictions or reduced functionality[9] (e.g. number of queries, volume of data retrieved, trial periods, intended use).
- Paid-only services (e.g. billable per query, volume of data, subscription models).

3.2 Data Sources

3.2.1 Services

During the research for the present dissertation, data from several online services was collected using their corresponding public APIs[10] (Table 3.1) to assess their suitability as urban data sources.

[7] But not trivially.

[8] Increasingly using the OAuth open standard.

[9] The combination of "free with restrictions" + "additional paid services" is often denominated *freemium* (a portmanteau of free + premium).

[10] Although some data was not available through an API and had to be retrieved using web scrapping.

Of these services, all but the Google Geocoding Service have a social networking component, and are discussed in the present chapter. These networks popularity rise and fall over time (Fig. 3.2), as shown by their individual evolution based on Google Trends data for Facebook[11], Flickr[12], Instagram[13], Panoramio[14] and Twitter[15], making their study a moving target:

- Instagram (owned by Facebook) is the only network enjoying a steady and sustained growth, and the most "trendy" service at the time of writing.
- Twitter began declining after eroding some of Facebook's user base, and is now being surpassed in popularity by Instagram.
- Facebook showed a decline from its maximum around 2013 as its alternatives diversified, although it keeps it dominant position unchallenged.
- Flickr (owned by Yahoo during the research and by SmugMug since April 2018) began a steady decline in 2012, probably as some of its users migrated to other services such as Facebook to fulfill their photo-sharing needs.
- Panoramio (owned by Google) showed a decline since 2010 until its closure in November 4, 2016.

At the time of writing, Facebook was the most popular social network[16] (Table 3.2) according to Alexa, attracting 12 times more web searches than Instagram[17] and 17 times more than Twitter in the past year, according to Google Trends data. However, Facebook data was not used in this research because its main focus was not on location data but on personal relationships, and also because of the difficulty in accessing its data.

3.2.2 Selection Criteria

The selected sources were chosen according to their capacity to provide a significant amount of accurate location data; since a significant amount of photographs taken by smartphones are automatically geotagged[18] and time-stamped, the choice was biased towards photo sharing applications (Panoramio, Flickr and Instagram).

Therefore, if we consider a photograph an *event* (determined by the user as something worth taking a picture of), the retrieved data is firmly anchored in space (the location of the picture-taking device on the surface of the earth) and

[11] Facebook data available at https://trends.google.com/trends/explore?date=all&q=Facebook
[12] Flickr data available at https://trends.google.com/trends/explore?date=all&q=Flickr
[13] Instagram data available at https://trends.google.com/trends/explore?date=all&q=Instagram
[14] Panoramio data available at https://trends.google.com/trends/explore?date=all&q=Panoramio
[15] Twitter data available at https://trends.google.com/trends/explore?date=all&q=Twitter
[16] And third most popular site behind Google and Youtube accoding to Alexa, with data available at *http://www.alexa.com/topsites* at the time of writing.
[17] Instagram is also part of Facebook since April 2012.
[18] Using remarkably accurate location data provided by the device multiple sensors.

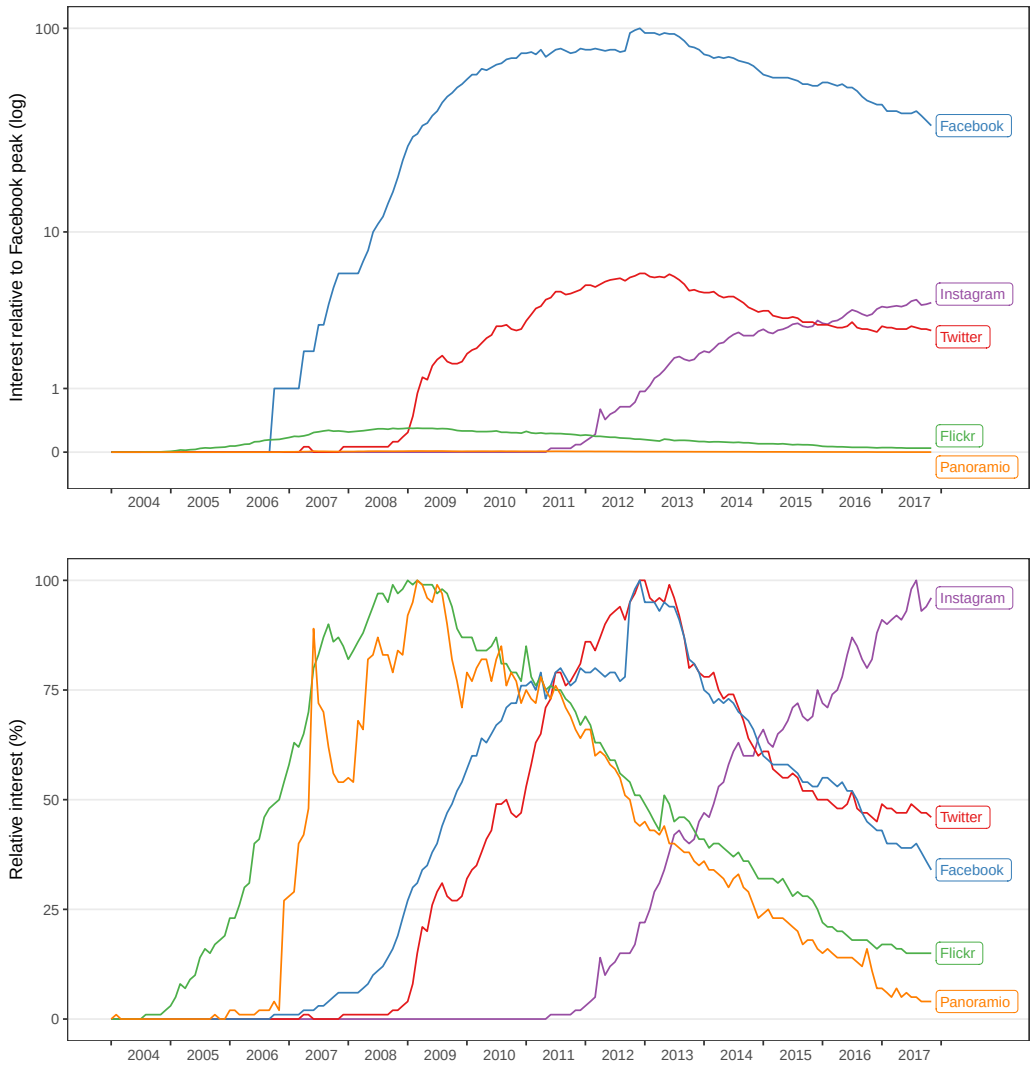


Figure 3.2: Evolution of the worldwide absolute (top) and relative (bottom) popularity of the researched sources since 2004 compared to the most popular service (Facebook), based on the volume of online search queries. Top figure uses a logarithmic scale. Bottom figure scales time series between zero and its corresponding peak. Source: Own work based on Google Trends data (retrieved on November 29, 2017).

Table 3.2: Ranks of the discussed social networks according to data from Alexa, and user base sizes according to their respective Wikipedia entries (data retrieved on July 30, 2017).

Source	Alexa rank	User base (millions)
Facebook	3	2,000
Twitter	12	319
Instagram	18	700
Flickr	322	87
Panoramio	8,684	4 (closed)

time (the instant when the picture was taken). In contrast, if the event is a tweet, it can provide the spatio-temporal coordinates of the user where and when it was *sent*, but may or may not correspond to the spatial or temporal context of the event being *referred to*.

Furthermore, images *can* be produced with much less effort[19] (just pointing a camera and pushing a button) and offer greater immediacy than composing a piece of text, even if it just contains 140 characters or less —as exemplified by the increasingly pictorial[20] and less verbose[21] millennials— allowing the collection of comparatively larger amounts of data.

Therefore, pictures shared on social networks can provide insights on the image of the city and its public spaces as perceived by its users, which select the most relevant landmarks according to their own preferences. However, despite this focus on geotagged picture content, other associated information or metadata was also collected when available, in addition to the location data:

- Temporal data (creation and/or upload).
- Textual descriptions (allowing the extraction of semantic content and language identification).
- Links to the actual image content (digital file with the picture).
- User profile (and his or her associated metadata).
- Device metadata (model, operating system).

[19] In addition, pictures can be perceived as less ideologically charged.
[20] As evidenced by the multitude of people taking "selfies".
[21] Manifested in the raising ☺ use in messaging.

Table 3.3: Retrieval dates and software used for data collection from all service APIs researched.

Service API	Retrieval date(s)	Software
Panoramio	June 8, 2016	Custom (R)
Flickr (images)	March 21 – 22, 2017	Custom (R)
Flickr (users)	March 23 – 25, 2017	Custom (R)
Instagram	August 19 – 20, 2017	Custom (R)
Twitter (streaming)	October 26 – 27, 2016	streamR
Twitter (search)	November 8, 2016 – October 31, 2017	twitteR
Google Geocoder	April 20 – 24, 2017	Custom (R)

3.2.3 Collected Data

The different API implementation of each source required developing custom software[22] to collect the available data, taking advantage of open-source libraries when available (Table 3.3). The retrieval process for each source will be described in the following subsections, sorted in the chronological order they were developed, which was roughly aligned with the difficulty in accessing their respective APIs:

1. Panoramio (see section 3.3 starting on page 59).
2. Flickr (see section 3.4 starting on page 64).
3. Instagram (see section 3.5 starting on page 71).
4. Twitter (see section 3.6 starting on page 77).

The data from the different sources covered very different time periods, from one single day streaming Twitter data to more than 13 years of Flickr user data (Table 3.4). The time frames of some sources overlapped for long periods of time such as Panoramio, whose almost 9-year range was completely overlapped by Flickr's (Fig. 3.3).

The volume of data retrieved was also variable, as well as the total unique records and number of fields (Table 3.5), being Flickr the largest with the over a million unique records with 69 fields each. The same service could return different number of fields depending on the API accessed (e.g. Twitter streaming and search).

For the Panoramio and Flickr APIs, the respective queries required a bounding box, which was defined as roughly the size of the minimum bounding rectangle (MBR) of the Barcelona city limits (Table 3.6), in the WGS[23] 84 global reference

[22] All the retrieval and analysis pipeline was implemented in the open-source R programming language.

[23] World Geodetic System.

Table 3.4: Temporal spans for the researched service APIs (earliest and latest dates, and total length in days).

Service API	Earliest date	Latest date	Days
Panoramio	October 3, 2005	August 5, 2014	3228
Flickr (images)	January 1, 2005	December 31, 2016	4382
Flickr (users)	February 4, 2004	March 16, 2017	4789
Instagram	Not applicable	Not applicable	-
Twitter (streaming)	October 26, 2016	October 27, 2016	1
Twitter (search)	November 1, 2016	November 31, 2017	365

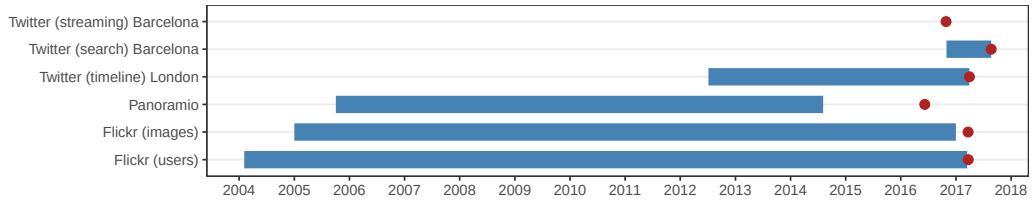


Figure 3.3: Temporal spans for the researched service APIs (as blue bars from the earliest recorded date to the latest) showing periods of overlapping temporal data, and corresponding dates of retrieval (as red dots). Longer periods at the bottom and shorter periods at the top. Source: Own work.

Table 3.5: Total unique records, number of fields and allocated memory for the data retrieved from the researched service APIs.

Service API	Unique records	Fields	Memory
Panoramio	80459	13	26.8 Mb
Flickr (images)	1166704	69	2.2 Gb
Flickr (users)	34283	23	40.1 Mb
Instagram	10980	4	21.3 Mb
Twitter Streaming	21943	42	14 Mb
Twitter Search	516990	16	179.3 Mb
Twitter Timeline (London)	1105	35	994.5 Kb

Table 3.6: Extent of the defined bounding box in the Panoramio and Flickr API requests.

Limit (edge)	Position	Facing	Decimal degrees	DMS
Min Latitude	Bottom	South	41.306105095 N	41° 18' 21.98" N
Max Latitude	Top	North	41.478298448 N	41° 28' 41.87" N
Min Longitude	Left	East	2.044216148 E	2° 2' 39.18" E
Max Longitude	Right	West	2.237585016 E	2° 14' 15.31" E

system for geospatial information used by the GPS satellite navigation system (EPSG:4326[24]).

3.2.4 Research Process Timeline

The following sections (3.3, 3.4, 3.5 and 3.6) will introduce the data sources used in the research, and outline the retrieval process for each of them. The sources will be described in the order they were developed, which followed a progression of increasingly difficulty in accessing the data (from a technical point of view).

Panoramio This was the first attempt to get data using an API in the research process. Since the calls were not authenticated and used a single method with few parameters, it was relatively straightforward to review the documentation to develop a script to retrieve the data using trial-and-error, and to develop strategies to sidestep its limitations. Unfortunately Panoramio data is not available any longer, and alternative sources of data are not as newbie-friendly (section 3.3).

Flickr The quality of the Panoramio data was unexpectedly good, which was encouraging and spurred the interest in investigating other sources, and Flickr was the service which seemed to share more similarities with it. In this case the API had much more functionality but as a result it was much more complex to use, and it required authentication. However the amount of data collected increased ten-fold, and provided much richer metadata (section 3.4).

Instagram The overwhelming popularity of Instagram, and the fact that it was a photo-sharing service like Panoramio and Flickr, made its API worth exploring. However, it resulted too restrictive to be useful for data research by third parties, and it only was possible to retrieve the locations from the Facebook place database, which was useful later in the research (section 3.5).

[24]The definition of EPSG:4326 is available at *http://spatialreference.org/ref/epsg/4326/* at the time of writing.

Twitter The limitations found using Instagram data prompted researching an alternative source. The Twitter API had enormous functionality at the expense of greater complexity, but the availability of open source software libraries allowed interacting with the API more easily than using purely home-grown methods as before. However, Twitter positional data was not comparable in terms of accuracy to the locations retrieved from Panoramio and Flickr. Despite this limitation, its capacity to capture semantic content using text mining tools was unique, es the imprecision issues could be significantly reduced using a broader scale of analysis (section 3.6).

3.3 Panoramio Data

3.3.1 Service Overview

Founded by two Spanish entrepreneurs, Panoramio[25] *was*[26] a platform for sharing geolocated pictures launched on October 3, 2005 and subsequently acquired by Google on June 27, 2007. The service popularity (Fig. 3.2) increased sharply in 2007 —around the time of its acquisition— but peaked in 2009, when it began to slowly decline until its eventual closure.

One of its distinct features was the capacity to overlay picture locations on Google Maps and Google Earth, with the objective of exploring virtually the pictures taken in a specific area[27]. Although neither its popularity nor the size of its user base (Table 3.2) was comparable to the other services explored in the research —which probably played an important role in its closure—, its focus on geotagged pictures could have provided very valuable data for urban research.

Panoramio closed operations on November 4, 2016 and was completely retired 12 months later[28], during the research phase of this dissertation, exemplifying the sometimes short-lived nature of online services, which are a moving target for research and can severely impair long-term investigations.

However, the MapSights[29] website hosts an ongoing project to preserve photos and location data from Panoramio. Furthermore, the Sightsmap[30] website provides worldwide coarse-resolution heatmap of photo density using data sourced from

[25] A snapshot of the Panoramio website as it appeared on April 27, 2016 is available at *http://archive.fo/yDNfV/image* at the time of writing.

[26] Since at the time of writing the service has ceased operations.

[27] It could be argued that Panoramio was a crowd-sourced precursor to Google Street View (launched in May 2005).

[28] However, the layer in Google Earth was available until January 2018.

[29] MapSights website is available at *http://mapsights.com/* at the time of writing.

[30] Sightsmap website is available at *http://sightsmap.com/* at the time of writing.

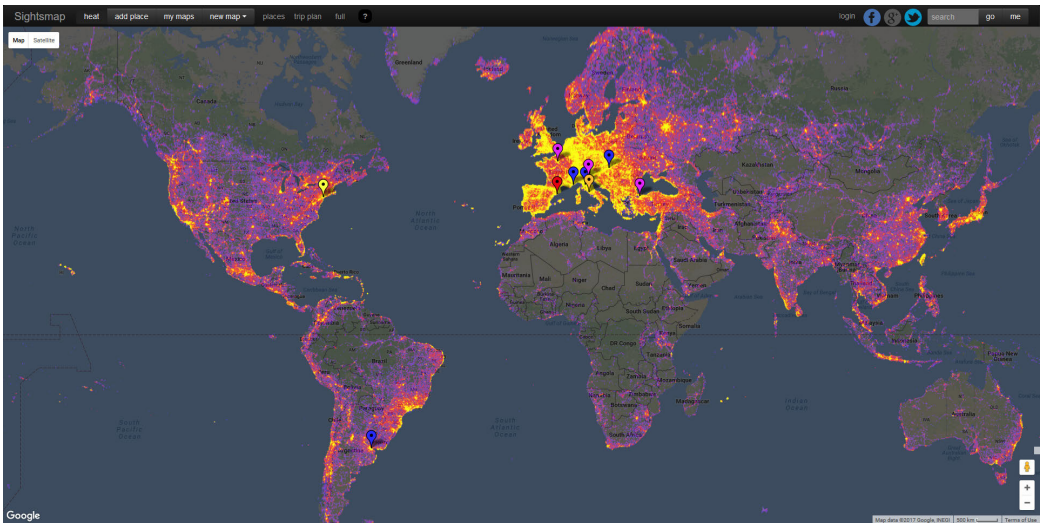


Figure 3.4: Screenshot of the Sightsmap website (retrieved on February 6, 2017) showing a worldwide heatmap of picture density. Barcelona appears as one of the top ten highlighted cities (red marker) with a rank of 3.

Panoramio (Fig. 3.4). In the map, 10 cities that are considered hot spots are highlighted and ranked according to their number of pictures taken (Table 3.7), and Barcelona appears in the third position behind New York City and Rome.

3.3.2 Panoramio API Features and Limitations

In contrast with the other services discussed (Flickr, Instagram and Twitter), access to Panoramio data did not require neither registration nor authentication (Table 3.1), the request format it used was minimalistic and the structure of its JSON response with the returned data was very simple. For these reasons it was the first service explored in this research, serving as a test-bed to develop retrieval and parsing methodologies, which were later refined and applied to other services.

Panoramio provided two APIs —which were deprecated in November 2016, shortly after the data was collected by the author on June 8, 2016 (Table 3.3)— that served two different and complementary purposes:

- Panoramio Data API[31], allowing the retrieval of Panoramio photos to display in an third-party website.

[31] An archived capture (May 10, 2016) of the Panoramio Data API documentation is available at *http://web.archive.org/web/20160510095912/https://www.panoramio.com/api/data/api.html* at the time of writing.

Table 3.7: Top ten cities according to their number of pictures taken, as they appear in the Sightsmap website (using data from Panoramio).

Rank	City
1	New York City
2	Rome
3	Barcelona
4	Paris
5	Istanbul
6	Venice
7	Monte Carlo
8	Florence
9	Buenos Aires
10	Budapest

- Panoramio Widget API[32], a JavaScript library to embed a widget with a user interface and search capabilities in external websites.

Only the Panoramio Data API[33] was used throughout this research, building the request URL from its single endpoint[34]. The API call was very simple, and did not require registration, authentication or encrypted transport, and allowed the following request parameters about the data to be returned:

- Set of pictures: either popular photos, all photos, or photos belonging to a specific user ID.
- Chunk of photos to retrieve, sorted according to their upload date (most recent first).
- Bounding box to filter pictures between a maximum and minimum latitudes and longitudes.
- Size of the picture: original, medium (default value), small, thumbnail, square, mini square.
- A "mapfilter" parameter to reduce clutter when showing locations on a map, avoiding almost-overlapping locations.

The Panoramio API was free for both commercial and non-commercial purposes,

[32] An archived capture (June 2, 2016) of the Panoramio Widget API documentation is available at *http://web.archive.org/web/20160602030004/http://www.panoramio.com/api/widget/api.html* at the time of writing.

[33] An overview can be found at *https://www.programmableweb.com/api/panoramio* at the time of writing.

[34] Located at *http://www.panoramio.com/map/get_panoramas.php* active during the retrieval phase of the research.

with a limitation of 100,000 queries through the API per day[35], slightly above the size of the retrieved dataset for Barcelona, which consisted in 82,477[36] records.

3.3.3 Available Data from Panoramio

The data retrieved through the Panoramio API was returned in JSON format, with a simple and predefined structure, without arbitrarily nested objects or variable length arrays. Each response envelope contained two key-value pairs (named "count" and "photos"):

- The first was the number of available pictures for the specific set and bounding box requested.
- The second consisted in an array with the data of each of the returned pictures in the requested chunk.

Because of the simplicity of the API, it not was possible to specify the desired data or metadata to be returned, and each of the retrieved pictures always included the following fields:

- A unique photo ID.
- The title given to the picture by its owner.
- URLs pointing to the picture location (page and image file).
- The picture longitude and latitude.
- The picture width and height.
- The date the picture was uploaded (day, month and year).
- Data about the owner: unique user ID, user name and profile URL.

3.3.4 Retrieved Panoramio Data

The Panoramio data was retrieved on June 8, 2016 (Table 3.3), requesting pictures within a defined bounding box (Table 3.6) enclosing the city limits of Barcelona. The retrieval process collected 82,477 records, of which 80,459 were unique (Table 3.5), covering a temporal span of almost 9 years (Fig. 3.3).

Although this amount was comparatively small compared to the rest of the services researched (in particular compared to the volume of data collected from Flickr), the focus of Panoramio on location data (specifically geotagged pictures) resulted in a good source of point data, less noisy than data collected from other services[37] (Fig. 3.5).

[35] Since the service was not authenticated, the author a assumed that it was limited according to the originating IP.

[36] Of which 80,459 were unique.

[37] Although in those cases, it could be somewhat compensated by their sheer amount of samples.

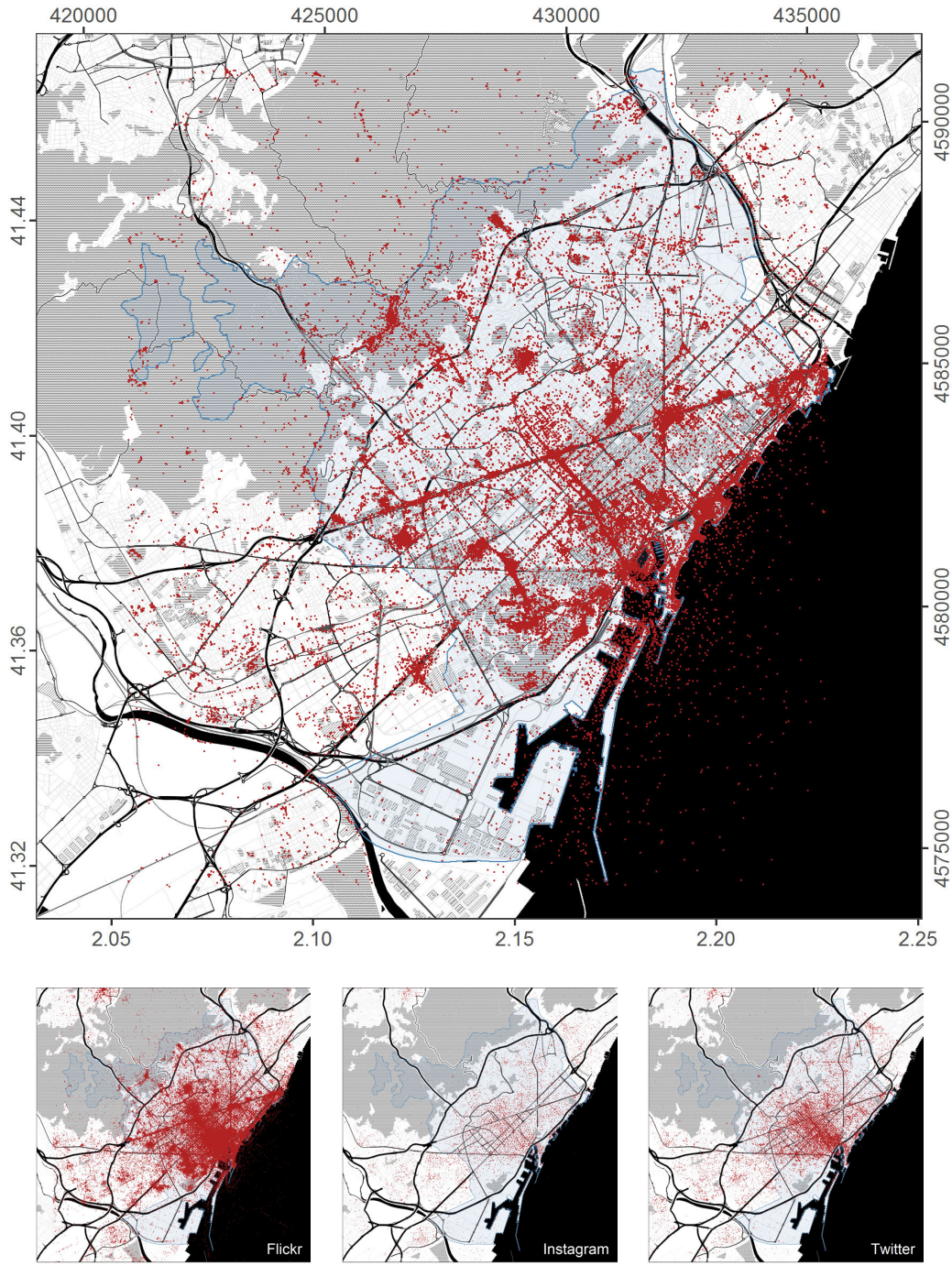


Figure 3.5: Overview of the 80,459 unique geotagged pictures retrieved using the Panoramio API (in red), overlaid on the official Barcelona municipal limits according to the ICGC (in faint blue) in the WGS 84 / Pseudo-Mercator – Spherical Mercator (EPSG:3857) projected coordinate system. Bottom miniature maps correspond to the equivalent Flickr (left), Instagram (center) and Twitter (right) retrieved data. Base map tiles by Stamen Design, under CC BY 3.0, using data from OpenStreetMap, under ODbL.

3.4 Flickr Data

3.4.1 Service Overview

Flickr[38] is an image[39] hosting service founded by the Canadian company Ludicorp on February 10, 2004 and acquired by Yahoo one year later (March 20, 2005), eventually replacing the Yahoo Photos service in September 20, 2007. After finishing the daft of this thesis, the service was acquired by SmugMug on April 2018, and later announced that the free option would be limited to store a maximum of 1,000 photos or videos[40].

The service popularity (Fig. 3.2) increased steadily since its launch, and peaked in 2009, when it began to slowly decline[41], as its features began to overlap with the photo-sharing functionality of Facebook, and later when its lack of integration with mobile devices lagged behind other services such as Instagram, which offered a more immediate experience.

However, it remains a popular service, with 87 million active users (Table 3.2) comprised mostly of photography enthusiasts who add more than 3.5 million new images every day.

The service allows users to upload and organize pictures which can be commented by its online community. Although it is possible to browse its catalog of public images without registration, the creation of an account is mandatory to upload content and use its social networking features, and provides the user with a personal profile page.

Flickr offers either free (advertisement supported) and paid services, and since 2009 allows users to receive payment when their picture content is used as stock photography, through a partnership with Getty Images.

Among the image metadata, Flickr allows storing location information for geotagged pictures, which users can use to produce a map mashup, overlaying their picture locations on Google Maps cartography using the iMapFlickr[42] service.

The availability of location information makes Flickr data suitable for visualizing urban phenomena, as has been explored by the artist and programmer Eric

[38] The Flickr website is available at *http://www.flickr.com/* at the time of writing.

[39] Although it allows hosting video content as well.

[40] According to the post "Free Flickr accounts slashed to 1,000 pictures; the rest will be deleted" published on ArsTechnica, available at *http://arstechnica.com/gaming/2018/11/free-flickr-accounts-slashed-to-1000-pictures-the-rest-will-be-deleted/* at the time of writing.

[41] According to the post in The Verge "The man behind Flickr on making the service 'awesome again'" available at *http://www.theverge.com/2013/3/20/4121574/* at the time of writing.

[42] iMapFlickr is available at *http://imapflickr.com/* at the time of writing.

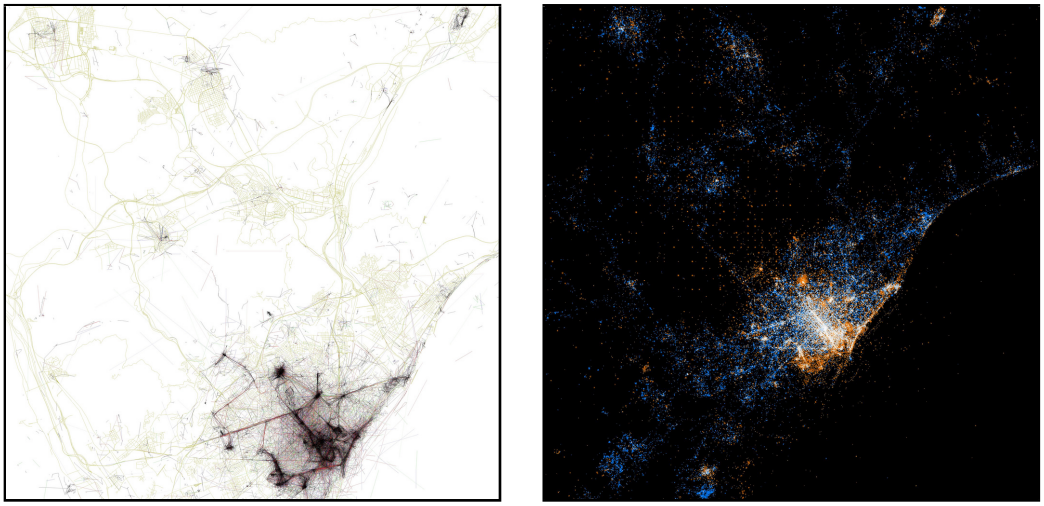


Figure 3.6: Eric Fischer's Barcelona images from the map series "The Geotaggers' World Atlas" (left) and "See something or say something" (right). Credit: Flickr user Eric Fischer (CC BY-SA 2.0).

Fischer[43], who appropriately publishes his maps in his own Flickr page[44].

In his series "The Geotaggers' World Atlas"[45], Fischer uses the photo locations from the public Flickr and Picasa search APIs to reveal the movement patterns in 100 cities, among which we can find Barcelona[46], and in another series named "See something or say something"[47], compares the relative abundance of Tweets (text) compared to Flickr images (pictures) in 50 locations[48], also including Barcelona[49] (Fig. 3.6).

In addition to location, data hosted on Flickr has also been used in research to analyze its image content (e.g. to analyze the change in the color content of

[43] CityLab has a profile interview of Eric Fischer titled "Mapmaker, Artist, or Programmer?" available at *http://www.citylab.com/design/2012/08/mapmaker-artist-or-programmer/3132/* at the time of writing.

[44] Eric Fischer's personal Flickr page is available at *http://www.flickr.com/photos/walkingsf* at the time of writing.

[45] "The Geotaggers' World Atlas" map series is available at *http://www.flickr.com/photos/walkingsf/sets/72157623971287575/* at the time of writing.

[46] The Barcelona map in "The Geotaggers' World Atlas" is available at *http://www.flickr.com/photos/walkingsf/4621767399/* at the time of writing.

[47] The "See something or say something" map series is available at *http://www.flickr.com/photos/walkingsf/albums/72157627140310742* at the time of writing.

[48] According to its description, "Red dots are locations of Flickr pictures. Blue dots are locations of Twitter tweets. White dots are locations that have been posted to both."

[49] he Barcelona map in "See something or say something" is available at *http://www.flickr.com/photos/walkingsf/5928667207/* at the time of writing

Table 3.8: Most popular tags for Barcelona, according to Flickr.

All time popular tags	Tag count
Catalonia (catalunya + catalonia + cataluña)	19,837
Spain (spain + españa)	10,973
street	6,450
architecture	5,137
art	4,420
night	4,260
water	3,928
gaudi	3,853
building	3,739
hdr	3,346
bcn	3,219

pictures throughout the yearly cycle[50]). Moreover, metadata added by users such as tags can also be valuable to find out the perception of a place[51] by its users (Table 3.8).

3.4.2 Flickr API Features and Limitations

At the time of retrieval, the Flickr API (Table 3.1) was free to use[52] non-commercially[53]. In contrast with the Panoramio API, some of its functionality was only available using an API key[54] —available upon registration— although some features required authentication —using the OAuth protocol[55]—.

Introduced in 2005, the API was very mature and its documentation comprehensive[56]. Almost all the functionality available on the Flickr website was accessible

[50] In the R-bloggers post of May 1, 2013 titled "Color analysis of Flickr images" by Benedikt Koehler available at *http://www.r-bloggers.com/color-analysis-of-flickr-images/* at the time of writing.

[51] The most popular tags for Barcelona in Flickr are available at *http://www.flickr.com/places/info/12693396* at the time of writing.

[52] According to its terms of service, available at *http://www.flickr.com/services/api/tos/* at the time of retrieval.

[53] At the time of retrieval, commercial use of the API was allowed only with prior permission.

[54] The Flickr API Keys application process was detailed at *http://www.flickr.com/services/api/misc.api_keys.html* at the time of retrieval.

[55] The Flickr OAuth authentication process was detailed in the documentation at *http://www.flickr.com/services/api/auth.oauth.html* at the time of retrieval.

[56] The API index page was available at *http://www.flickr.com/services/api/* at the time of retrieval.

through it[57], either interfacing directly or through unofficial third-party API kits, generally in the form of software libraries available for most popular programming languages (e.g. Python, Java). The Flickr API endpoint[58] provided a list of named parameters:

- The calling method (required).
- A number of arguments specific to each method (optional).
- The API key (required).

At the time of writing, there were 221 methods listed in the API documentation, grouped into 40 categories ranging from data search and retrieval to photo and profile management. Each method had different limitations regarding its authentication requirements, usage limits and privacy restrictions. Only two of these methods were used to retrieve the data during the research: the *flickr.photos.search* method and the the *flickr.profile.getProfile* method, discussed in the next subsection.

3.4.3 Available Data from Flickr

The picture data was retrieved through the Flickr API using its *flickr.photos.search* method[59]. This method queries the database in order to return a list of photos matching a specific set of criteria, passed as arguments in the API call.

Since the main research focus was the spatial distribution of the geotagged pictures, this search method was especially useful because it allowed requesting image data corresponding to a specific location in multiple ways:

- Inside a bounding box (as the coordinates of the north, south, east and west edges of a "rectangle" oriented along the axis of a geographic reference system).
- Within a specified distance from a given coordinate (longitude and latitude).
- A Where On Earth Identifier (WOEID), a unique 32-bit reference identifier maintained by Yahoo.
- A Flickr place ID (an identifier specific to the Flickr service).

Furthermore, beyond the delimitation of the geographic region where the images were requested from, other location-related parameters were optionally available in the request:

[57] According to The Flickr Developer Guide available at *http://www.flickr.com/services/developer/* at the time of retrieval.

[58] The entry point to the Flickr REST API was *http://api.flickr.com/services/rest/* at the time of retrieval.

[59] Described in detail in the corresponding entry for the *flickr.photos.search* method in the API documentation available at *http://www.flickr.com/services/api/flickr.photos.search.html* at the time of writing.

- Defining a minimum accuracy in a range 1–16, being 11 the city level and 16 the street level.
- Include only geotagged pictures (discarding images without longitude and latitude coordinate pairs).
- Restricting the results to only pictures taken either indoors or outdoors (provided that this information was available).

Finally, beyond the parameters related to the picture location, other search parameters were also available to further specify the required attributes of the returned image data in the API request:

- Restrict pictures to a specific user or gallery.
- Search for a specific string in tags, title or description.
- Return results within a range of dates when the picture was either taken or uploaded.
- Specific restrictions such as license, privacy or adult content.
- Filter the type of media (e.g. photographs, screenshots, video).
- Specify additional metadata to return (instead of the default).

However, some valuable information related to the images did not belong to the pictures themselves, and instead was stored as an attribute of the user[60] that took/uploaded them. Therefore, the profile info of all the users who had uploaded at least one picture within the bounding box during defined time span were retrieved using the API method *flickr.profile.getProfile*[61], provided that he or she had made at least part of this information publicly available. In addition, since some of the profile data was not available through the API[62], additional information was retrieved scraping[63] the HTML content of the users' public profile pages.

3.4.4 Retrieved Flickr Data

The Flickr picture data was collected between March 21 and 22, 2017 (Table 3.3), and covered a 12-year period (Table 3.4). Since the main interest was their location, the pictures were requested within a defined bounding box (Table 3.6), which

[60] Or in Flickr parlance, the owner.

[61] Described in detail in the corresponding entry for the *flickr.profile.getProfile* method in the API documentation available at *http://www.flickr.com/services/api/flickr.profile.getProfile.html* at the time of writing.

[62] Notably sex and relationship status.

[63] While the legality of web scraping of public websites is controversial, a California federal court rejected a LinkedIn claim that scraping was hacking against startup hiQ, according to https://arstechnica.com/tech-policy/2017/08/court-rejects-linkedin-claim-that-unauthorized-scraping-is-hacking/

matched the region previously defined to retrieve the data through the Panoramio API.

Since the search was not authenticated[64], the query only returned the pictures flagged as "public", with a total of 1,166,704 unique records with 69 variables[65] each (Table 3.5), among which the most useful were the following:

- Owner (which allowed linking the data from the corresponding user profile).
- Location (longitude, latitude, estimated accuracy)[66].
- Date and time when the picture was taken (including estimated granularity).
- Title and description of the picture.
- Tags (including automatically generated machine tags).
- URL linking to the stored picture (in multiple sizes and aspect ratios).

All retrieved pictures belonged to 34,283 unique users (Table 3.5), resulting in an average of 34 pictures taken per user within the region of interest. After querying the information of each of these users through the API[67] —no personally identifiable information was kept—, the following information was available for roughly 50% of them, although it was not possible to determine whether it was always authentic:

- Links to other social media accounts belonging to the user (personal website, Facebook, Twitter, Tumblr, Instagram, Pinterest).
- First and last name.
- User self-description.
- Occupation.
- Hometown and place of residence (city and country).
- Sex and status.

The Flickr dataset contained 14.5 times more points (Table 3.5) than the Panoramio dataset (Fig. 3.5). Even more dramatically than in the case of the Panoramio data, the number of points introduced serious challenges for their representation and interpretation due to overplotting (Fig. 3.7). The methodologies developed to overcome or mitigate these issues will be discussed in chapter 6.

[64] Only the API Key was used.
[65] After parsing the JSON responses.
[66] Flickr uses the Geo microformat for marking up WGS84 geographical coordinates, the daft specification is available at *http://microformats.org/wiki/geo* at the time of writing.
[67] And in some cases scraping their public profile web pages.

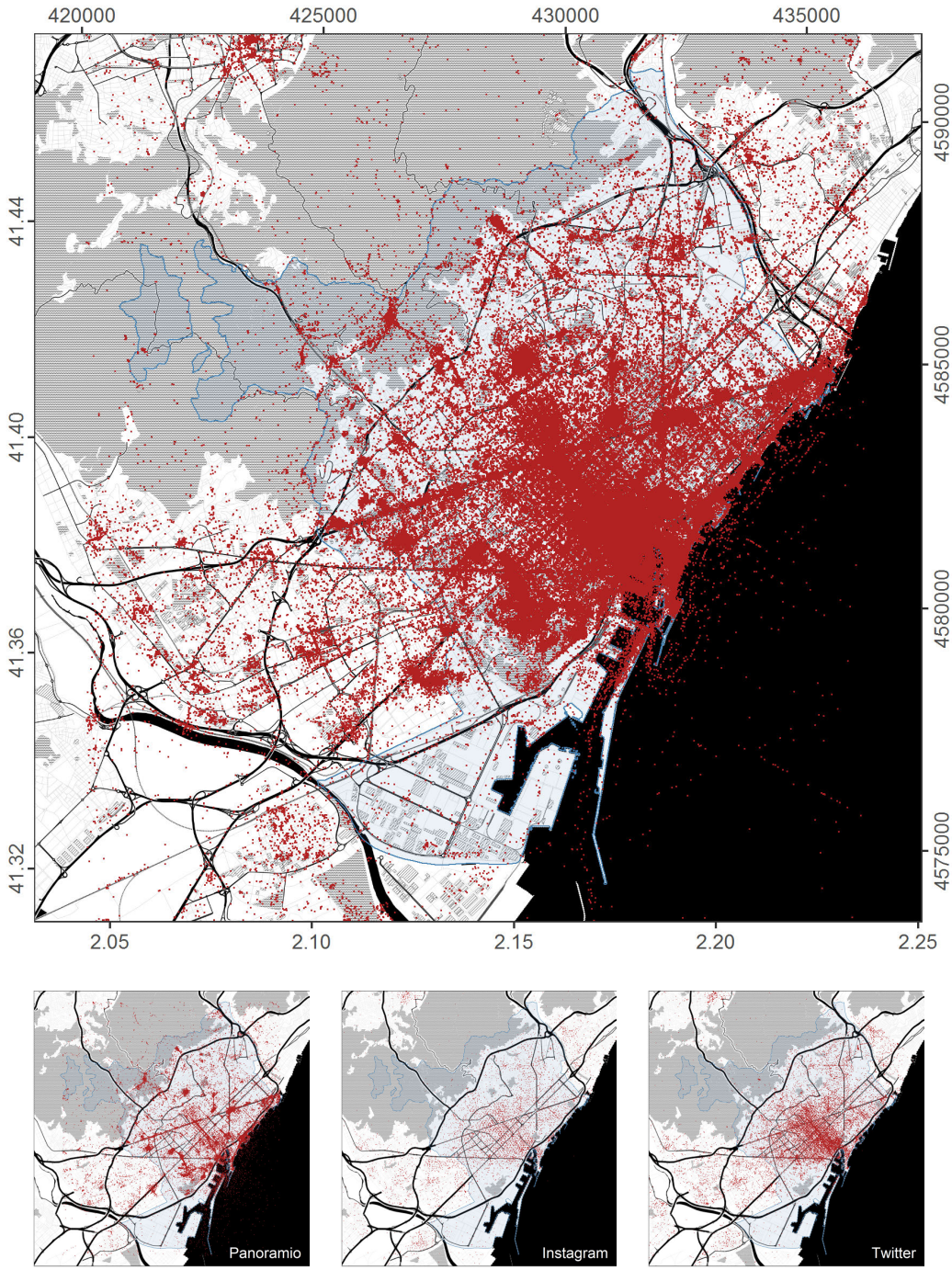


Figure 3.7: Overview of the 1,166,704 unique geotagged pictures retrieved using the Flickr API (in red), overlaid on the official Barcelona municipal limits according to the ICGC (in faint blue) in the WGS 84 / Pseudo-Mercator – Spherical Mercator (EPSG:3857) projected coordinate system. Bottom miniature maps correspond to the equivalent Panoramio (left), Instagram (center) and Twitter (right) retrieved data. Base map tiles by Stamen Design, under CC BY 3.0, using data from OpenStreetMap, under ODbL.

3.5 Instagram Data

3.5.1 Service Overview

Named after the portmanteau of "instant camera" and "telegram", Instagram[68] is a mobile-centric[69] photo-sharing[70] platform launched in October 6, 2010. The service was acquired by Facebook in April 2012.

Initially available only as an iOS[71] app, it became available for devices using the Android Operating System in April 2012. It became popular because of its capacity to enhance pictures —which initially had a distinctive square aspect ratio[72]— with high quality photographic filters.

According to data retrieved from Google Trends (Fig. 3.2), the service has attracted increasingly higher search queries since 2012, and is the only among the analyzed services exhibiting an upward trend in popularity at the time of writing.

Instagram has an user base of 700 million accounts (Table 3.2). The demographics of Instagram users are predominantly female[73] and young[74]. The service is free to use, generating revenue through advertising and paid analytic tools for business accounts.

The service allows its users to link their account to other social media services. In addition to uploading and sharing their pictures, users can add a title and multiple hashtags to their posts, and optionally include location information. However, the functionality to view custom maps of the geotagged pictures taken by users was removed in September 2016 because the service was not very widely used[75].

Because of its geotagging functionality, Instagram should be an excellent source of data to explore spatio-temporal phenomena in urban settings. In addition, since its users focus in sharing image-centric content, their locations are firmly attached to a specific place, and because of its focus on mobile devices, its immediacy anchors the recorded events in time. Furthermore, the text content of the titles and hashtags provides additional semantic context to location and temporal data.

[68] The Instagram website is available at *http://www.instagram.com/* at the time of writing.
[69] Although it can also be accessed through a web browser since November 2012.
[70] The functionality to share short 15-second videos was introduced in June 2013.
[71] Apple mobile operating system used by iPhones and iPads (collective known as iDevices).
[72] Emulating the classic Polaroid pictures.
[73] 68% of its users are women.
[74] 90% of the users are under the age of 35.
[75] Discussed in the Mashable post of September 6, 2016 "Instagram is killing photo maps" by Emma Hinchliffe available at *http://mashable.com/2016/09/06/instagram-kills-photo-maps/* at the time of writing.

3.5.2 Instagram API Features and Limitations

The Instagram API is not designed for data research, and focuses on interoperability with third-party software components to *extend* the core Instagram functionality, *without replicating* it[76]. It requires developers to create an account[77] and register an application, providing a short description of their intended use of the API.

The API is well documented[78] but not as feature-rich as other services, following the deprecation of their real-time API on November 17, 2015, according to the online changelog[79]. Instagram provides libraries for Python[80] and Ruby[81], but they are not actively maintained[82]. However there exist third-party software libraries to access the service using popular programming languages[83] such as JavaScript, Python or Ruby.

The service requires authentication[84] using OAuth 2.0 over a secure SSL connection (Table 3.1), to receive an access token to produce an authorized request. It allows requests to endpoints grouped in seven categories[85]: users, relationships, media, comments, likes, tags and locations. Data is requested including the following information in the call:

- The API endpoint.
- A number of named arguments.
- The access token.

In contrast with the other researched services, Instagram places developers in a

[76] According to its Platform Policy, available at *http://www.instagram.com/about/legal/terms/api/* at the time of writing.

[77] Using a regular user account, according to the Instagram documentation page on registration, available at *http://www.instagram.com/developer/register/* at the time of writing.

[78] The Instagram Developer Documentation is available at *http://www.instagram.com/developer/* at the time of writing.

[79] Deprecation of realtime subscriptions for tags, locations and geographies for apps created on or after November 17, 2015 available at *http://www.instagram.com/developer/changelog/* at the time of writing.

[80] The official Python library repository hosted on GitHub is available at *http://github.com/Instagram/python-instagram* at the time of writing.

[81] The official Ruby library repository hosted on GitHub is available at *http://github.com/Instagram/instagram-ruby-gem* at the time of writing.

[82] According to the Instagram documentation page on libraries, available at *http://www.instagram.com/developer/libraries/* at the time of writing.

[83] For a comprehensive directory of libraries see the list on ProgrammableWeb available at *http://www.programmableweb.com/api/instagram-graph/libraries* at the time of writing.

[84] Described in detail in the authentication section of the Instagram developer documentation available at *http://www.instagram.com/developer/authentication/* at the time of writing.

[85] According the API Endpoints section of the Instagram developer documentation available at *http://www.instagram.com/developer/endpoints/* at the time of writing.

limited "sandbox" environment[86]. Sandbox mode allows accessing the API but limits the available data[87] —effectively isolating the sandbox from Instagram data— and the rate limits[88], requiring the submission of the app for review to receive approval to operate with real live data.

3.5.3 Available Data from Instagram

The majority of the functionality offered by the Instagram API caters to the social networking aspect of the service, and involves dealing with users and their relationships, and managing the comments on the posted media, as well as their "likes".

The more useful tools for gathering urban data are the Media[89], Location[90] and Tags[91] endpoints, either of which offer the following functionality:

- Get information on media[92], tags[93] or locations[94].
- Get a list of recent media with a tag[95] or in a location[96].
- Search within an area media[97], tags[98] or locations[99].

However, only the last feature (locations search) appeared to be functional in sandbox mode, returning the public locations[100] indexed in the Facebook database. Despite the growing relevance of the service, these limitations to access the

[86] Described in detail in the sandbox mode section of the Instagram developer documentation available at *http://www.instagram.com/developer/sandbox/* at the time of writing.
[87] It only allows receiving data from up to 10 authorized sandbox users.
[88] Limited to 500 requests per hour, according to the limits section of the Instagram developer documentation available at *http://www.instagram.com/developer/limits/* at the time of writing.
[89] Media endpoints are detailed in the Instagram developer documentation at *http://www.instagram.com/developer/endpoints/media/* at the time of writing.
[90] Location endpoints are detailed in the Instagram developer documentation at *http://www.instagram.com/developer/endpoints/locations/ints/media/* at the time of writing.
[91] Tag endpoints are detailed in the Instagram developer documentation at *http://www.instagram.com/developer/endpoints/tags/* at the time of writing.
[92] Information on media is retrieved through the */media/{media-id}* endpoint.
[93] Information on tags is retrieved through the */tags/{tag-name}* endpoint.
[94] Information on locations is retrieved through the */locations/{location-id}* endpoint.
[95] The list of recent media with a tag is retrieved through the */tags/{tag-name}/media/recent* endpoint.
[96] The list of recent media in a location is retrieved through the */locations/{location-id}/media/recent* endpoint.
[97] Media is searched through the */media/search* endpoint.
[98] Tags are searched through the */tags/search* endpoint.
[99] Locations are searched through the */locations/search* endpoint.
[100] According to the Instagram help page "How do I create a new location?" available at *http://help.instagram.com/1618893218361276* at the time of writing.

Instagram data[101] motivated seeking more developer-friendly ways to access Instagram data, which in the end was accessed indirectly through the Twitter API for those users with linked Twitter and Instagram accounts (discussed in section 3.6).

3.5.4 Retrieved Instagram Data

The Instagram locations data were initially retrieved on June 23, 2016 followed by a second attempt with an improved methodology on August 19–20, 2017 (Table 3.3). The locations were collected within the same bounding box (Table 3.6) used to retrieve the Panoramio and Flickr data, for comparison purposes.

However, the Instagram API did not allow searching inside a bounding box, offering only geographic search for locations within a distance around a geographic coordinate. The parameters of this search were the following:

- Longitude and latitude coordinate pairs of the search center.
- User definable radius in meters (with a default of 500, and a maximum of 750).
- Number of returned locations (with a default of 20 and maximum[102] of 30).

Since the 750 meter radius was too small to cover the whole bounding box, and the number of results would be insufficient because of the limit in the number of returned locations, the bounding box was covered with a regular hexagon grid which ensured the minimum circle overlap (Fig. 3.8), and therefore required fewer API calls.

Since the centers had to be specified in geographic coordinates but the radius was defined in meters, the retrieval required the following process:

1. Projecting the bounding box into projected coordinates (EPSG:25831).
2. Tessellating the projected bounds with a hexagonal grid (with a separation of $\sqrt{3} \cdot radius$, being *radius* the search distance in the API request).
3. Convert the centers back into geographic coordinates (EPSG:4326).

Looping through the 11,889 centers with a 100 meter radius required around 30 hours due to the service 500 requests per hour limit in sandbox mode, and resulted in 349,903 records, of whom 10,980 were unique (Table 3.5). The retrieved data included only 4 variables: location ID, name, longitude and latitude.

The locations show the locations collected by Facebook with their users' input, and correspond mainly to business locations and landmarks. In contrast with

[101] Discussed in the January 5, 2017 post "Social Media Data – Instagram Pulls Back on API Access" on BrightPlanet available at *http://brightplanet.com/2017/01/instagram-data/* at the time of writing.

[102] Determined empirically, since the parameter was undocumented.

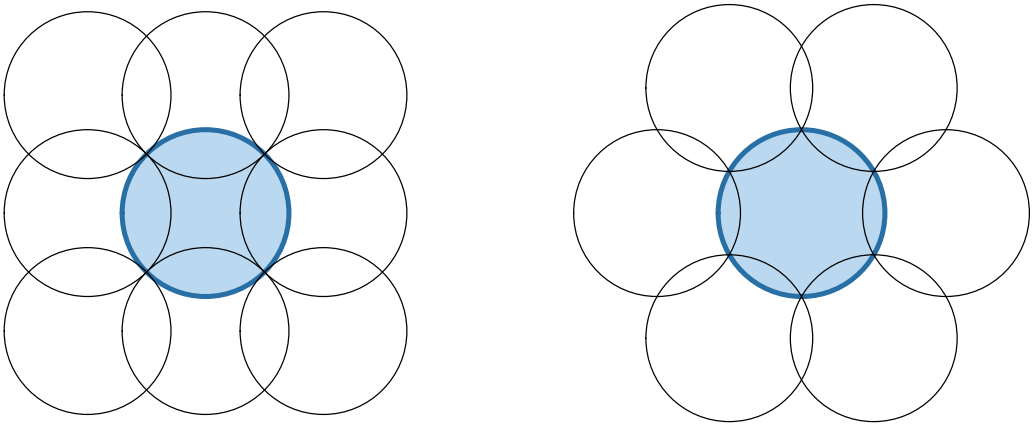


Figure 3.8: Patterns with the same radius packed in a square (left) and hexagonal (right) grid. The hexagonal packing provides a better overlap to coverage ratio. Source: Own work.

picture information, these data allows identifying “deserts” where users have located few attractive spots (Fig. 3.9).

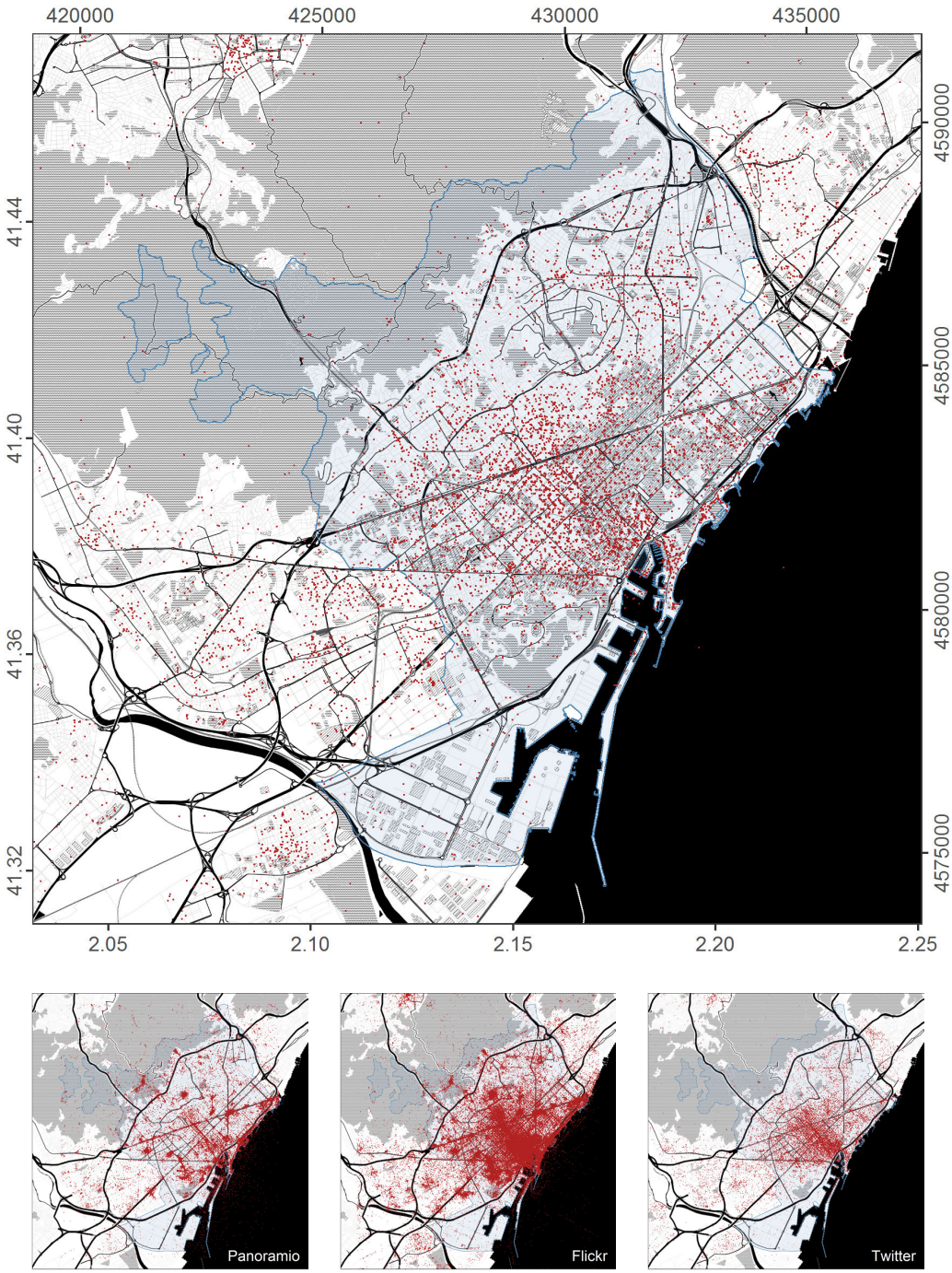


Figure 3.9: Overview of the 10,980 unique locations retrieved using the Instagram API (in red), overlaid on the official Barcelona municipal limits according to the ICGC (in faint blue) in the WGS 84 / Pseudo-Mercator – Spherical Mercator (EPSG:3857) projected coordinate system. Bottom miniature maps correspond to the equivalent Panoramio (left), Flickr (center) and Twitter (right) retrieved data. Base map tiles by Stamen Design, under CC BY 3.0, using data from OpenStreetMap, under ODbL.

3.6 Twitter Data

3.6.1 Service Overview

Sometimes nicknamed "the SMS of the Internet", Twitter[103] is challenging to define, having been described as an online news service, a social networking community, or a microblogging service. It was introduced publicly on July 15, 2006 after a short prototype phase that began on March 21, 2006.

Twitter became the top word of year 2009, according to the Global Language Monitor[104]. Although Google trends data shows a decline in its popularity since 2013 (Fig. 3.2), the service remains very popular (Table 3.2) and it has become a valuable news source and discussion forum[105].

Twitter users can post status messages (tweets), limited to the service distinctive maximum 140 character length, which was raised to 280 on November 2017[106]. Only registered users can post tweets, but unregistered users can also read them, provided they are designated as public. Users also have the possibility to subscribe to other users' tweets (timelines), becoming "followers".

Users have the possibility to like/favorite the messages of other users, and to forward a message to other users (a process known as "retweet"). In addition, they have the possibility to mention another username (known as "mentions", prefixed with "@") or to include a number of metadata tags (known as "hashtags", prefixed with a "#").

Although the service initially focused on text content, the service now also offers the possibility to attach images or videos, as well as links to web or media content (usually shortened[107] due to message length constraints).

Twitter users are mostly older adults, who perceive the service as more "serious", in contrast with other services like Instagram who are used by younger individuals. The gender distribution is split almost evenly[108] between women (47%) and men (53%) and more tweets are published using mobile devices[109] (86%) than using its

[103] The Twitter website is available at *http://twitter.com/*

[104] The Global Language Monitor Top Words of the 21st Century list is available at *http://www.languagemonitor.com/category/top-words-of-21st-century/* at the time of writing.

[105] As exemplified in the 2016 U.S. presidential election.

[106] Details of the rollout can be found in the Twitter blog post "Tweeting Made Easier " of November 7, 2017 available at *http://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html* at the time of writing.

[107] For example with Twitter own shortening service at *http://t.co* at the time of writing.

[108] According to Hootsuite data, available at *http://blog.hootsuite.com/twitter-demographics/* at the time of writing.

[109] According to 2013 data published on April 3, 2014 by The Wall Street Journal, available at *http://blogs.wsj.com/digits/2014/04/03/data-point-social-networking-is-moving-on-from-the-desktop/* at the time of writing.

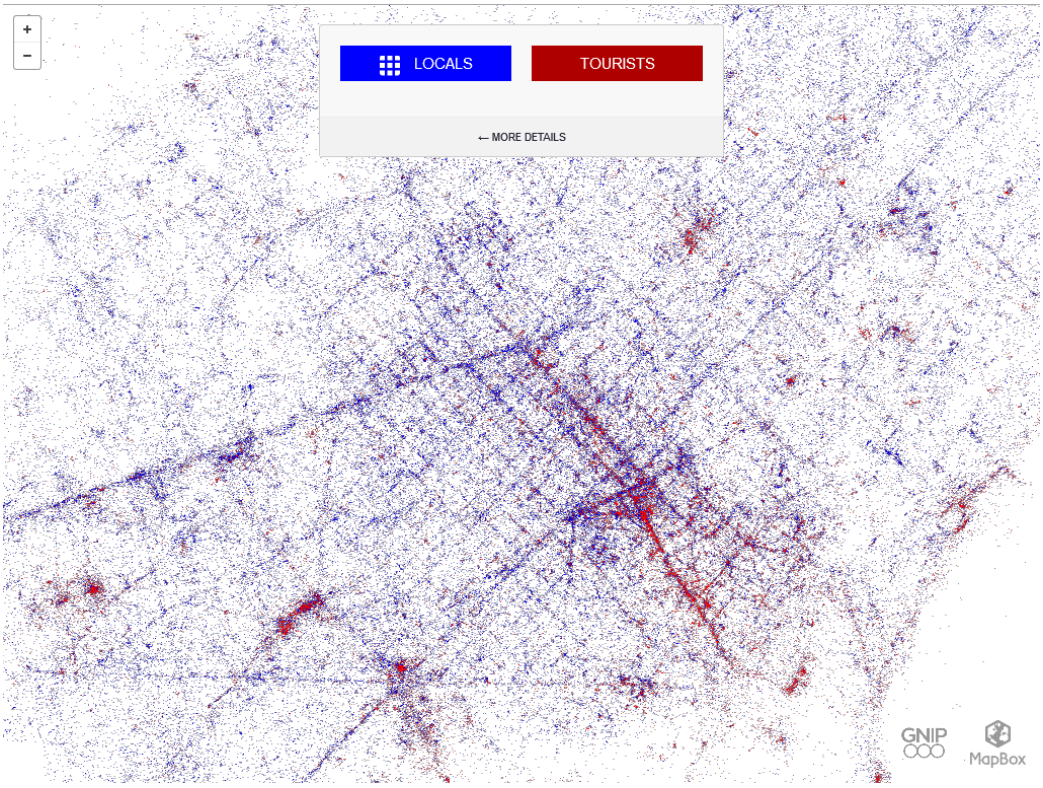


Figure 3.10: Map of 3 billion tweets (every geotagged tweet since September 2011), zoomed onto the Barcelona region. Blue dots represent "locals" whereas red dots represent "tourists". Source: Screenshot of the Locals & Tourists website developed by Eric Fischer, retrieved on August 15, 2017.

desktop web interface.

Urban phenomena can be analyzed through the locations of geotagged tweets, as in the work of Eric Fischer in his "Locals & Tourists"[110] interactive map[111], where locals —who post in one city for one consecutive month— are distinguished from tourists —whose tweets are centered in another city— according to they spatio-temporal behavior (Fig. 3.10).

In addition to the location of geotagged tweets, data from Twitter can be valuable in urban research for two unique aspects, distinct from image hosting services like Fickr (Fig. 3.6):

[110]This map uses the Twitter firehose data through Gnip, according to the post on the Mapbox blog "Visualizing 3 Billion Tweets" available at *http://www.mapbox.com/blog/visualizing-3-billion-tweets/* at the time of writing.

[111]The interactive "Locals & Tourists" map is available at *http://www.mapbox.com/bites/00245/locals/* at the time of writing.

- The semantic content of the status updates, which can be analyzed with text mining tools.
- The capacity to follow in real-time events as they unfold, which is unprecedented in urban analysis.

3.6.2 Twitter API Features and Limitations

The Twitter API is one of the most feature-rich and open API, attracting developer interest since its introduction. These developers innovate and improve the features of the service leveraging their access to Twitter data, and in some cases these developed technologies have been subsequently acquired by Twitter itself. Twitter offers three avenues to access to their data[112]:

- REST API[113].
- Streaming API[114].
- Twitter Firehose[115].

The REST and Streaming APIs are very well documented[116] and free to use, but only provide limited access to their data, as unrestricted access to all data is a paid service offered by the Twitter's enterprise API platform Gnip[117].

The REST API provides access to Twitter data with limitations[118] according to which of the 67 available endpoints[119] is being used[120], providing functionality similar to the Twitter client. Most of the data retrieval is done trough the search

[112]The differences are summarized in the blog post "Twitter Firehose vs. Twitter API: What's the difference and why should you care?" available at *http://brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/* at the time of writing.

[113]The Twitter REST API documentation is available at *http://dev.twitter.com/rest/public* at the time of writing.

[114]The Twitter Streaming API documentation is available at *http://dev.twitter.com/streaming/overview* at the time of writing.

[115]The Twitter Firehose is accessed through Gnip and is documented at *http://support.gnip.com/apis/* at the time of writing.

[116]The Twitter documentation index for developers is available at *http://dev.twitter.com/docs* at the time of writing.

[117]Gnip belongs to Twitter since April 2014, and its website can be found at *http://gnip.com/* at the time of writing.

[118]Details of the limits are available at *http://developer.twitter.com/en/docs/basics/rate-limiting* at the time of writing.

[119]At the time of writing.

[120]A comprehensive list of the rate limits per window for each of the endpoints is available at *http://developer.twitter.com/en/docs/basics/rate-limits* at the time of writing.

API[121]. In the case of the search endpoint[122] the limits are:

- 180 requests[123] inside a 15 minute window, each request returning up to 100 tweets.
- Returns only a sample[124] of recent Tweets.
- Restricts tweets published in the past 7 days[125].

In contrast, the streaming API provides low latency (near real-time) access to the global stream of Twitter data, and the messages are pushed to the client keeping a persistent HTTP connection open[126]. The streaming API has two modes of operation for public streams[127]:

- The filter endpoint[128] returns returns public statuses that match one or more filter predicates[129]:
  - A comma separated list of user IDs (up to 5000).
  - A comma separated list of keywords to track (up to 400).
  - A set of bounding boxes to track (up to 25).
  - A specific language (seems to have been deprecated at the time of writing).
- The sample endpoint[130] returns a small random sample of all public statuses[131] (around 1% of the public Tweet volumes at any time[132]).

Both free APIs require authentication using OAuth and setting up a Twitter app[133], and responses are provided in JSON format (Table 3.1).

[121] Documentation in the search API is available at *http://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets* at the time of writing.

[122] The Twitter search resource URL is *https://api.twitter.com/1.1/search/tweets.json* at the time of writing.

[123] Only when using User authentication, when using Application-only authentication the limit is 450.

[124] Multiple request attempts return the same sample.

[125] According to some sources, between 6 and 9 days.

[126] Only a single connection can be made at any time, and if more attempts to connect are made, the oldest connection is dropped.

[127] The streaming API documentation is available at *https://developer.twitter.com/en/docs/tweets/filter-realtime* at the time of writing.

[128] The Twitter realtime streaming resource URL is *https://stream.twitter.com/1.1/statuses/filter.json* at the time of writing.

[129] Details on the available parameters are available in the API reference at *http://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html* at the time of writing.

[130] The Twitter realtime sample resource URL is *https://stream.twitter.com/1.1/statuses/sample.json* at the time of writing.

[131] Details are available in the API reference at *http://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample* at the time of writing.

[132] According to the Twitter Developers forum post available at *http://twittercommunity.com/t/potential-adjustments-to-streaming-api-sample-volumes/31628* at the time of writing.

[133] Registration page available at *http://apps.twitter.com/* at the time of writing.

3.6.3 Available Data from Twitter

There are several libraries available to retrieve Twitter data for the most popular programming languages. In the development of the research the following packages were used in the R programming language (although the Tweepy[134] Python library was also considered):

- streamR 0.2.1 [270], which supports the Stream API only, was used to monitor the tweets located inside the Barcelona bounding box during 24 hours.
- twitteR 1.1.9 [271], which supports the REST API only, was used to retrieve the tweets around Barcelona from November 8, 2016 to October 31, 2017 (on a weekly basis).
- rtweet[135] 0.4.0 [272], which supports the Stream and REST APIs, was used to retrieve the timeline and list of followers of selected Twitter accounts, but the results were not used in the present research (some were published separately).
- httr 1.3.1 [273] was used for authenticating with the rtweet and twitteR packages.
- ROAuth 0.9.6 [274] was used for authentication when using the streamR package.

Both the Streaming and the REST APIs returned a plethora of metadata beyond the text content of the tweet (Table 3.5), although not all the fields were always available for every single tweet collected:

- The tweet content itself, and its unique ID.
- The user who authored it, his or her unique ID and some profile information (e.g. name, city, country).
- Time, date and timezone when the status was updated.
- Location of the geographic coordinates of the tweet (if the device was geo-enabled).
- Extracted lists of #hashtags, @mentions and URLs.
- Language.
- Type of device (e.g. Mac, iPhone) and client application.
- Metadata to keep track of the interactions between users (e.g. replies, retweets, favorites, quotes).

[134]The Tweepy Python library is available at *http://www.tweepy.org/* at the time of writing.

[135]The author contributed a patch to this package, the pull request was available at the time of writing on GitHub at https://github.com/mkearney/rtweet/pull/28

3.6.4 Retrieved Twitter Data

The Twitter data was collected in two stages (Table 3.3) using two complementary strategies with very different retrieval windows (Table 3.4):

1. Probing during 24 hours the tweets produced inside a bounding box enclosing Barcelona (Table 3.6), using the filter method in the Twitter Stream API.
2. Retrieving the tweets corresponding to the two keywords with highest proportion of tweets that included location information (from the previous stage), using the search method in the Twitter REST API.

In the first phase, the Twitter stream was monitored during 1440 minutes[136] in a bounding box of double the size[137] of the one used to retrieve the Panoramio, Flickr and Intagram data (Table 3.6). This stream contained 21,950 tweets (15.24 tweets per minute on average) of which 2,208 included longitude and latitude coordinates (about 10%). To maximize the number of geotagged tweets, these responses were analyzed to identify the sources with a high tweet volume that also had a significant proportion of location information.

Therefore, the sources "Twitter for Android" and "Twitter for iPhone" were discarded because even though their clients produced a large number of tweets, only a small percentage of them had location data (around 3% and less than 1% respectively).

On the other hand, the sources labeled as "Instagram" and "Foursquare" (through its Swarm[138] app) were selected because among all the sources with a high ratio of geotagged tweets[139] they were the ones with the highest tweet volume (Fig. 3.11).

The second phase began on November 8, 2016 (Table 3.3), and consisted in the weekly[140] collection of Twitter data through its search REST API during a complete year period. The process retrieved an average of roughly 43,000 tweets weekly, among which around 31,000 (72%) were unique[141] (Fig. 3.12). Of the unique weekly tweets, around 10,000 (32%) included location data on average. (Table 3.9).

The query always used the same keywords ("instagram" and "foursquare") and retrieved tweets in batches of less than 18,000[142] with a 15-minute cooldown

[136] Using a persistent connection on a dedicated computer.

[137] And therefore four times the area.

[138] According to the Swarm website, available at *http://www.swarmapp.com/* at the time of writing, "Swarm is a fun, interactive way to keep a record of all the places you go".

[139] In the sample obtained, both Instagram and Foursquare had a 100% score, as all tweets coming from their clients had location information.

[140] Because of the maximum one-week-old-tweet limitation imposed by Twitter.

[141] There was a significant overlap every week.

[142] Because of the limit of 180 requests (which returned a maximum of 100 tweets each) in a 15 minute window.

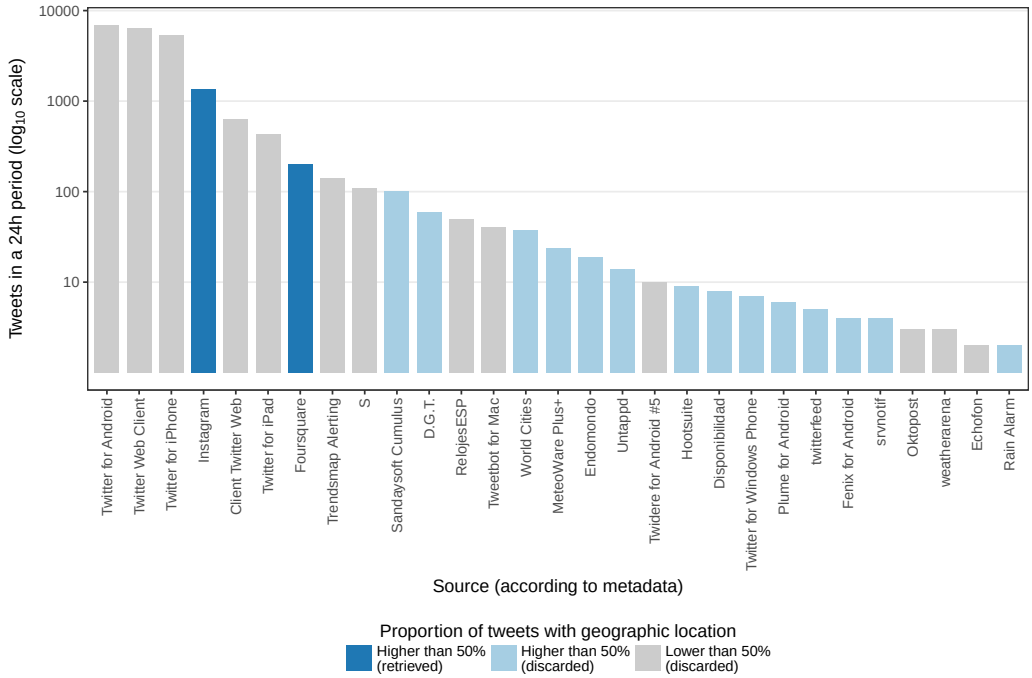


Figure 3.11: Distribution of the 21,950 tweets (in a logarithmic scale) according to their source, retrieved from the Twitter Streaming API in a 24-hour period starting at 13: 38 on October 26, 2016. Of the sources with more than 50% of tweets with location information (shades of blue), the top two were selected to be retrieved weekly during one year (darker blues). Source: Own work based on data retrieved from Twitter through its Streaming API.

Table 3.9: Total volume of tweets and weekly averages retrieved through the Twitter REST API. Of all collected tweets about 72% were unique, and among unique tweets 32% had location data, on average.

	Total	Weekly average
Retrieved tweets	2,250,317	43,128
Unique tweets	1,627,786	31,197
Unique tweets with coordinates	516,990	9,908
Downloaded data	600.6 Mb	11.51 Mb

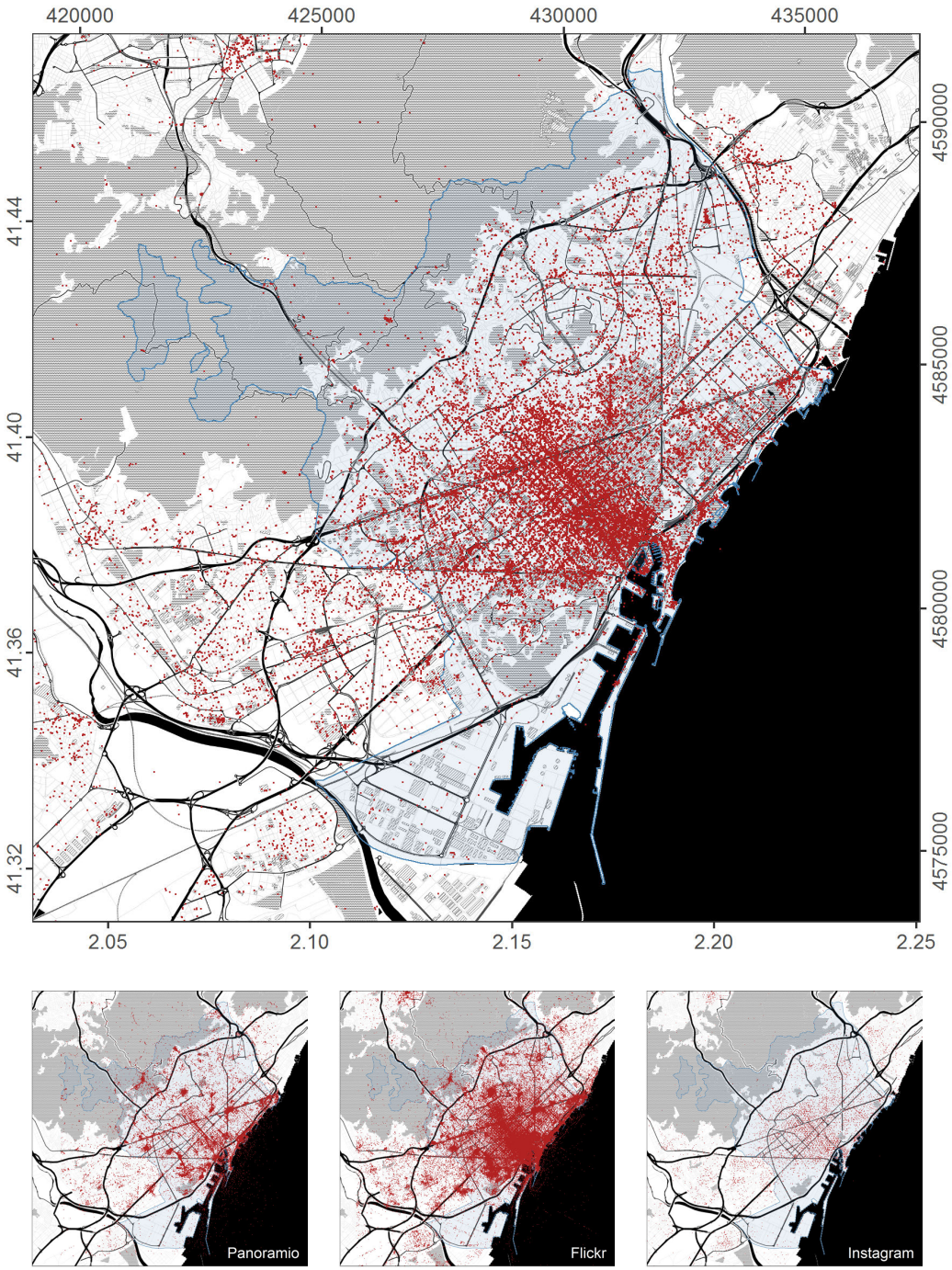


Figure 3.12: Overview of the 516,990 unique locations retrieved using the Twitter API (in red) in a 356-day period, overlaid on the official Barcelona municipal limits according to the ICGC (in faint blue) in the WGS 84 / Pseudo-Mercator – Spherical Mercator (EPSG:3857) projected coordinate system. Bottom miniature maps correspond to the equivalent Panoramio (left), Flickr (center) and Instagram (right) retrieved data. Base map tiles by Stamen Design, under CC BY 3.0, using data from OpenStreetMap, under ODbL.

period after each batch. The tweets were searched within a radius which included Barcelona because the Twitter REST API did not allow to specify a bounding box. This search radius overlapped[143] the bounding box used to retrieve the Panoramio, Flickr and Instagram data (Table 3.6).

3.7 Assessing Retrieved Locations

3.7.1 Location Data Across Services

The analysis of the retrieved sources revealed significant differences in the location data extracted from the each service. One of the factors was the amount of collected observations (Table 3.5), which differed by two orders of magnitude (Fig. 3.13) between service with the highest unique location count (Flickr, discussed in section 3.4) and the one with less retrieved locations (Instagram, discussed in section 3.5).

Besides the quantity and focusing on the qualitative aspects, while the amount of locations retrieved from Twitter (discussed in section 3.6) was remarkable —especially considering the relatively shorter time span (Table 3.4) it covered (Fig. 3.3)—, the quality of the location data it provided (Fig. 3.12) did not match the precision of the picture locations collected from Panoramio (Fig. 3.5) (discussed in section 3.3), despite the former being almost an order of magnitude larger.

3.7.2 Outliers

The large number of samples was expected to balance individual biases through the law of large numbers, relegating small variations to background noise and amplifying the most prevalent collective behaviors. However, users with a large

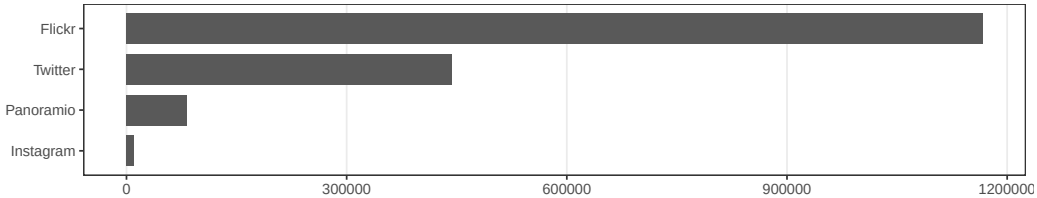


Figure 3.13: Number of unique locations retrieved from the Flickr, Twitter, Panoramio and Instagram services during research. Source: Own work.

[143]The collection methodology was modified on August 22, 2017 to correct a bug in the retrieval code that requested tweets within a slightly misaligned window.

number of images had the potential to distort the results because exceptionally large biases would not be easily canceled-out by the majority, for example:

- Incorrectly geotagged pictures where the location coordinates were the home or workplace of their owner.
- Pictures whose assigned timestamps were the upload date instead of the moment they were taken.
- Cases where users did not adjust the timezone of their picture-taking devices when traveling, and left it in their home location.

While any of the users could randomly be in any of these situations, it was expected that with a sufficient number of samples the aggregate data remain be unaffected. However, if one or more of the most active users had any of this biases, the estimations of the spatial distribution of events and/or the analysis of the temporal behavior of users could be compromised.

Therefore, to avoid this potential distortions, users whose number of pictures of Barcelona was larger than six standard deviations away from the mean picture count of all users were discarded in most analyses. This threshold was chosen because of the similarity with the Six Sigma quality management method to ensure a manufactured good is reasonably free from defects.

3.7.3 Service Comparison

To compare the location data across the researched services, the locations of their geotagged data were overlaid on an ortophoto. The aerial photograph was converted to grayscale and its brightness was increased to provide spatial context without becoming distracting. The locations were made partially transparent[144] to reveal the urban texture underneath.

However, the resulting maps were adequate for interactive exploration but not suitable for a printed document, especially in a small format such as this dissertation. Instead, a graphical summary of the retrieved locations was produced, arranging in a 4 x 3 grid a total of 12 maps of selected landmarks of Barcelona, selected because of their outstanding picture density compared to their immediate surroundings. For comparison purposes, all map tiles were at the same 1:10,000 scale, using the ETRS89 / UTM zone 31N (EPSG:25831) projection.

The comparison matrices were produced for the two services with geotagged images —Panoramio (Fig. 3.14) and Flickr (Fig. 3.15)—, and in the case of Flickr an additional version where the locations were made 90% transparent[145] was

[144] The whole layer was made semitransparent, not the individual points.

[145] And therefore, ten or more overlapping locations were needed to produce a point with the same saturation as the other examples.

also included (Fig. 3.16) to crudely alleviate overplotting (which will be further discussed in section 6.2.1).

In the case of non-picture data, Twitter locations (Fig. 3.17) were also included for comparison but did not reveal any significant pattern in any of the landmarks, suggesting that it was more suitable for broader scales of analysis. Instagram locations (Fig. 3.9) was exuded because it did not provide *event* data but the locations of actual *places* instead.

El Camp Nou The Panoramio data features more pictures in the west stand (under cover) than in the rest of the seating areas, and fewer in the playing field, while the Flickr data increases the amount of locations, which are more spread around the stadium, as well as the surrounding areas.

Rambla de Mar The outline of the bridge that gives access to the Maremàgnum shopping center is revealed in the Panoramio locations, as well as the neighboring Christopher Columbus monument (west), through the accumulation of pictures around the roundabout where it stands. The Flickr data shows the same pattern but the amount of pictures obscures the shapes. In both cases a few locations appear in the harbor waters.

Park Güell The Panoramio data shows a substantial accumulation of pictures at the entrance, specially in the Casa del Guarda and where the Dragon Stairway is located. The perimeter of the main terrace and the path that leads to the Casa Museu Gaudí are also visible. The Flickr data is much more abundant, and shows more clearly the paths taken around the park, as well as an accumulation of pictures in the street that leads to the park (Carrer de Larrard). In both cases the nearby Turó de les Tres Creus is also visible.

Plaça d'Espanya The Panoramio data outlines the shape of the Las Arenas shopping center designed by the London architectural studio Rogers, Stirk, Harbour and Partners, which is much more visible in the Flickr data. Despite being a circular building, it is possible to identify some locations[146] (south and north-west) where the majority of pictures are taken from. Both datasets also show the circular shape of Plaça d'Espanya[147], which provides access to Avinguda de la Reina Maria Cristina, one of the Fira de Barcelona[148] venues.

La Sagrada Família In the Panoramio locations it is visible the (almost) absence of pictures in the pond of Plaça de Gaudí (North-East) and the pedestrian paths in Plaça de la Sagrada Família (South-West). The Flickr locations are much more dense and also provide a hint of the heavy photographed axis

[146] Interestingly, Twitter locations also show an accumulation of locations in the south quadrant.
[147] At the intersection of Paral·lel Avenue and Gran Via de les Corts Catalanes.
[148] Barcelona trade fair, home to the Mobile World Congress at the time of writing.

of Avinguda de Gaudí.

La Pedrera Both the Panoramio and the Flickr locations show many pictures taken *inside* the Casa Milà. From outside, a significant amount of pictures of the facade are taken from the opposite sidewalk of Passeig de Gràcia, which is only possible because the facade of La Pedrera is uncharacteristically not occluded by trees, unlike like the rest of the buildings in this avenue.

Torrre Agbar Now known as Torre Glòries, the brand new skyscraper designed by Jean Nouvel and Fermín Vázquez clearly appears in the Panoramio locations. The Flickr locations also show the same pattern but allows identifying the vantage points from where the tower is photographed. In addition, the Design Museum of Barcelona designed by Oriol Martorell, Oriol Bohigas and David Mackay can also be identified in the Flickr dataset.

Parc del Tibidabo The Panoramio and Flickr locations show the location of the Temple Expiatori del Sagrat Cor as well as the amusement park premises, and reveal the locations where the views of Barcelona from the Tibidabo mountain are preferred.

Castell de Montjuïc This fortress and museum was once the location of the cannons that bombarded the city, and now provides a spot to take pictures of the city skyline, as revealed by the Panoramio and Flickr data, which also reveal the routes taken by the visitors to access the castle through the drawbridge from the uphill paths.

Plaça de Catalunya The locations in this area reveal the movement patterns of people as they move back and forth between Passeig de Gràcia, La Rambla or Avinguda del Portal de l'Àngel, avoiding the unpaved paths, while taking photographs —the majority from the center of the square, but also near the fountain and the Francesc Macià monument—. Interestingly, the Apple store[149] appears as an emerging landmark.

Illa de la Discòrdia The Block of Discord has buildings by four notable *Modernisme* architects (Domènech i Montaner, Gaudí, Puig i Cadafalch and Sagnier), and according to the location data from Panoramio and Flickr, Gaudí is unsurprisingly the most popular. Many of pictures are taken from inside of the Casa Batlló, but in contrast with La Pedrera, most of the pictures of the facade are taken from the same sidewalk as the buildings.

Parc de la Ciutadella The pictures concentrate around the Monumental Waterfall, where the majority of the pictures are taken favoring its longitudinal axis. Some pictures are taken around the central lake[150], along the pedestrian paths and in the axis between the center of Plaça de Joan Fiveller and the main facade of the Catalan Parliament.

[149] Open to the public since July 28, 2012.

[150] Or inside, since small boats are allowed.

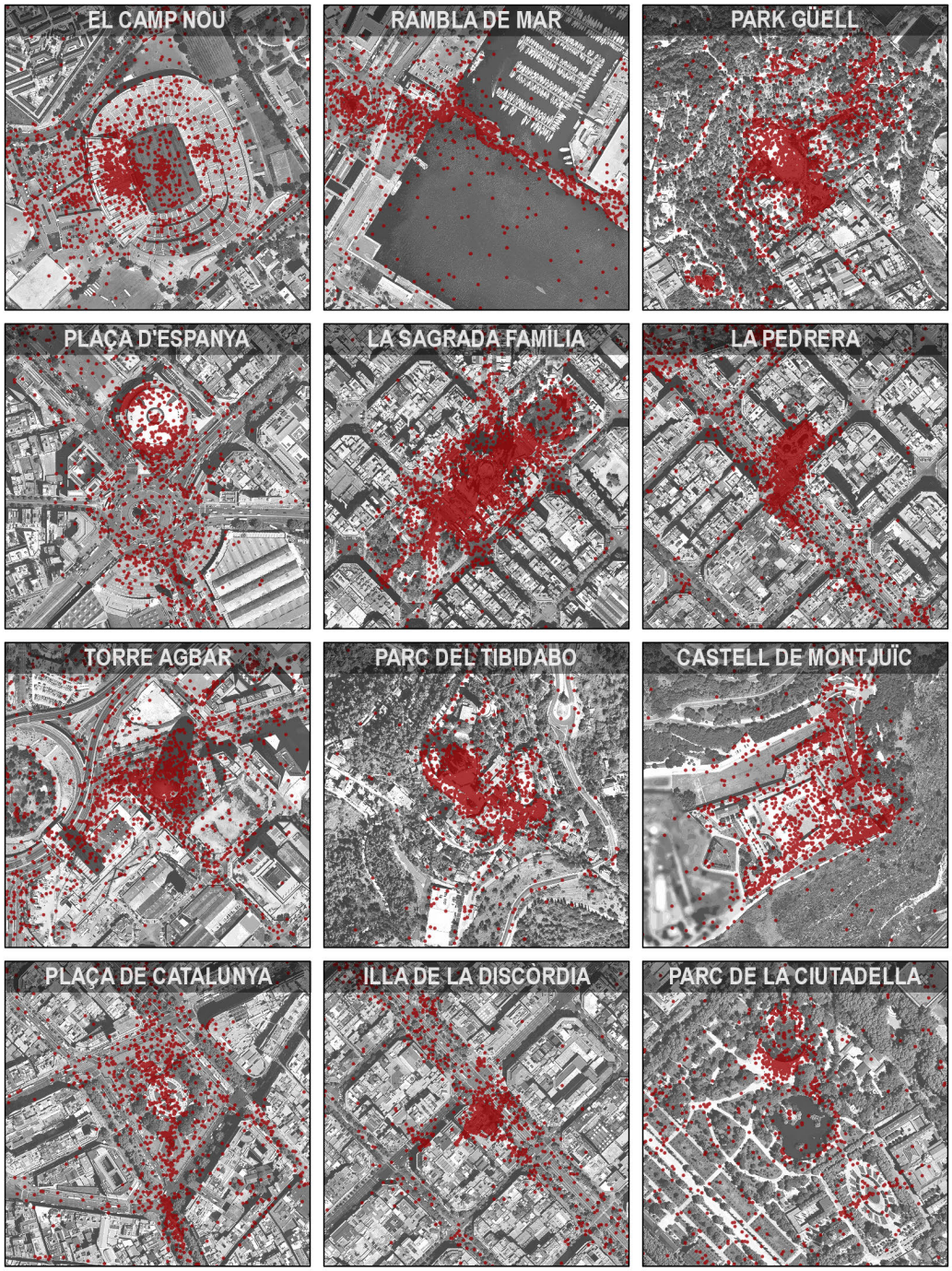


Figure 3.14: Landmarks showing the locations of geotagged pictures retrieved from the Panoramio API (all tiles in the grid at 1:10,000 scale). Base cartography by Cartogràfic i Geològic de Catalunya (ICGC), under CC BY 4.0.

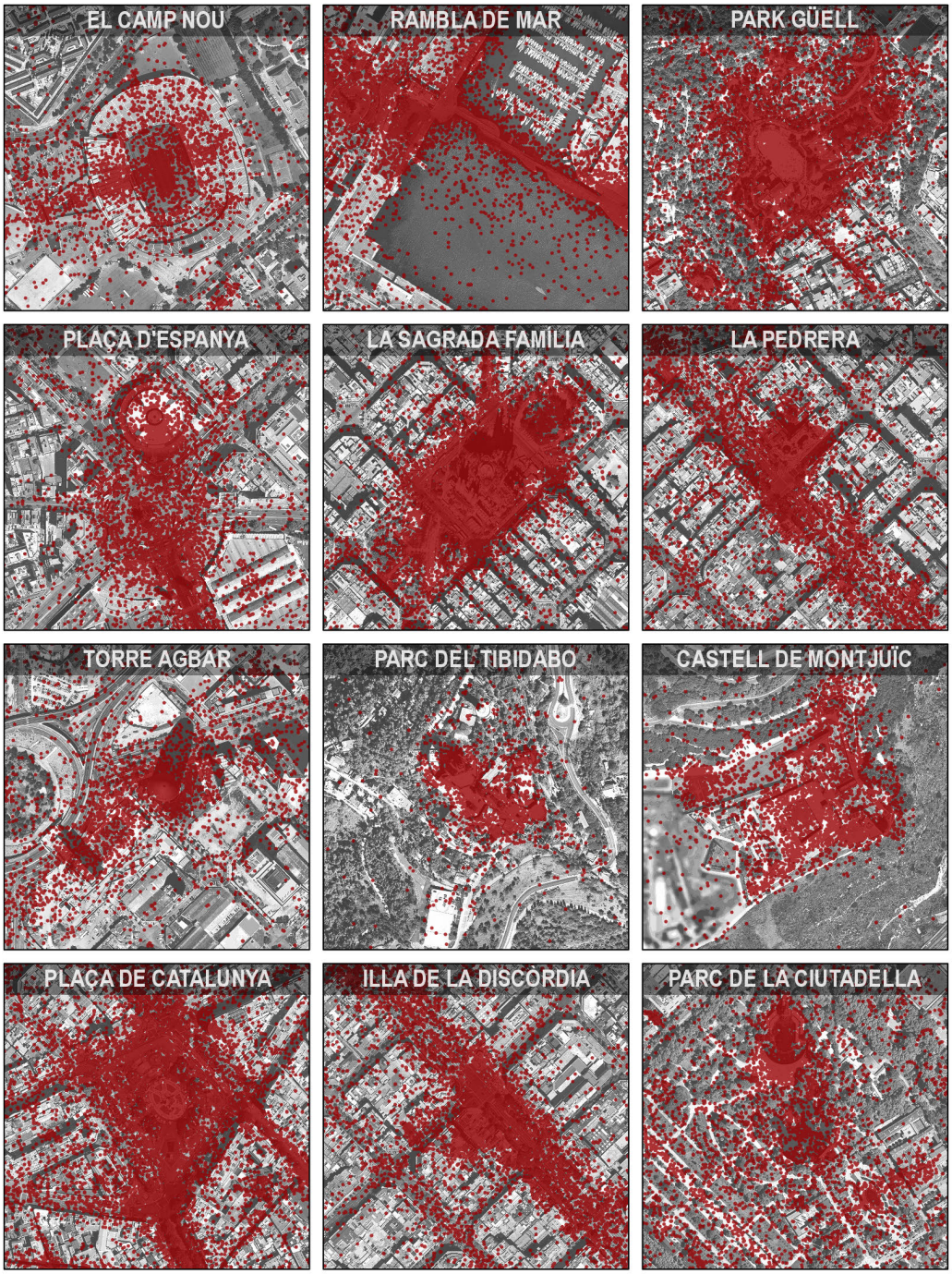


Figure 3.15: Landmarks showing the locations of geotagged pictures retrieved from the Flickr API (all tiles in the grid at 1:10,000 scale). Base cartography by Cartogràfic i Geològic de Catalunya (ICGC), under CC BY 4.0.

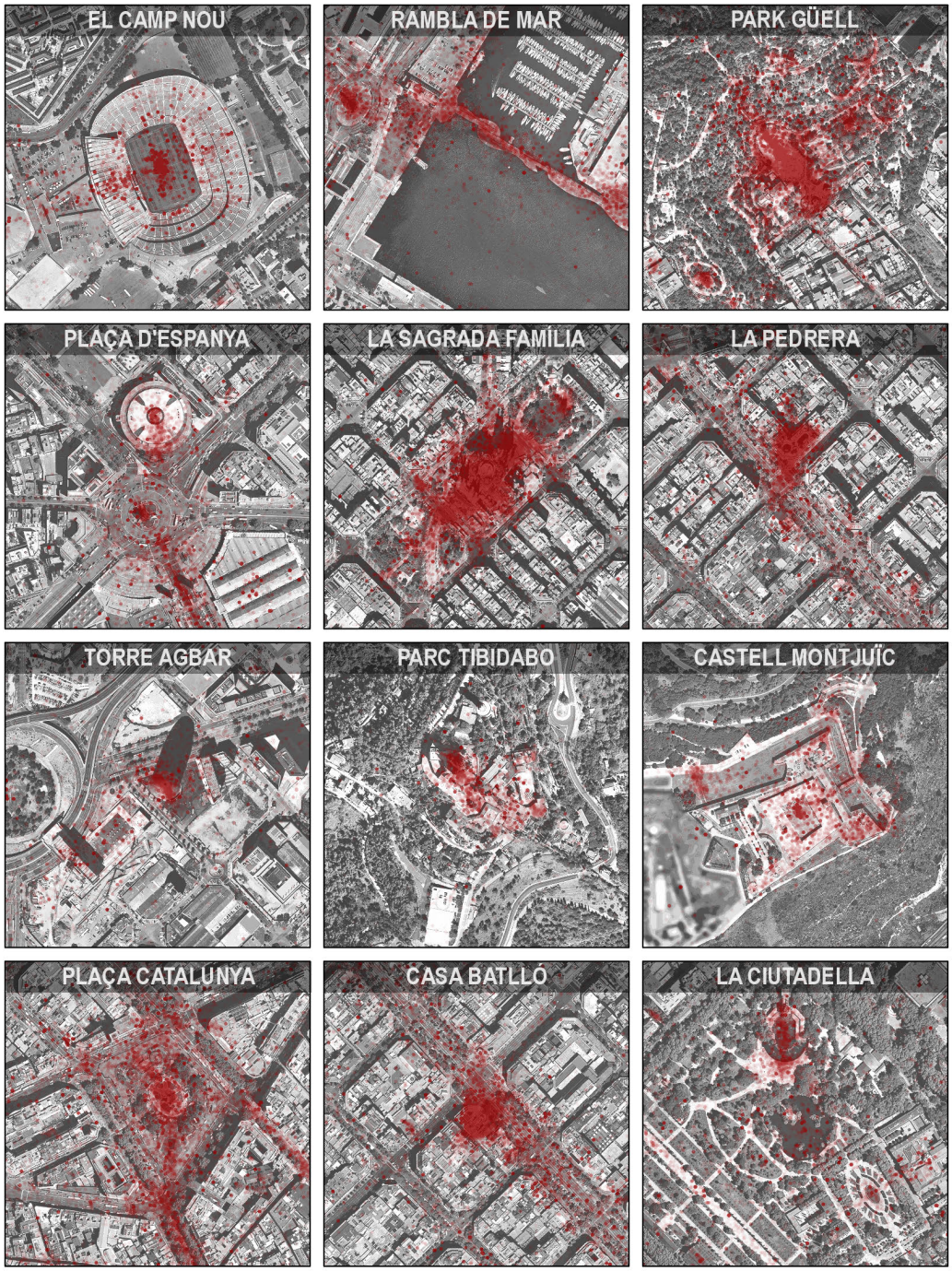


Figure 3.16: Landmarks showing the locations of geotagged pictures retrieved from the Flickr API (all tiles in the grid at 1:10,000 scale). Locations are semitransparent (90%) and full saturation means ten or more overlapping locations. Base cartography by Cartogràfic i Geològic de Catalunya (ICGC), under CC BY 4.0.

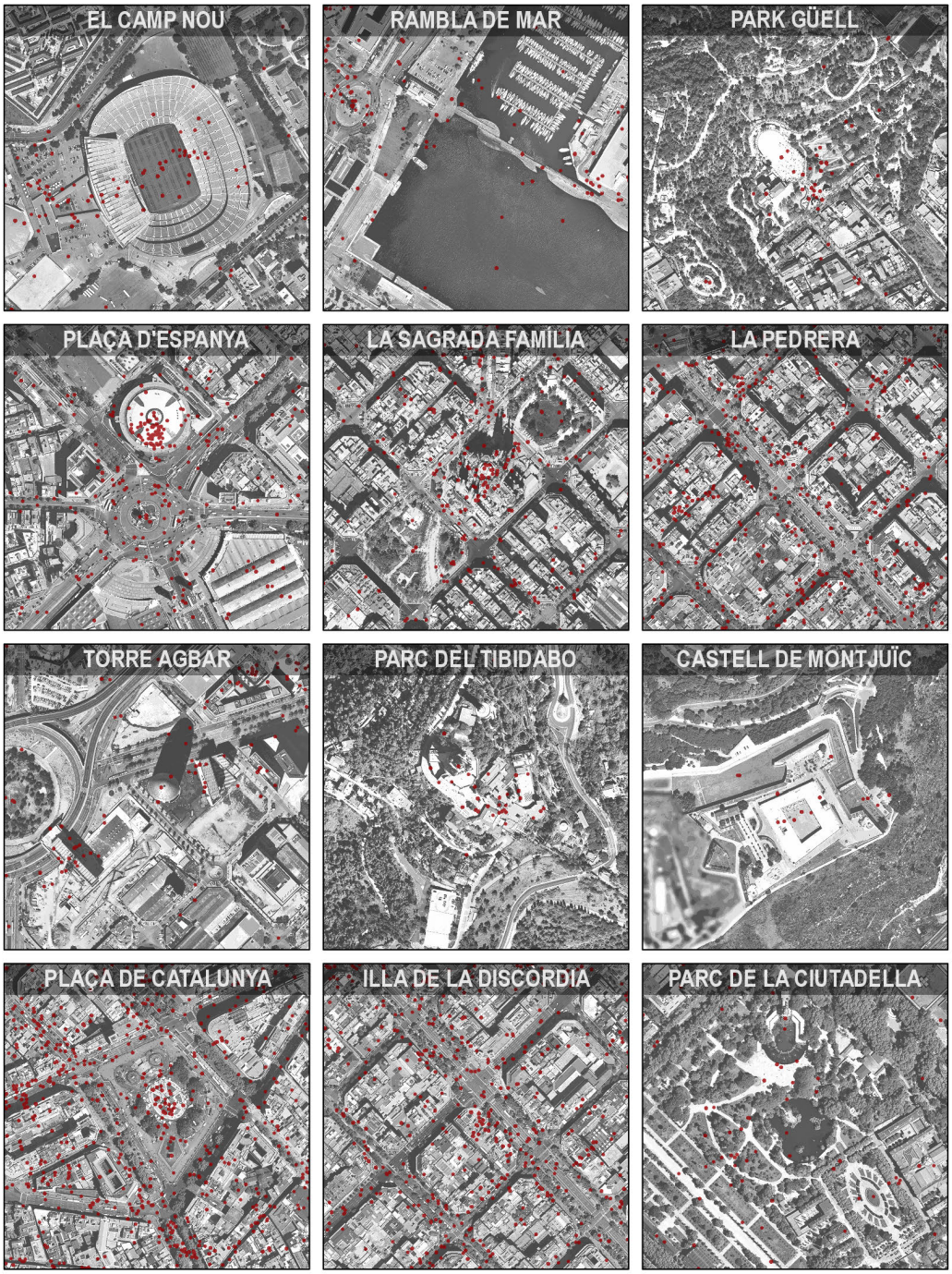


Figure 3.17: Landmarks showing the locations of geotagged status messages retrieved from the Twitter API (all tiles in the grid at 1:10,000 scale). Base cartography by Cartogràfic i Geològic de Catalunya (ICGC), under CC BY 4.0.

3.8 Conclusions

3.8.1 Differences and Similarities

Beyond differences in location accuracy (discussed in section 3.7), data collected from the four analyzed services —discussed in the previous sections 3.3, 3.4, 3.7 and 3.6— significantly different in other aspects, and each of the services had some distinctive features that influenced the volume, velocity, variety and veracity (discussed in section 2.3.2) of data and metadata that was available for retrieval:

- The size of the user base (Table 3.2), or more specifically, the number of *active* users, can limit the number of potential samples available and therefore the signal to noise ratio.
- The *engagement* of the users in the community (measured as the number of minutes per day using the service), can influence the immediacy of the data collected (e.g. Twitter users engaged several times daily while most Flickr users only posted media on special occasions).
- The demographics of the users (gender, age, status, origin) have the potential to bias the results, unless the focus of the research is aligned with this bias, in which case it could become an advantage.
- The service orientation (e.g. social, business) can also bias the results as users can adopt different roles depending on their online activity, and choose their preferred services accordingly.
- The type of media shared (e.g. picture, text) that is the main focus of the service can also play an important role, as it can influence the spatial (on-site or off-site) and temporal (immediate or delayed, frequent or sparse) context of the user when sharing the media.
- Technical limitations of the devices (battery consumption, sensors) and the software they run (background processes, permissions) can determine the capacity to collect some kinds of metadata accurately.

In addition, since the business model[151] of most these services is based on leveraging their data to produce revenue[152] (e.g. through target advertising), they are increasingly reluctant[153] to providing unlimited free access to their digital vaults to third parties, and some features can be unavailable through their API[154], at least without payment (Table 3.1).

[151] Since at the very least have to cover their operating costs.
[152] Hence the expression "If you're not paying for the product, you are the product."
[153] Furthermore, the access conditions can change as the result of acquisitions, pivoting to more profitable services to ensure their economic viability, initial public offerings (IPOs), etc.
[154] Although in some cases the reason is the lack of maturity of its implementation.

3.8.2 Representation and Interpretation

There does not seem to exist a one-size-fits-all source of data capable of answering all possible research questions, especially considering the complexity of urban research, as each source can have a unique intertwined combination of advantages and disadvantages, depending on the perspective of the research subject. However, these issues can be somewhat alleviated adopting a multi-faceted approach, taking advantage of several complementary data sources instead of a single one.

The volume and diversity of the retrieved data brought important challenges beyond the scope of GIS technologies currently used in urban analysis [95]. Over the next chapters, multiple approaches are developed within this new urban data paradigm, focusing on multiple scales, topologies and dimensions:

- A Global Perspective (chapter 5)
- The City Scale (chapter 6)
- The Neighborhood Scale (chapter 7)
- A Network Approach (chapter 8)
- The Temporal Dimension (chapter 9)

Chapter 4

Determining User Origins

> "I will tell the story as I go along of small cities no less than of great. Most of those that were great once are small today; and those that in my own lifetime have grown to greatness, were small enough in the old days"
>
> Herodotus in "The Histories"

4.1 Introduction

4.1.1 Background

The spatial distribution of digital traces left by users of online communities is one of the research objectives of this dissertation, and chapter 3 focused on discussing the methodology to collect digital traces left on Barcelona by the users of different social media services, using their respective APIs to obtain the desired data.

The processes described were largely asymmetric, essentially querying the services as if they were databases, using relatively simple queries[1] which returned large volumes of data. Hitherto, the strategies revolved on the best approaches to maximize the amount of collected data, building these queries intelligently to avoid the API limitations and collect the most representative data for research.

This chapter discusses a fundamentally different use of a Web APIs which employs the opposite mechanism, and instead of a simple query that returns many

[1] The queries could be informally translated into "retrieve all available location data in Barcelona".

records with a relatively simple data structure, places many queries that retrieve a relatively small amount of data with a complex structure, which needs to be properly parsed and aggregated.

In the multi-scale approach discussed in section 3.8.2, the broadest possible scale —corresponding to the global scale— is discussed in 5. However, the analyzed spatial distribution is not of the pictures taken in Barcelona but of the people who took these pictures.

4.1.2 Chapter Outline

Before the origin of visitors could be analyzed, it was necessary to convert their locations into geographic coordinates and classify them into geographic units. This chapter describes the methodology of this conversion, and it is divided in the following sections:

- Section 4.2 details the process to collect the profile data from the users who took a picture of Barcelona and posted it on Flickr.
- Section 4.3 introduces geocoding services and how these services can convert a location name to a coordinate pair and/or administrative units, querying a database through a Web API.
- Section 4.4 discusses the process to geocode the hometowns of the Flickr users and the results obtained.
- Section 4.5 describes the criteria to define the categories into which the geocoded coordinates can be classified.
- Section 4.6 explains the aggregation process to convert the locations into the categories defined in section 4.5.

4.2 Source Data

4.2.1 User Location Data

All three services from which it was possible to download media —Panoramio, Flickr and Twitter, but not Instagram— included in their metadata the unique identifier of the account the data originated from, which in some cases made possible to extract information about the account holder's location (obviously not the current real-time location, but the stated place of residence according to the records of the corresponding service[2]).

[2]The location is generally filled in during the registration process.

Panoramio All picture records (discussed in section 3.3) included in their metadata a link to the URL of the corresponding user's profile, but the profile page did not include the user's location information. Its API did not provide a method to access this information either.

Flickr The retrieved pictures from Flickr (discussed in section 3.4) contained the unique identifier of the owner of the picture (the person who uploaded the media). The Flickr API allowed requesting the public information of any profile through this user identifier, which in many cases included location information.

Instagram This service (discussed in section 3.5) did not provide access to neither media information nor user metadata outside of its sandbox mode.

Twitter All retrieved tweets (discussed in section 3.6) included the screen name of the user who wrote it. The Twitter API allowed retrieving the information from any profile through this screen name. Besides one of the pieces of information prominently featured in the user's profile page is his or her location (provided that this information is public).

Among the two services from which the location of their users was available (Flickr and Twitter), Flickr was selected in this research because its users appeared to state more truthfully their actual location, the provided location data was more detailed (Table 3.5), and it covered a substantially longer time span (Table 3.4).

4.2.2 Flickr User Profile Data

The 1,166,704 unique records of picture data retrieved on March 21–22, 2017 contained 34,283 unique user identifiers (Table 3.5). Each of these user identifiers were queried in the period between March 23–25, 2017 (Table 3.3) using three different methods (Table 4.1):

- The *flickr.people.getInfo* method of the Flickr API.
- The *flickr.profile.getProfile* method of the Flickr API.
- Retrieving information from the raw HTML code of each corresponding user profile pages, through the link obtained in the results of the *flickr.people.getInfo* API method.

After iterating through the list of all unique users, the *flickr.people.getInfo* method[3] of the Flickr API returned 34,279 records[4] with 23 variables each. Among the returned data, three pieces of information were related to the user location (Table 4.1):

[3]The *flickr.people.getInfo* method is described in detail at *http://www.flickr.com/services/api/flickr.people.getInfo.html* at the time of writing.

[4]As 4 users had left the service or changed their profiles to private in the two-day period between picture retrieval and user data retrieval.

Table 4.1: Flickr user location-related data obtained using the three discussed retrieval methods.

Name	Type	Method	Valid records	% valid
location	place	flickr.people.getInfo	14,720	43%
timezone	metadata	flickr.people.getInfo	24,687	72%
profileurl	link	flickr.people.getInfo	34,279	~100%
hometown	place	flickr.profile.getProfile	10,456	30%
city	place	flickr.profile.getProfile	12,969	38%
country	place	flickr.profile.getProfile	13,302	39%
hometown	place	HTML scrapping	12,470	36%
currently	place	HTML scrapping	14,722	43%

- Location, formatted as "{City}, {Country}".
- Timezone information, composed of three elements:
  - Label[5]
  - Offset[6] from the Coordinated Universal Time (UTC)
  - Unique timezone identifier[7]
- A URL pointing to the page with the public profile page of the user, which also contained location data among other information.

Alternatively, iterating over the list of user identifiers, the *flickr.profile.getProfile* method[8] of the Flickr API returned 18 variables, of which three contained location-related information. Of all the 34,283 unique users, 15,246 records (44%) had at least one of the following three pieces of information (Table 4.1): Hometown, City or Country.

Finally, the content of the public profile pages was scraped using the R package rvest [275], iterating through all the pages linked in the profile URLs obtained in the previous steps[9] and extracting the desired information —some of which was location-related (Table 4.1)— from their HTML code using XPath. The retrieved information was stored as elements of a description list[10]:

- Hometown.
- Place of current residence, formatted as "{City}, {Country}".
- Sex (male, female, other or unknown).

[5]For example a label would appear as "Brussels, Copenhagen, Madrid, Paris".
[6]For example an offset would appear as "+01:00".
[7]For example a timezone identifier would appear "Europe/Brussels".
[8]The *flickr.profile.getProfile* method is described in detail at *http://www.flickr.com/services/api/flickr.profile.getProfile.html* at the time of writing.
[9]From the results of the *flickr.people.getInfo* API method.
[10]Elements (`<dt>` and `<dd>` tags) inside a `<dl>` tag.

- Relationship status (open, taken, single or unknown).

4.3 Geocoding Services

4.3.1 Geocoders

Geocoding is the process of transforming the textual description of an address —complete or partial— into the geographic location of a point on the surface of the Earth in longitude and latitude coordinates. The opposite process, converting a geographic coordinate into an address or place name, is called reverse geocoding.

The software component that performs this operation is called a geocoder. The specific requirements of geocoders, which operate on massive global data sets —that must be kept up-to-date—, but ingest and return relatively small-sized chunks of data, makes them suitable to be implemented as web services in a client-server model.

The accuracy of the geocoding process depends on the nature of the query, the comprehensiveness of the database, and the specific details of the software implementation. During the research, three geocoders with global coverage were tested:

- The Google Maps Geocoding API.
- The Data Science Toolkit[11] (DSK) Google-style geocoder, which takes the same arguments and returns the same data structure as the former[12].
- The Open Street Map Nominatim API[13].

4.3.2 The Google Maps Geocoding API

The Google Maps Geocoding API[14] provides geocoding and reverse geocoding functionality through an HTTP interface. The standard usage of the free[15] geocoder has the following limitations[16]:

[11] Available as a service or as Virtual Machine at *http://www.datasciencetoolkit.org/* at the time of writing.

[12] The emulation details are available at *http://www.datasciencetoolkit.org/developerdocs* at the time of writing.

[13] Details on Nominatim are available at *http://wiki.openstreetmap.org/wiki/Nominatim*

[14] The Google Maps Geocoding API details are available at *http://developers.google.com/maps/documentation/geocoding/* at the time of writing.

[15] The paid Google Maps APIs Premium Plan offer higher usage limits of 100,000 requests per 24 hours.

[16] Details on the limits of the Google Maps Geocoding API are available at *http://developers.google.com/maps/documentation/geocoding/usage-limits* at the time of writing.

- 2,500 requests per day[17].
- 50 requests per second.

Like the majority of web APIs, the response is JSON or XML formatted (Table 3.1). The response returns two root elements, "status" —with the status code of the request[18]— and "results" —a variable-length array[19] of geocoded addresses and geometry information—. *Each* of the returned results in the array can contain:

- An array of address component objects of different types:
  - street number
  - locality
  - intermediate administrative and/or political divisions
  - country
  - postal code
- The complete human-readable address assembled from the address components.
- The geometry information:
  - location in geographic coordinates (latitude and longitude)
  - bounding box of the corresponding geometry (coordinates of its left, right, top and bottom edges)
  - method through which the coordinate was approximated[20]

4.4 Geocoding Methodology

4.4.1 Geocoder Results

The 10,908 unique user locations —combinations of city and country— were geocoded using the Google Maps Geocoding API on April 20–24, 2017 using a custom script developed from scratch in the R programming language (Table 3.3). The queries returned between zero and ten elements in the array of address components (Table 4.2), although the majority had only one single element.

These address components included a total of 23 different address component types (Table 4.3), from which the data of interest could be extracted. After the geocoding process, the following address component types were selected for each unique location, which were subsequently transferred to the corresponding user —and through the transitive relation, to his or her pictures— for further analysis:

[17] And at the time of writing, $0.50 per 1,000 additional requests, up to 100,000 daily.
[18] Which is "OK" if no errors occurred.
[19] It can also be a single-element array, or even an empty array.
[20] For example, it can be an exact street address (parcel level), interpolated along a road segment, or the geometric center of a larger element (e.g. an administrative division polygon).

Table 4.2: Number of addresses (elements in the address component arrays) returned by the Google Maps Geocoding API for each of the 10,889 unique locations queried.

Array elements	Count	Percent
0	331	3.04%
1	10,344	94.99%
2	147	1.35%
3	19	0.17%
4	14	0.13%
5	10	0.09%
6	5	0.05%
7	1	0.01%
8	6	0.06%
9	3	0.03%
10	9	0.08%

- Country.
- Administrative Level 1[21].
- Locality.
- Coordinates (latitude and longitude).
- Bounding boxes around the coordinates.

While not all administrative unit data was available for each and every returned location, their corresponding approximate bounding boxes and geographic coordinates were always returned by the geocoder. The coordinates were later spatially joined to the political and administrative boundaries cartography (discussed in section 4.5), while the bounding boxes provided a measure of the location uncertainty of each result (Fig. 4.1).

4.4.2 Geocoding Caveats

The original location data consisted in two pieces of information (localities and countries). This information was sometimes incomplete, in some cases being available just the locality or the country, but not both. Fortunately the geocoder was able to successfully discriminate what part of the address information it was dealing with.

Additionally, since the scope of the geocoded addresses was global, there were some potential issues with character encoding (e.g. location names spelled in

[21]In the case of Spain, it corresponds to the Autonomous Communities.

Table 4.3: Types of returned address component types returned by the Google Maps Geocoding API (queries returned between 0 and 10 address components) in alphabetical order.

Address component type	Count	Percent
administrative_area_level_1	9,847	21.91%
administrative_area_level_2	8,335	18.54%
administrative_area_level_3	1,683	3.74%
administrative_area_level_4	388	0.86%
administrative_area_level_5	1	<0.00%
airport	24	0.05%
bus_station	2	<0.00%
campground	2	<0.00%
colloquial_area	112	0.25%
continent	1	<0.00%
country	10,543	23.46%
establishment	113	0.25%
floor	4	0.01%
locality	8,992	20.00%
neighborhood	154	0.34%
political	288	0.64%
postal_code	2,774	6.17%
postal_code_suffix	22	0.05%
postal_town	755	1.68%
premise	30	0.07%
route	523	1.16%
street_number	335	0.75%
subpremise	21	0.05%

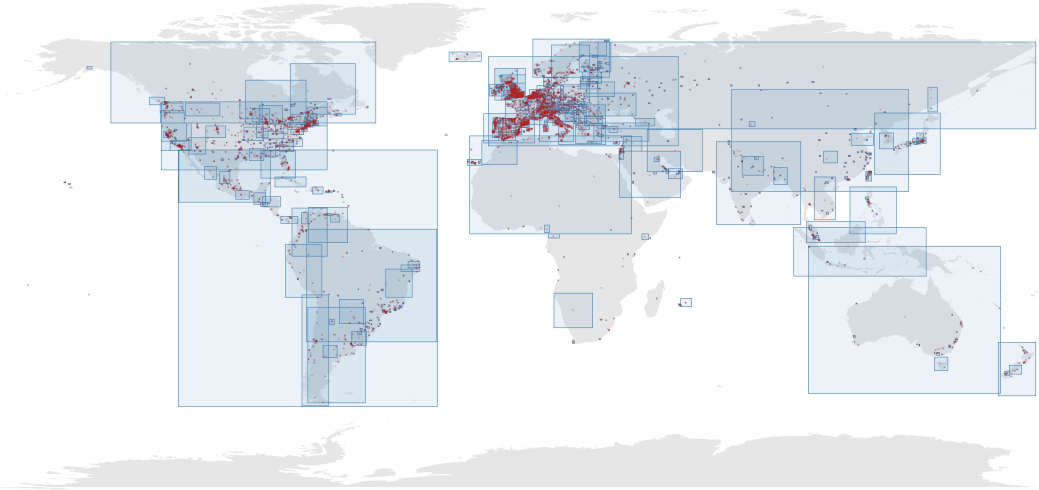


Figure 4.1: Map of the 5,693 unique geographic locations of Flickr users with geotagged pictures of Barcelona. Red dots are the approximate coordinates of the results, and blue boxes are their corresponding bounding boxes. Map uses the Plate Carrée Equirectangular projection (EPSG:4326 / EPSG:32663). Source: own work, with base cartography from Natural Earth, in the public domain.

CJK[22] or Cyrillic). Fortunately, the geocoding engine was able to handle queries encoded in UTF-8 automatically, and returned the translated place names in English, and was even capable of handling most misspellings.

Another issue was the ambiguity of queries that matched multiple places sharing the same name (e.g. "Guadalajara", which can be eithr a city[23] in Spain or in Jalisco, Mexico). To mitigate this issue, the query was region biased, and the script instructed the geocoder to search locations preferably in Spain, where the users were more likely to originate from.

The ambiguity of the query could also cause the geocoder to return multiple hits (Table 4.2) when different administrative levels had the same name[24] (e.g. "New York" is a city as well as a State, and "Barcelona" can be the name of either a city or a province). In these cases, since the geocoder appeared to return the candidates sorted according to their likelihood —as determined by the system—, the first result was always selected.

Finally, some locations in disputed territories did return a coordinate and a city name, but without a country assigned. In these cases, some information (such as the continent) had to be assigned through the location of their coordinates within

[22] CJK is a collective term for the Chinese, Japanese, and Korean languages.
[23] And an Autonomous Community of the same name.
[24] Street names can be named after cities as well.

the geometry of the administrative and political boundaries of the cartography.

4.4.3 Conflict Resolution

The origin data was scattered over six fields from three different sources (Table 4.1) that were not always available. When there was a clash, the source to determine the origin of the users was selected in descending order according to the amount data provided per source, obtaining at the end of the process the origins of 17,146 users[25]:

1. The concatenation of "city" and "country" from *flickr.profile.getProfile*, obtaining the origins of 14,721 users (the "location" from *flickr.people.getInfo* was a subset of these data).
2. The "hometown" in the scraped profile page, obtaining the origins of 12,470 users (the "hometown" from *flickr.profile.getProfile* was a subset of these data).
3. The "hometown" from *flickr.profile.getProfile*, obtaining the origins of 10,456 users.

4.5 Categorization

4.5.1 Categorization Criteria

The origin information consisted on a pair of geographic coordinates with a bounding box corresponding to the viewport extent required to visualize the complete location (with sizes ranging from a neighborhood to a whole continent). Almost all the user origins included at least a locality and a country, but some of them lacked one, or exceptionally both.

The total number of unique countries was 137, and the unique localities were 2,929. This data allowed analyzing the users' origin by country or by locality, but had to be categorized into larger geographic units for further analysis (discussed in chapter 5). Since the data was from visitors of a single location —the case of study, Barcelona—, the locations were categorized according to increasingly larger geographic rings around the city.

Municipal The administrative limits of the city of Barcelona, as the innermost ring or kernel.

Metropolitan It was defined as the municipalities that belong to the Àrea Metropolitana de Barcelona (Barcelona Metropolitan Area), excluding the city of Barcelona itself (i.e. the municipal ring).

[25]The three sources overlapped partially.

Regional Included the municipalities within the 7 "comarques" (regions) of the Àmbit Metropolità de Barcelona (and proposed Vegueria de Barcelona), excluding the ones that belong to the Àrea Metropolitana de Barcelona (i.e. the metropolitan ring).

Autonomic It was defined as the Autonomous Community of Catalonia, excluding the Àmbit Metropolità de Barcelona (i.e. the regional ring).

National The Kingdom of Spain, including overseas territories, but excluding the Autonomous Community of Catalonia (i.e. the autonomic ring).

European The nations included in the Schengen Area, but excluding Spain (i.e. the national ring).

International The locations that do not belong to any of the previous regions (i.e. the European ring).

4.5.2 Cartographic Data Sources

Political and/or administrative boundaries have historically been object of dispute, from the country level down to the municipal level. There are few cartographic data sources that maintain an up-to-date cartography of these boundaries, specially at the global level (countries) or local level (municipalities). The following sources were used as cartographic overlays to categorize the geocoded locations according to the region they lied within (Table 4.4).

Global boundaries For the worldwide administrative and political divisions, the public domain Natural Earth[26] data was used, accessed through the R package rnaturalearth 0.1.0 [276]. The Natural Earth public domain[27] database provides vector and raster cartographic data at 1:10m (detailed), 1:50m (medium), and 1:110m (broad) scales.

Municipal boundaries Used for the municipalities in the Autonomous Community of Catalonia, they were based in the working file of municipal limits elaborated by the Institut Cartogràfic i Geològic de Catalunya (ICGC). These boundaries were based on the administrative divisions cartography[28] at 1:5,000 scale.

[26] The Natural Earth website is available at *http://www.naturalearthdata.com/* at the time of writing.

[27] The primary authors, Tom Patterson and Nathaniel Vaughn Kelso, and all other contributors renounce all financial claim to the maps according to *http://www.naturalearthdata.com/about/terms-of-use/* at the time of writing.

[28] Territorial organisation of Catalonia in municipalities, counties, vigeriates and provinces, available at *http://www.icgc.cat/en/Public-Administration-and-Enterprises/Downloads/Geoinformation-layers/Administration-boundaries* at the time of writing.

Table 4.4: References used to determine the geographic scope of each category.

Category	Geographic unit	Cartography	Geocoder
Municipal	City of Barcelona	ICGC	Locality
Metropolitan	Àrea Metropolitana de Barcelona	ICGC	-
Regional	Àmbit Metropolità de Barcelona	ICGC	-
Autonomic	Catalonia	ICGC	Level 1
National	Kingdom of Spain	Natural Earth	Country
European	Schengen Area	Natural Earth	-
International	Rest of the World	Natural Earth	-

4.5.3 Geographic Units

Using the categorization criteria discussed in section 4.5.1, seven regions were defined from the two base cartographic layers (Table 4.4), assigning an attribute to each polygon denoting their inclusion into the corresponding geographic unit —centered on Barcelona—, where the larger units did not include the area included the smaller ones.

Barcelona The municipal level was defined using the multi-part feature class in the ICGC cartography corresponding to the code[29] "080193", which matched the geocoded address[30] component type whose locality[31] had the value of "Barcelona". It appears in the map (Fig. 4.2) shaded in beige.

Àrea Metropolitana de Barcelona The metropolitan level was defined as the set of municipalities as listed (Table 4.5) in the Àrea Metropolitana de Barcelona (AMB) official website[32], excluding Barcelona. It appears in the map (Fig. 4.2) surrounding Barcelona as the lightest shade of blue .

Àmbit Metropolità de Barcelona The regional level was defined as the seven regions (comarques) included in the Àmbit Metropolità de Barcelona and proposed Vegueria de Barcelona (Table 4.6), excluding the municipalities that belong to the AMB. It appears surrounding the Àrea Metropolitana de Barcelona in a darker shade of blue (Fig. 4.2).

Catalonia The autonomic level was defined as the rest of the ICGC municipal base not included in any of the previous geographic units, which corresponded to the geocoded responses where the value of "Administrative

[29] According to the Catalan Government statistics bureau (idescat) available at *http://www.idescat.cat/emex/?id=080193* at the time of writing.

[30] The geocoder also recognized variants such as "BCN".

[31] Barcelona itself, a place within Barcelona, or (very rarely) a postal address in Barcelona.

[32] The metropolitan municipalities list is available at *http://www.amb.cat/web/area-metropolitana/municipis-metropolitans* at the time of writing.

Area Level 1" matched the regular expression "Catal.*a"[33]. It appears in the map (Fig. 4.2)as the darkest shade of blue.

Spain The national level was defined as the locations where the locations were inside a polygon of the Natural Earth cartography whose ISO 3166-1 alpha-3 code was "ESP", which were the same whose country address component in the geocoder response was "Spain".

Schengen Area The European level was defined as the nations that belong to one of the 26 states (Table 4.7) of the Schengen Area[34] (at the time of writing) —which allow the free and unrestricted movement of people, goods, services, and capital—, but excluding Spain. In the map (Fig. 4.3), the countries currently in the Schengen Area appear in shades of green, while the rest of the EU countries appear in shades of orange. Some territories in Denmark[35], France[36], Netherlands[37], Norway[38] and Spain[39] have special statuses.

Rest of the World The largest geographic unit was defined as the locations not included in any of the areas discussed above.

4.6 Aggregation Levels

4.6.1 Defined Levels

Each of the unique user origins were assigned different categories to group them into larger geographical units for analysis. The categories were the following, in roughly ascending order from the smallest to the largest:

City Obtained from the locality in the geocoder result. In the case of multiple coordinates sharing the same locality (e.g. "Barcelona" or "BCN", "Poblenou" and "Diagonal 649" would produce the same locality with different coordinates), the latitude and longitude coordinates were assigned by consensus, selecting the most frequent pair, for each city name and country combination, to avoid grouping Barcelona, Spain with its namesake Barcelona, Venezuela together (Table 4.8).

[33]The regular expression matches either "Catalonia", "Catalunya" or "Cataluña".

[34]The list of countries in the Schengen Area is available at the Wikipedia entry at *http://en.wikipedia.org/wiki/Schengen_Area* at the time of writing.

[35]Excluded Greenland and the Faroe Islands.

[36]Excluded overseas departments and collectivities.

[37]Excluded Aruba, Curaçao, Sint Maarten and the Caribbean Netherlands.

[38]Excluded Svalbard.

[39]With special provisions for Ceuta and Melilla.

Table 4.5: List of the 36 municipalities in the Àrea Metropolitana de Barcelona (including Barcelona itself) according to their official AMB website, sorted according to their 2016 population.

Name	Comarca	Population (2016)
Barcelona	Barcelonès	1,608,746
L'Hospitalet de Llobregat	Barcelonès	254,804
Badalona	Barcelonès	215,634
Santa Coloma de Gramenet	Barcelonès	117,153
Sant Cugat del Vallès	Vallès Occidental	88,921
Cornellà de Llobregat	Baix Llobregat	86,072
Sant Boi de Llobregat	Baix Llobregat	82,402
Viladecans	Baix Llobregat	65,779
Castelldefels	Baix Llobregat	64,892
El Prat de Llobregat	Baix Llobregat	63,457
Cerdanyola del Vallès	Vallès Occidental	57,543
Gavà	Baix Llobregat	46,266
Esplugues de Llobregat	Baix Llobregat	45,733
Sant Feliu de Llobregat	Baix Llobregat	44,086
Ripollet	Vallès Occidental	37,648
Sant Adrià de Besòs	Barcelonès	36,496
Montcada i Reixac	Vallès Occidental	34,802
Sant Joan Despí	Baix Llobregat	33,502
Barberà del Vallès	Vallès Occidental	32,832
Sant Vicenç dels Horts	Baix Llobregat	27,961
Sant Andreu de la Barca	Baix Llobregat	27,434
Molins de Rei	Baix Llobregat	25,359
Sant Just Desvern	Baix Llobregat	16,927
Corbera de Llobregat	Baix Llobregat	14,168
Badia del Vallès	Vallès Occidental	13,482
Castellbisbal	Vallès Occidental	12,277
Montgat	Maresme	11,621
Pallejà	Baix Llobregat	11,348
Cervelló	Baix Llobregat	8,861
Tiana	Maresme	8,553
Santa Coloma de Cervelló	Baix Llobregat	8,073
Begues	Baix Llobregat	6,736
Torrelles de Llobregat	Baix Llobregat	5,933
El Papiol	Baix Llobregat	4,075
Sant Climent de Llobregat	Baix Llobregat	4,024
La Palma de Cervelló	Baix Llobregat	3,000

Table 4.6: List of the seven regions (comarques) included in the Àmbit Metropolità de Barcelona, and their corresponding number of municipalities within and outside the AMB.

Region	Municipalities	Within AMB	Outside AMB
Alt Penedès	27	-	27
Baix Llobregat	30	22	8
Barcelonès	5	5	-
Garraf	6	-	6
Maresme	30	2	28
Vallès Occidental	23	7	16
Vallès Oriental	39	-	39

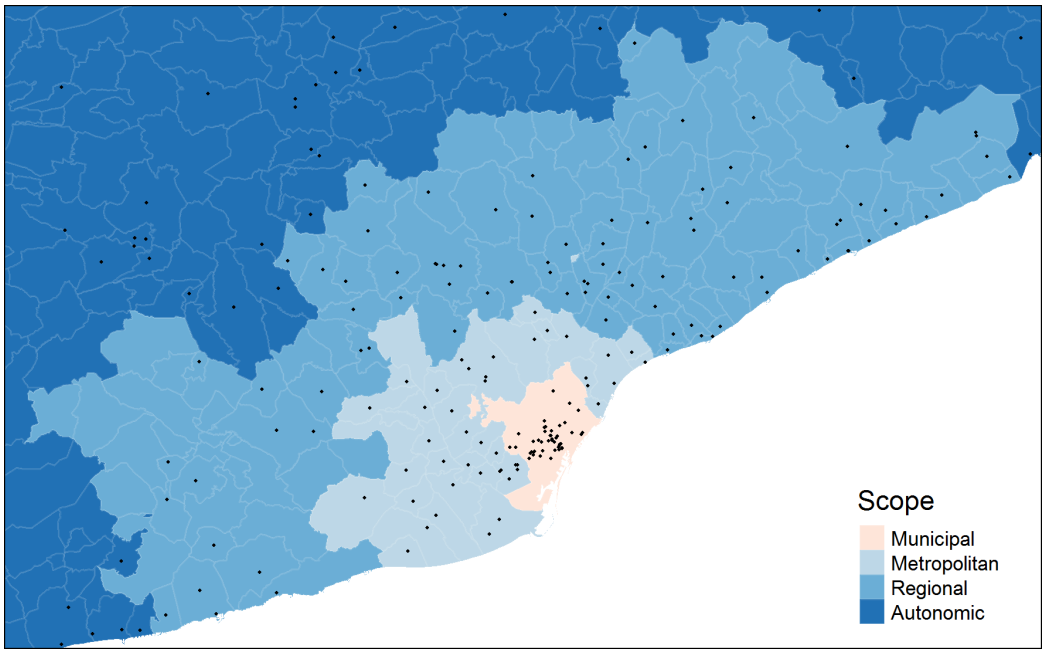


Figure 4.2: Map of the nearest geographical units defined around Barcelona, with distinct Flickr user locations overlaid as black dots. Map uses the ETRS89 UTM zone 31N projection (EPSG:25831). Source: own work based on IGCG cartography, under CC BY 4.0.

Table 4.7: Lists of the 26 nations in the Schengen Area which allow the free and unrestricted movement of people, goods, services, and capital.

Name	ISO-alpha3 code	EU Membership
Austria	AUT	EU
Belgium	BEL	EU
Czech Republic	CZE	EU
Denmark	DNK	EU
Estonia	EST	EU
Finland	FIN	EU
France	FRA	EU
Germany	DEU	EU
Greece	GRC	EU
Hungary	HUN	EU
Iceland	ISL	Non-EU
Italy	ITA	EU
Latvia	LVA	EU
Liechtenstein	LIE	Non-EU
Lithuania	LTU	EU
Luxembourg	LUX	EU
Malta	MLT	EU
Netherlands	NLD	EU
Norway	NOR	Non-EU
Poland	POL	EU
Portugal	PRT	EU
Slovakia	SVK	EU
Slovenia	SVN	EU
Spain	ESP	EU
Sweden	SWE	EU
Switzerland	CHE	Non-EU

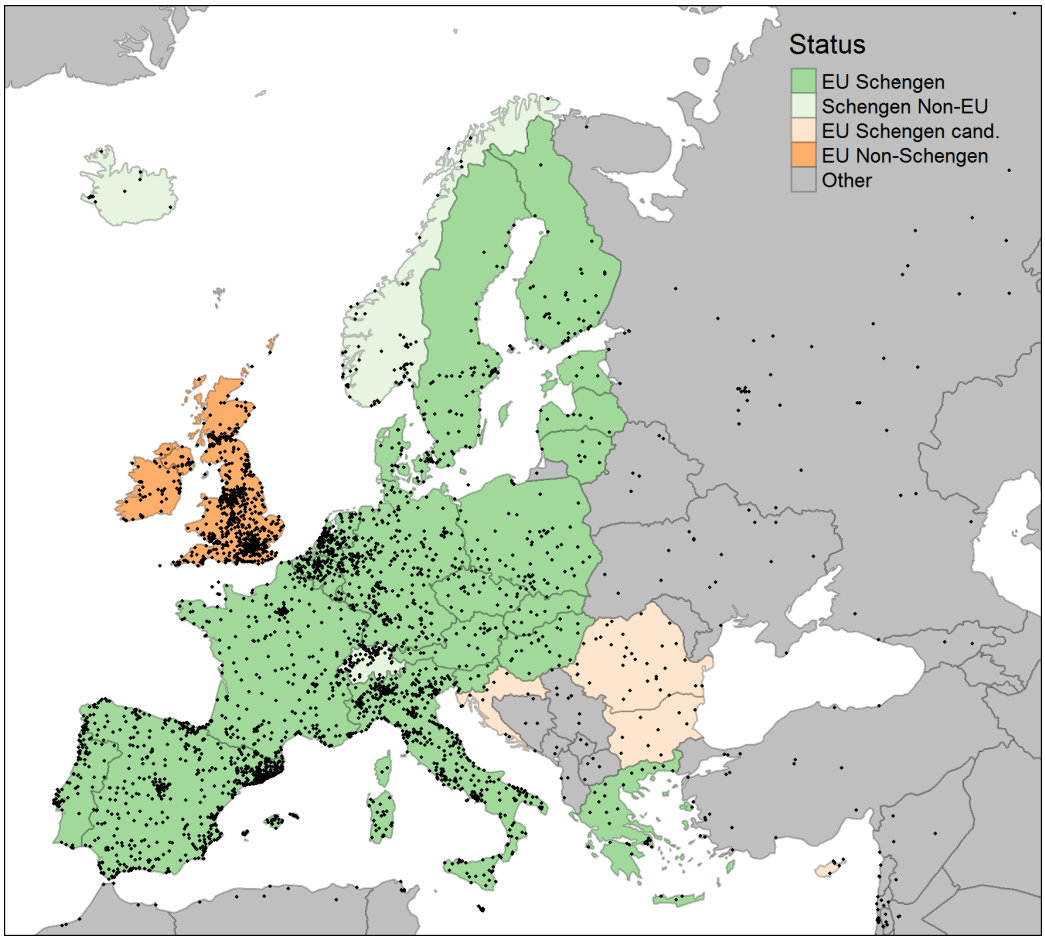


Figure 4.3: Nations in the Schengen Area with distinct user locations overlaid as black dots. Map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035), with the same projection and bounding box as the maps published by Eurostat. Source: own work based on cartography from the "tmap" R package, using Natural Earth data, in the public domain.

Table 4.8: Top ambiguous locality names (appearing more than twice) in the geocoder results.

Locality	Namesakes
Hamilton	4
Cambridge	3
Córdoba	3
Newport	3
Venice	3

Map Unit This code broke the countries into constituent units. The code was unique for each polygon, unlike the country code. It allowed overseas territories in countries with territories in multiple continents (e.g. France and French Guiana) to include the correct region attribute.

Country Each location was assigned the corresponding ISO 3166-1 alpha-3 code[40] —used by the ISO 3166 Maintenance Agency— using the Natural Earth cartography.

World Region This code grouped map units into broad geographical regions: a) continents, b) continents according to the United Nations, c) sub-regions according to the United Nations M49 standard[41], and d) regions according to the World Bank country groups, although only the latter was used in the analysis (Fig. 4.4).

Scope Corresponding to the classification discussed in section 4.5.1, with categories defined as seven increasingly larger geographic rings centered on Barcelona: 1) Municipal, 2) Metropolitan, 3) Regional, 4) Autonomic, 5) National, 6) European, and 7) International.

4.6.2 Level Assignment

All these categories (except the city, which was obtained directly from the geocoder) were assigned to each location using the coordinate returned by the geocoder, according to which polygon —in the administrative and political boundaries cartography— it lied within(Fig. 4.5).

However, when using worldwide cartography, some precision issues arose that could result in a point outside any administrative boundary, in which case it

[40]The country code list is available from the United Nations Trade Statistics Knowledgebase at *http://unstats.un.org/unsd/tradekb/Knowledgebase/Country-Code* at the time of writing.

[41]More details on the standard country or area codes for statistical use (M49) is available at *http://unstats.un.org/unsd/methodology/m49/* at the time of writing.

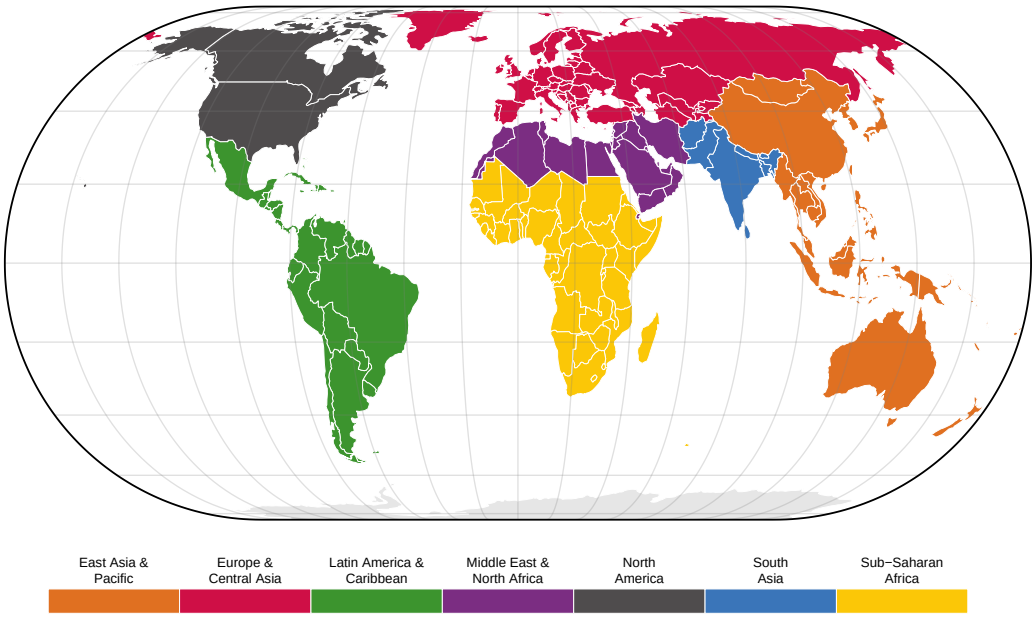


Figure 4.4: World Regions classified according to World Bank analytical grouping, using the same color scheme used in their publications. Map uses the equal-area World Eckert IV projection (ESRI:54012). Source: own work based on cartography from Natural Earth, in the public domain.

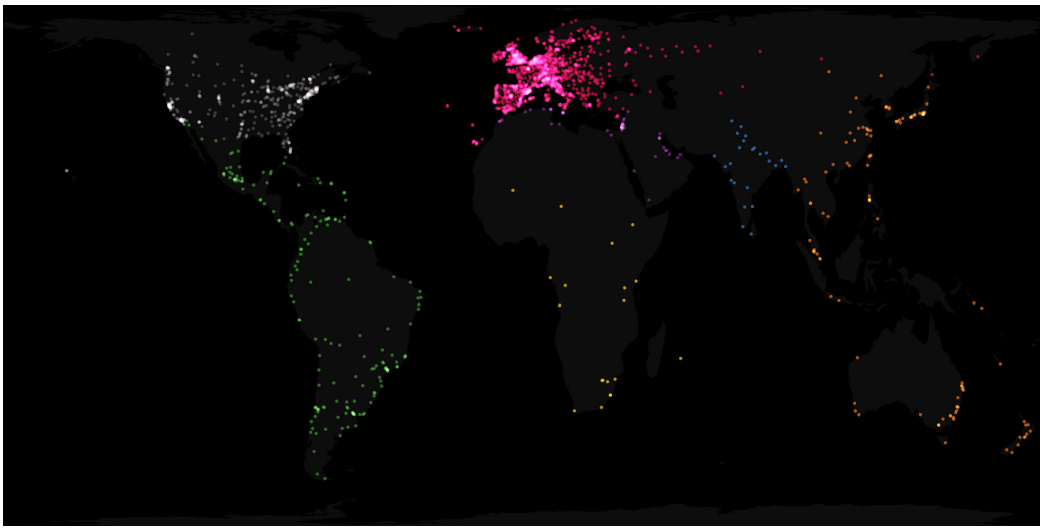


Figure 4.5: Global locations of collected Flickr users, colored according to the World Bank region classification, using the same color scheme as figure 4.4. Dots are overlaid using the QGIS screen blending mode. World map uses the Hobo–Dyer cylindrical equal-area projection. Source: own work based on cartography from Natural Earth, in the public domain.

would not be assigned any category:

- Continental waters such as lakes where not polygon was present.
- Insufficient resolution of the coast lines.
- Small islands that were not represented.

Therefore, when aggregation attributes were assigned by location —all attributes discussed in section 4.6.1 except the city—, they were assigned to the *nearest* polygon feature adopting the following heuristic, using a port[42]of the GEOS[43] (Geometry Engine, Open Source) to the R language through the rgeos 0.3-23 package [277]:

1. The points were checked whether they fell within (intersected) the boundary of any polygon and if they did, they were assigned the attributes of this polygon.
2. A distance matrix was built between the unmatched points and the polygons.
3. The attributes of its closest polygon (smallest distance value in the distance matrix) were assigned to each unmatched point.

[42]Which itself its a C++ port of the Java library JTS Topology Suite, available at *http://github.com/locationtech/jts* at the time of writing.

[43]The GEOS (Geometry Engine Open Source) project website is available at *http://trac.osgeo.org/geos* the time of writing.

Table 4.9: Summary of number of users and corresponding picture counts, according to their origin, classified according to the criteria discussed in section 4.5.3 in the geographic units discussed in section 4.5.3.

Origin	Users	Pictures
Municipal	2,247	204,641
Metropolitan	311	48,187
Regional	361	29,325
Autonomic	413	23,017
National	1,897	50,545
European	5,139	111,300
International	6,778	223,021
Undetermined	17,137	476,668

After the geocoding process, the users with origin information were about 50%[44] of the total number of users, which accounted to almost 60%[45] of all the retrieved pictures (Table 4.9).

[44]Locations of 17,146 out of 34,283 retrieved users.
[45]Locations of 690,036 out of 1,166,704 retrieved users' pictures.

Chapter 5

A Global Perspective

> "Cities are 2% of the earth's crust, but they are 50% of the world's population"
>
> Carlo Ratti

5.1 Introduction

5.1.1 Background

The classic approach in spatial analysis answers questions related to *where*, insomuch as Esri, one of the leading drivers in GIS technology uses the slogan "The Science of Where" as their slogan[1]. However, this chapter focuses on the global scale, turning the classic spatial analysis approach inside-out and changing the research question from "where" into "*from* where", using the service users' hometown location data.

As discussed in section 1.1.3, one of the reasons of the reasons Barcelona was selected as the case of study was its popularity among visitors among global cities (Table 3.7). Being a powerful attractor can be beneficial for a city, as tourism-related services can be generators of economic activity[2], but can also

[1] As appears in the company promotional video available at *http://youtu.be/XrU8GX7manc* at the time of writing.

[2] As seen in the quick expansion or Airbnb in Barcelona, as published in "La rápida expansión de Airbnb en Barcelona, en un mapa temporal" available in La Vanguardia at *http://www.lavanguardia.com/local/barcelona/20170726/4378476565/expansion-airbnb-barcelona-mapa.html* at the time of writing.

cause issues in the daily lives of its inhabitants [278] and in the long run even lead to tourismphobia [279].

Social networks offer a unique perspective [280] to understand the behavior of visitors [281], as one of the most popular activities of tourists is taking pictures, which are increasingly posted on social networks [11]. This approach can provide a new research avenue to understand travel on a continental [282] or even global scale [283], without relying on airport statistics alone.

This chapter focuses on using the location data from Flickr users who took pictures of Barcelona —aggregated according to their region of origin as discussed in chapter 4— to study the regional distribution of the visitors of the city. However, since the research focuses on a single service, the regional variation in the popularity of the service can potentially bias the results, in addition to the other inescapable biases of social network data (e.g. gender, age, occupation). Despite these limitations, the developed visualization methodology should prove valuable in other scenarios, particularly in exploratory research.

5.1.2 Chapter Outline

This chapter discusses the data pipeline developed to convert the raw user data (its retrieval and pre-processing is discussed in chapter 4) into information, and it is divided in the following sections:

- Section 5.2 discusses the more suitable visualization strategies for large amounts of data at a global scales, focusing on geographic maps and treemaps, but also discussing scatter plots and dot plots.
- Section 5.3 analyzes the global distribution of Flickr users who took a geotagged picture of Barcelona and their corresponding pictures, focusing on their cities of origin, using treemaps aggregated into 1) global regions according to the World Bank classification, and b) geographic rings around Barcelona. The connection between the number of pictures and users is also explored to identify possible outliers capable of distorting the results.
- Section 5.4 studies the worldwide distribution of users and pictures, aggregated by country (reducing the group variance discussed in 5.3), as well as the variation in the pictures per user ratio among countries.
- Sections 5.5 and 5.6 try to explain the effect of the confounding variables of the population and Gross Domestic Product (GDP) of the countries of visitors, respectively.
- Section 5.7 proposes an very approximate method to determine the expected contribution of visitors to the economic growth of city of Barcelona according to their country of origin, as an example of a practical application.

5.2 Visualizing World-Scale Data

5.2.1 Representation and Interpretation Challenges

As a first step to extract information from the location of the users' origins, it was necessary to explore the appropriate strategies to visualize the relationship of this location with other explanatory variables of interest effectively. The first challenge was the representation of worldwide data —not necessarily cartographic— within the size limitations of a sheet of paper or a computer screen, optimizing the available space.

To solve this issue, the main strategy was the aggregation of data into larger geographic units (discussed in sections 4.5 and 4.6), classified according to which unit each location was within or —in the cases where no intersection was found— its nearest one. After the classification, summary statistics (e.g. counts, means or medians) could be computed for locations aggregated into these units:

- Countries (subdivided into map units).
- World regions.
- Geographic rings centered on Barcelona.

However, these aggregation units were very different (Table 5.1) in their sizes, number of user locations that fell within them, and their associated data (e.g. population), in many cases with very long tails in their histograms, and it was therefore necessary to adjust the variables:

- Taking the base-10 logarithm of the variable.
- Using the median instead of the mean (which is more robust to the presence of outliers).
- Standardizing the units (using their z-score).
- Suppressing the units with a small number of samples, considering them as noise and making them equivalent to unavailable data.
- Classifying the values using the Fisher-Jenks natural breaks algorithm, maximizing differences *between* classes while minimizing differences *within* classes.

5.2.2 Cartography

One of the main issues when visualizing the origins of users as cartographic data was the uneven distribution of locations, including some regions with a high data density but also large areas without any data[3]. This issue was was partially solved through the aggregation into larger entities, and the computation of the

[3]The water surface is 70.8% of the total area or the planet.

Table 5.1: Descriptive statistics of the distributions of the analyzed dependent (picture and user counts) and independent (area, population and GDP) variables per country.

Attribute	Range	$\log_{10}$(max/min)	Skewness	Kurtosis
Picture count	1 – 355,715	5.55	10.95 >> 0	127.56 >> 3
User count	1 – 5,229	3.72	8.33 >> 0	82.47 >> 3
Area	0.01 – 16,953,094	9.23	6.27 >> 0	47.34 >> 3
Population	4 – 1,338,612,968	8.52	10.02 >> 0	110.95 >> 3
GDP	0.3 – 15,094,000	7.70	9.83 >> 0	117.79 >> 3

corresponding summary statistics. Following this aggregation, two visualization families were possible:

Choropleth maps as polygons representing political and administrative boundaries. According to the "grammar of graphics" [170] definition, the available graphical attributes are limited because shape, position, orientation and size are already used by the geometry of the polygon entities [284], leaving only fill (color shading or patterning) as the vehicle to convey the information of a single variable or, exceptionally, a limited combination of two variables [285]. Arguably border attributes (width, line type and color) could also be used, but become unsuitable because of overlap issues since in the case of countries they are topologically coincident by definition. When color values represent densities, it is advisable that the projection of the map preserves the area relationship of entities.

Bubble maps as points representing cities of origin. In this case the geometry is collapsed into a 0-dimensional entity (point) in 2D[4] space, and therefore size, color and symbol type are available to convey information driven by the attributes. To overcome the overplotting issues, the points were plotted using the screen[5] blend mode, where pixels become brighter as more points are overlaid on top of one another, producing a "bloom" effect.

5.2.3 Cartographic Pitfalls

The aggregation into polygons can introduce biases in what is known as the Modifiable Area Unit Problem (MAUP) [286], where arbitrary divisions of space can distort the results of spatial aggregation[6]. In the case of the countries this

[4]Geographic space is topologically 2D although it is on a 3D sphere.
[5]This blend mode is available in most image editing applications.
[6]And also invalidate statistical hypothesis testing.

problem is moot because their borders can be considered fixed, but could become an issue in the case of the definition of regions or geographic scopes discussed in section 4.6.

In addition, since choropleth maps rely solely on color to convey information about a single variable, the choice of colors [287] and the definition of breaks [288] can also introduce confusion in the interpretation of maps [180]. To avoid these issues, perceptually balanced color schemes were used [289] and a histogram with colored bands [290] was provided next to the legend to show how the breaks were distributed within the range of the variable.

In the case of categorical variables, the colors were consistent between the different types of visualizations (world region colors in scatter plots and treemaps) or shared across variable types (colors for user and picture data across choropleth maps, bubble maps, and dot plots).

Finally, when mapping large areas of the world on a flat shape, the fundamental problem of cartography regarding the impossibility of any map projection of preserving both angles[7] and areas[8] becomes apparent. Therefore, the choice of projection plays an important role [291, 292, 293]. Three *authalic* (equal-area) projections were used to visualize the cartographic data:

Eckert IV Described in 1906 by the German geographer Max Eckert as one of a series of three pairs of projections, it is an equal-area pseudocylindrical map projection. This projection is used in figures 5.9, 5.10, 5.15, 5.16, 5.18, 5.19, 5.22, 5.23 and 5.26, as well as figure 4.4 in chapter 4.

Hobo–Dyer Created in 2002 by cartographer Mick Dyer and commissioned by Bob Abramms and Howard Bronstein, this cylindrical equal-area projection is a modification of the 1910 Behrmann projection, with the north and south standard parallels at 37°30’ and an aspect ratio close to 2. This projection is used in figures 5.2 and 5.3 (top), as well as figure 4.5 in chapter 4.

Lambert Azimuthal Announced in 1772 by the Swiss mathematician Johann Heinrich Lambert, this azimuthal equal-area projection maps a sphere to a disk. The ETRS89-LAEA (EPSG:3035) is recommended for statistical mapping in Europe and is used in figures 5.2 and 5.3 (bottom), as well as figure 4.3 in chapter 4.

5.2.4 Treemaps

A treemap is a space-filling visualization of hierarchical structures, using a series of nested rectangles, whose areas are proportional to a numerical associated

[7] In conformal maps, any angle is preserved in the projected image.
[8] Equal-area maps are also called equivalent or authalic.

variable. Each rectangle in a level (branch) of the hierarchical structure (tree) can be recursively tiled into smaller rectangles that belong to a sub-level (sub-branches), until the lowest level is reached (leaves).

Treemaps (Fig. 5.1) can be considered a special case of product plots[9], a framework for visualizing tables of counts, proportions and probabilities [294], along with mosaic plots [295].

These plots can be particularly useful to visualize the proportion of a part in relation of the whole because by construction make very efficient use of space, and therefore have traditionally been used to display the allocated space of files inside the different nested folders of a computer file system.

One of the most common treemap applications is the visualization of the economic output of a region by sector, which in turn can be broken down by subsector. This visualization is used (Fig. 5.1a) by The Observatory of Economic Complexity[10] by Alexander Simoes at the Massachusetts Institute of Technology and the The Atlas of Economic Complexity[11] at Harvard University.

Another treemap example is the Newsmap[12] visualization mashup (Fig. 5.1b), a news aggregation tool that uses a treemap to display news headlines inside tiles, with sizes proportional to their popularity and coded according to their category[13].

In this chapter treemaps are used for figures 5.4, 5.5, 5.6, 5.7, 5.13 and 5.14, to visualize patterns that would otherwise remain hidden. The diagrams were produced with the R package treemap[14] version 2.4-2 by Martijn Tennekes [296], although the ggplot2 extension treemapify[15] version 2.3.2 [297] was also considered. The following conventions were used:

- Two levels were used, a higher hierarchical level or broad category (world regions or geographic scopes) and a lower hierarchical level or detailed category (countries or cities).
- The tiles of the same broad category were packed in the shape of a rectangle (outlined with a thicker line) and shared the same color.

[9] And therefore a distant relative to the pie chart.

[10] The Observatory of Economic Complexity is available at *http://atlas.media.mit.edu/* at the time of writing.

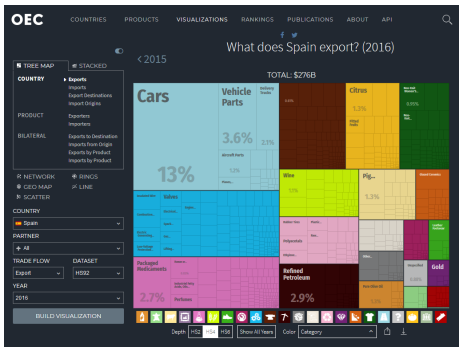
[11] The Atlas of Economic Complexity is available *http://atlas.cid.harvard.edu/* at the time of writing.

[12] Newsmap is available at *http://newsmap.jp/* at the time of writing.

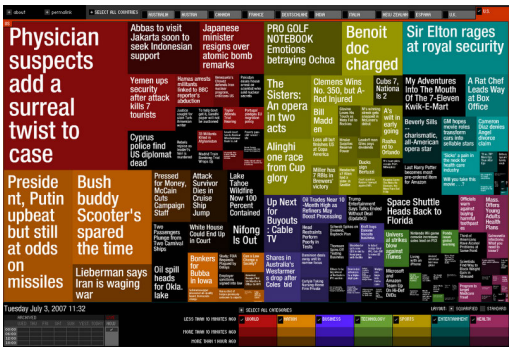
[13] The color-coded categories in Newsmap are World, National, Business, Technology, Sports, Entertainment, and Health.

[14] The treemap package is available on GitHub at *http://github.com/mtennekes/treemap* at the time of writing.

[15] The treemapify package is available on GitHub at *http://github.com/wilkox/treemapify* at the time of writing.



(a) Treemap of the 2016 Spanish exports. Source: The Observatory of Economic Complexity (MIT) under CC BY-SA 3.0.



(b) Screenshot of the Newsmap website. Source: Blog entry posted on November 15, 2007 at newsmap.blogspot.com.

Figure 5.1: Examples of treemaps showing the relative sizes of elements in a collection intuitively and with an efficient use of space.

- The tiles of the detailed categories were labeled while the colors of the broad categories were explained in a legend.
- The size indicated the proportion of the total, and the sum of areas always summed 100%
- The "squarified treemaps" algorithm [298], designed to produce tiles with an aspect ratio closer to one, was used.
- Items in the detailed category that were lower than a specified threshold —defined as the percentages that would result in tiles with a printed area of less than about 1cm²— were combined into their own "small elements" tile in each broad category (if applicable).

5.2.5 Avoiding Treemap Clutter

There were unique 2,973 cities, which had to be displayed in a treemap tiling a 13.5 x 18 cm area; if all pieces were equal, this would yield a size of just about 0.08 cm² per tile. However, some proportions were significantly larger than others, and therefore some tiles would be even smaller.

To avoid cluttering the display with very small tiles, proportions that would result in tiles with an area of less than a 1cm² were combined into a single "small elements" tile per category.

This operation did not alter neither the dimensions of the elements in the higher hierarchical level (scopes) nor the sizes of the other tiles in the secondary level (cities). However, in some cases all elements of a category of the higher hierarchical level would belong to the "small tiles" definition; in these cases, the category

was not further partitioned into smaller tiles.

5.2.6 Other Visualization Strategies

Despite its advantages, on some occasions treemaps were not the most adequate visualization strategy, and other —more standard— graphs were used to complement or substitute them.

Scatter plots were used in figures 5.8, 5.17, 5.20, 5.21, 5.24 and 5.25, with several enhancements because of the large number or elements to display:

- On some plots the top or bottom data points were labeled and or colored for identification of the outstanding values.
- Some plots used faceting [299] to visualize scatter plots side-by-side according to their category, while in others dot color was used for the same purpose.
- The axis included log ticks when a logarithmic scale was used.

Finally, Cleveland dot plots —not to be confused with dot plots [300]— were used in figures 5.12, 5.11 and 5.27 as an alternative to bar charts —reducing non-data ink usage [160]— to display the top elements of a list, as an improved solution instead of tables in the case of skewed distributions.

5.3 Cities of Origin

5.3.1 Source Data

The second level of aggregation (discussed in section 4.6) was the hometowns of Flickr users, resulting in a total of 2,973 unique cities where the users —who took a picture of Barcelona and posted it publicly— stated they were residing. After the process discussed in chapter 4, each of these unique cities included the following data:

- Data about the locality:
  - Name of the locality.
  - Geographic coordinates of the city.
- Data about the geographic context of the locality:
  - World region according to the World Bank classification.
  - Country and Map unit.
  - Geographic scope (e.g. metropolitan, regional).
- Summary statistics for each locality:
  - Number of pictures.
  - Number of users.

- Mean pictures per user.
- Median pictures per user.
- Standard deviation of the pictures per user.

5.3.2 Geographic Distribution per Cities

The locations of the user's hometowns were mapped using two different projections (discussed in section 5.2.2). Both projections were authalic, the top one encompassing the global distribution of locations, while the bottom one focused on Europe, where the locations mere more densely concentrated.

Two sets of maps were produced, one for users (Fig. 5.2) and another one for pictures (Fig. 5.3), using two different color schemes consistent with the other maps discussed in this chapter for user (blueish) or picture (reddish) data. Otherwise, both sets of maps shared the same projection and bounding boxes.

The size of the points in the map was proportional to their z-score —distance, measured in standard deviations, of the count of users or pictures in each city from the set mean—. The range of the z-score was linearly mapped (stretched) to a scale-dependent range where the largest size was four times larger than the smallest.

The points were overlaid using the screen blend mode, where the color values of two inputs are inverted[16], the inverted values are multiplied together, and the result of the multiplication is inverted again, using the formula below where a and b are the color values (the operation is symmetric and the input order does not influence the output):

$$f(a, b) = 1 - (1 - a) \cdot (1 - b)$$

These locations were overlaid on a generalized[17] world map from the Natural Earth cartography at the 1:110 million scale, using a dark gray over a black background to emphasize visibility of the locations while providing some geographic context through the outlines of the continents.

The results showed the major populated areas or the world, biased towards the countries with a higher GDP and the European geographic scope, where travel to Barcelona is cheaper because of the closer distance and more convenient thanks to the Schengen agreement.

In both maps, some geographic features such as the Apennine Mountains or the Alps were visible through the *lack* of Flickr users. However, the cities where its

[16] In image processing software, inversion consists in inverting the range, and is sometimes called "one minus".

[17] Generalization is the simplification of polygons or line features in a map.

users produced proportionally high or low amounts of pictures were difficult to observe using side by side comparisons, and a scatter plot (Fig. 5.8) was more suitable[18] to explore this relationship (discussed later in section 5.3.5).

The overview maps provided a broad perspective of the distribution of the origins of users, but the quantity and the uneven distribution of points introduced visual noise and clutter that made its interpretation difficult. To alleviate these issues, the locations were aggregated under two different criteria, discussed in the next two subsections:

- World regions (section 5.3.3)
- Geographic scopes (section 5.3.4)

5.3.3 Regional Distribution of Cities

To analyze the regional distribution of cities where Flickr users resided, two treemaps were produced —for the proportions of users (Fig. 5.4) and pictures (Fig. 5.5) from the corresponding total— classifying the cities into one of the seven world regions defined by the World Bank (discussed in section 4.6), according to which region they were geometrically within. These categories were displayed in the treemaps[19]. The described approach but analyzing countries instead of cities is discussed in section 5.4.4.

As was to be expected, the users from Barcelona itself took proportionally more pictures of the city than it was expected from their proportion of users, since they were likely to spend more time in Barcelona than users from other cities, and therefore have more opportunities to take a picture. It was expected that pictures from residents would me more candid; the content of the picture, however, was not analyzed.

All the users from Middle East & North Africa, South Asia and Sub-Saharan Africa came from a city with a proportion of users low enough to be grouped in the "small tiles" category. In addition, their tiles in the picture counts were even smaller, suggesting that they took proportionally less pictures on average.

In the Europe & Central Asia, non-domestic users from London, Madrid, Paris and Rome were the most abundant, while in the picture count cities nearby Barcelona (Sabadell, Santa Coloma de Gramanet, Cornellà de Llobregat) were also prominently featured. This can be explained because these cities might have a minority of users that takes many pictures, which is more likely if these users can afford to travel to Barcelona easily and frequently.

[18] As discussed in section 5.2.6.

[19] With the same color scheme as figure 4.4 on page 113.

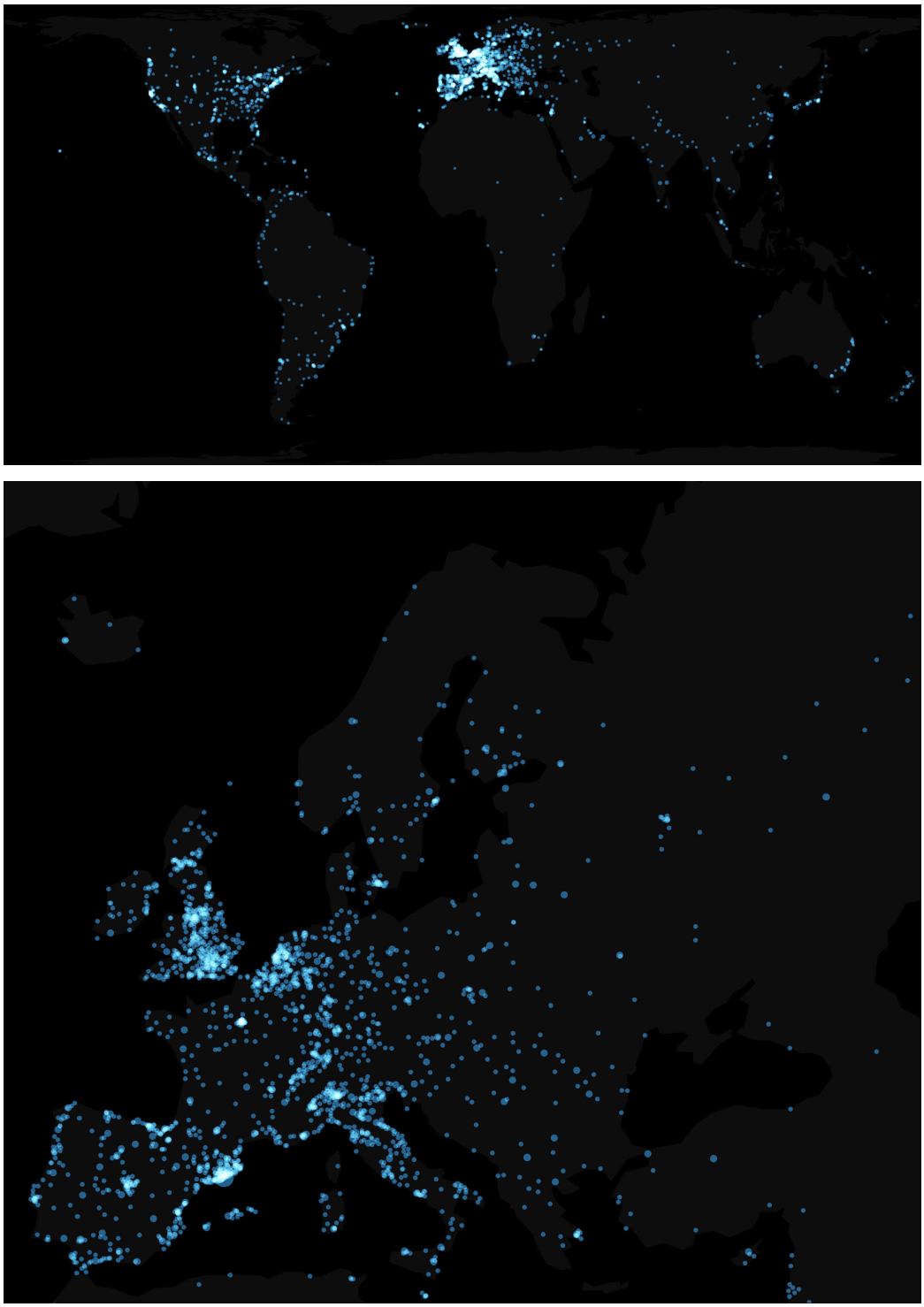


Figure 5.2: Map of all the cities where Flickr users who took at least a picture of Barcelona reside. Dot area of each city is proportional to the z-score of its corresponding number of users. Dots are overlaid using the QGIS screen blend mode. World map uses the Hobo–Dyer cylindrical equal-area projection, Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on cartography from Natural Earth, in the public domain.

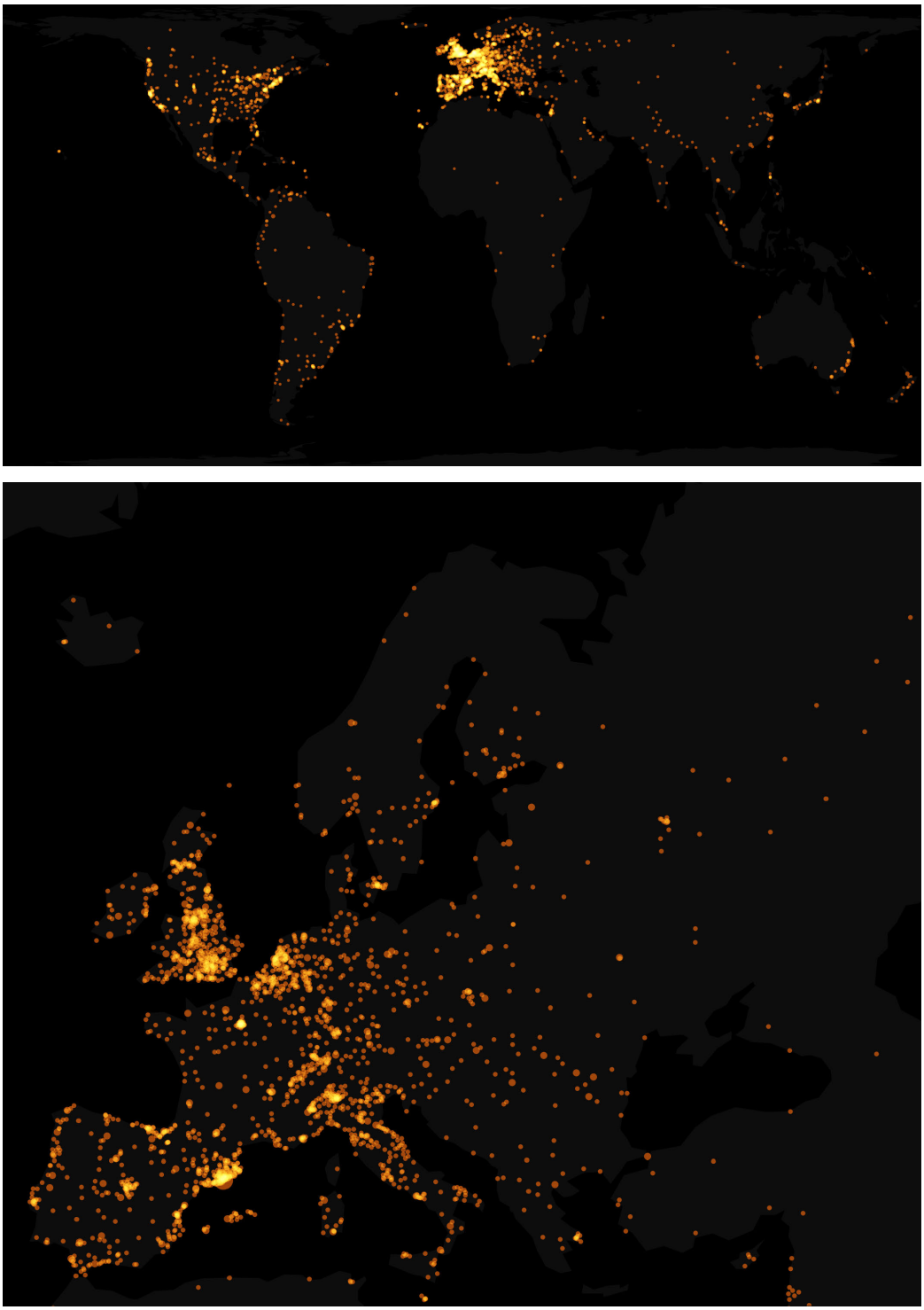


Figure 5.3: Map of all the cities where Flickr users who took at least a picture of Barcelona reside. Dot area of each city is proportional to the z-score of the number of pictures taken. Dots are overlaid using the QGIS screen blend mode. World map uses the Hobo–Dyer cylindrical equal-area projection, Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on cartography from Natural Earth, in the public domain.

The same phenomena can also be observed in the Latin America & Caribbean, where a lot of pictures seem to come from users in Porto Alegre, but this city does not even appear in the user counts, unlike Santiago and São Paulo which keep roughly the same proportions in both treemaps.

These outliers suggest that user counts are better suited to be studied using treemaps because they are less likely to be distorted by "super-users" (users with unusually large collections of pictures). These distortions can be better identified using an scatter plot, as discussed in section 5.3.5.

5.3.4 Cities Distribution within Influence Rings

The second approach classified the same cities of origin into geographic scopes (discussed in section 4.5.1), corresponding to seven regions (rings) of diminishing social and economic integration centered around Barcelona. Two treemaps were produced, for the user (Fig. 5.6) and picture (Fig. 5.7) counts of the cities within these regions.

In these treemaps, since the chosen visibility threshold was the same (discussed in section 5.2.5), the cities that appeared were —by construction— the same set as in their world regions treemap counterparts of users (figures 5.4 and 5.6) and pictures (figures 5.5 and 5.7) counts, with the same sizes —although their aspect ratios could be different— but grouped with a different criterion.

Therefore, the conclusions regarding the cities discussed in the previous subsection are the same, but in this case the effect of the distance on the presence of outliers —cities with many users who took a lot of pictures of Barcelona— from the Metropolitan and Regional rings is much more noticeable.

In this case, the treemaps also allow visualizing the relative proportion of visitors from Europe and the rest of the world, each accounting for about one third of all Flickr users who took a picture of Barcelona (Fig. 5.6). This proportion is maintained in the picture counts (Fig. 5.7) from international visitors but not in the case of users from the Schengen Area, who on average took about half the number of pictures than their international counterparts.

5.3.5 Ratio Distribution among Cities

The disparity between figures 5.6 and 5.7 prompted to investigate the connection between number of pictures taken and number of users of each city. This relationship was explored to identify potential outliers, using a scatter plot (Fig. 5.8).

In the scatter plot, a point was placed for each city in the position of its user count in the horizontal axis and their corresponding number of pictures in the vertical

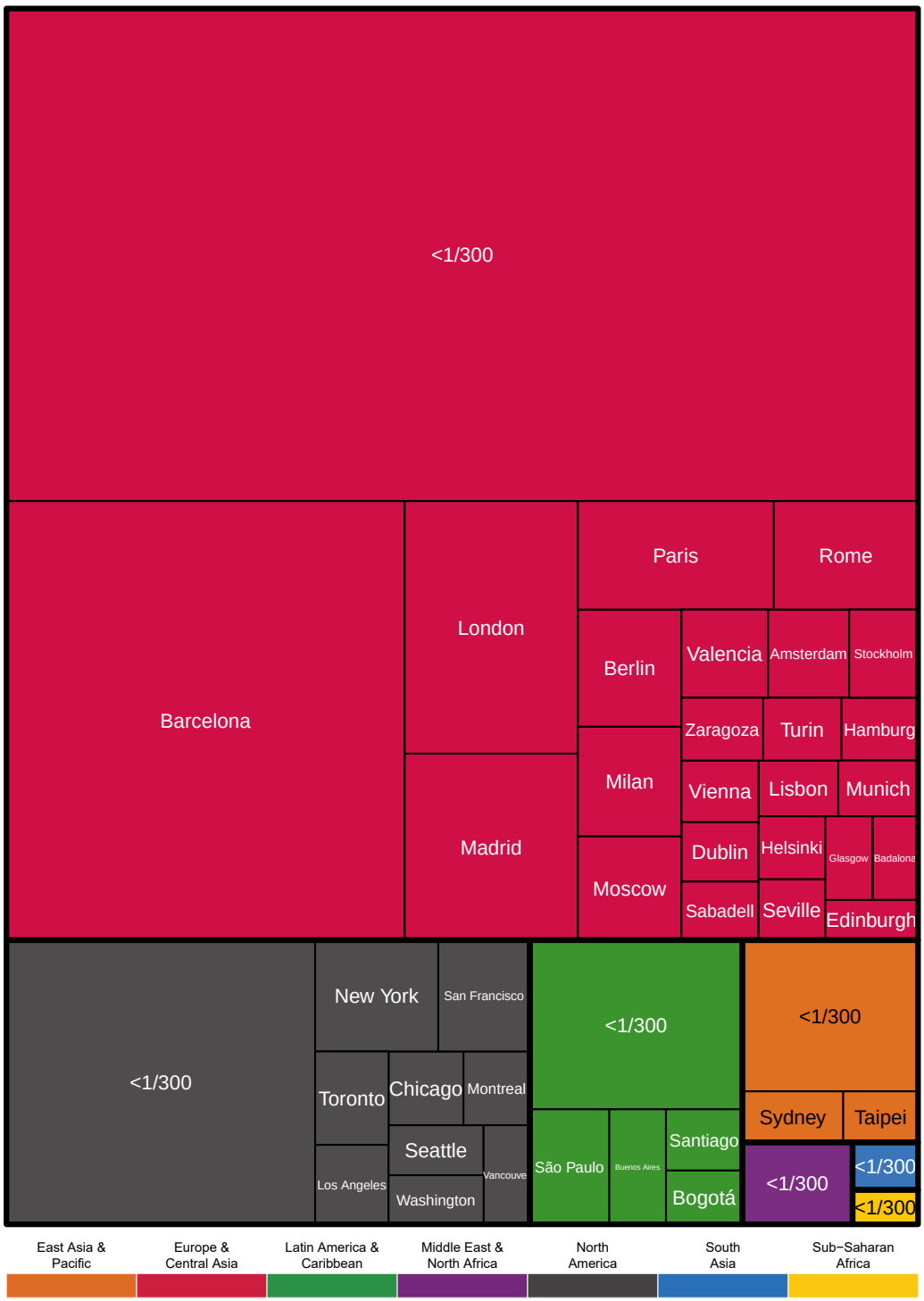


Figure 5.4: Treemap of the number of Flickr users with geotagged pictures of Barcelona, according to the city they reside in. The cities are grouped according to their region following the World Bank classification, and colored with the same scheme as figure 4.4 on page 113. In each region, cities with less than 1/300 of total users are collapsed into a single tile.

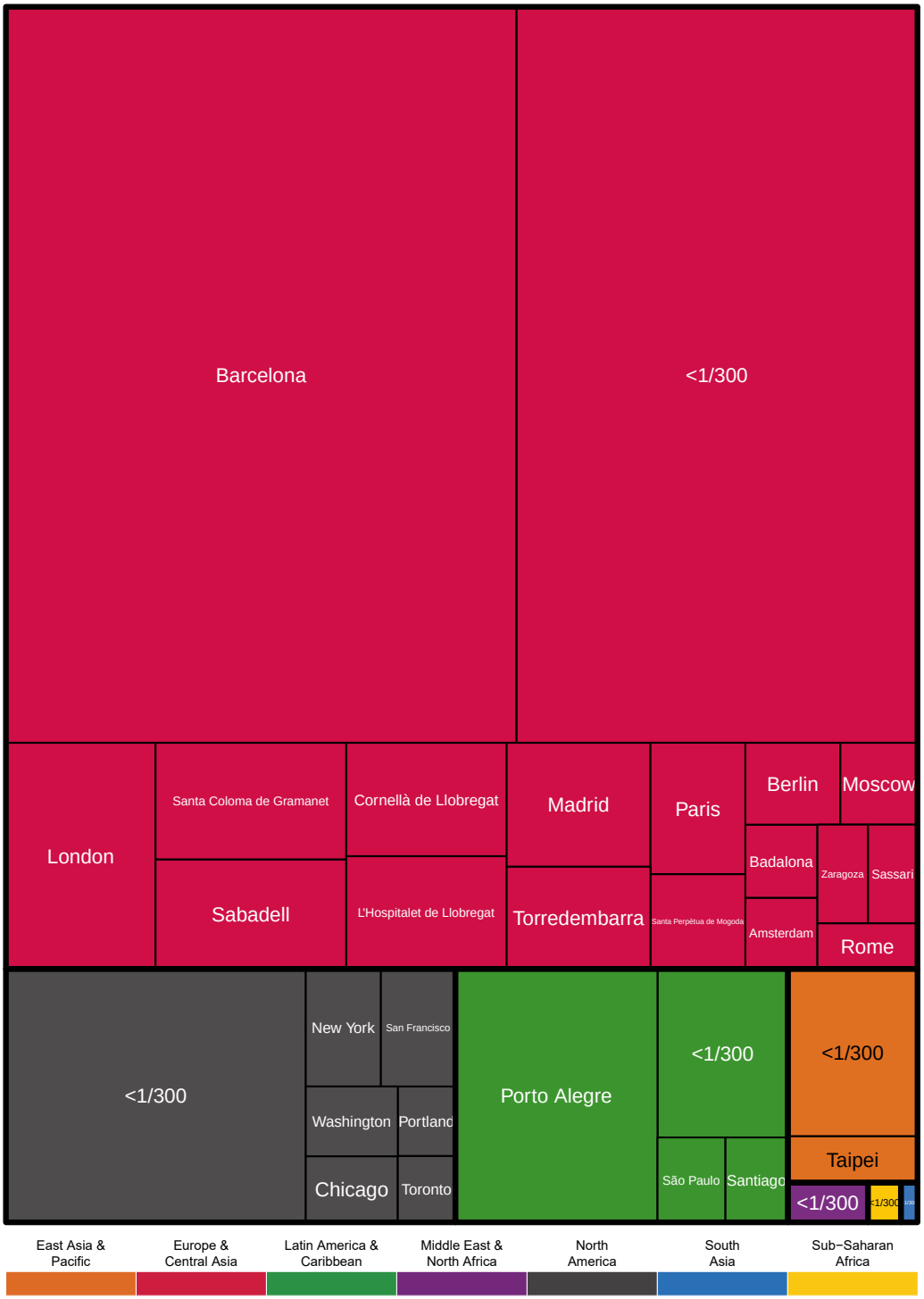


Figure 5.5: Treemap of the number of geotagged pictures of Barcelona posted on Flickr, according to the city of residence of their author. The cities are grouped according to their region following the World Bank classification, and colored with the same scheme as figure 4.4 on page 113. In each region, cities with less than 1/300 of total pictures are collapsed into a single tile.

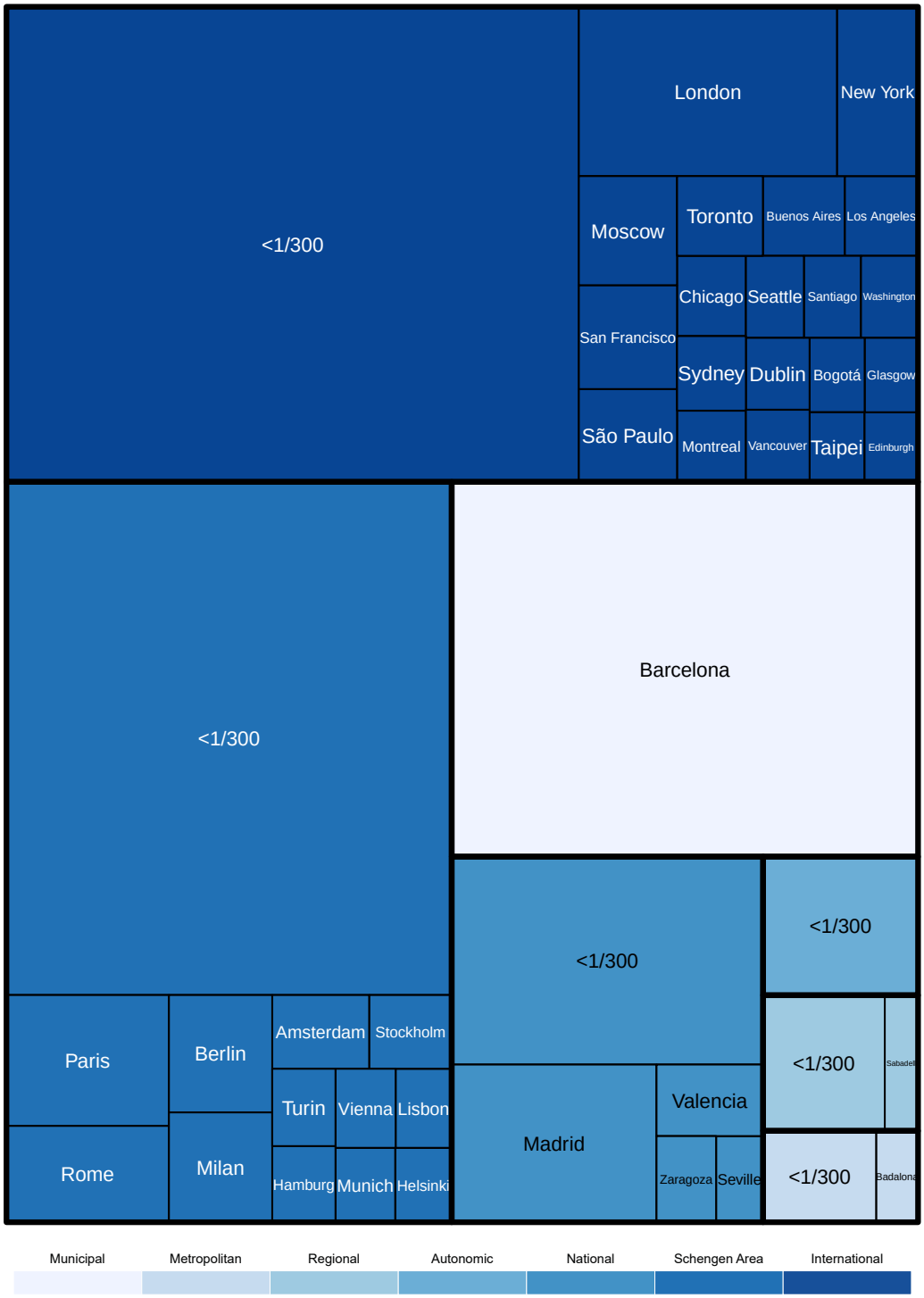


Figure 5.6: Treemap of the number of Flickr users with geotagged pictures of Barcelona, according to the city they reside in. The cities are grouped according to their scope as discussed in section 4.5. In each region, cities with less than 1/300 of total users are collapsed into a single tile.

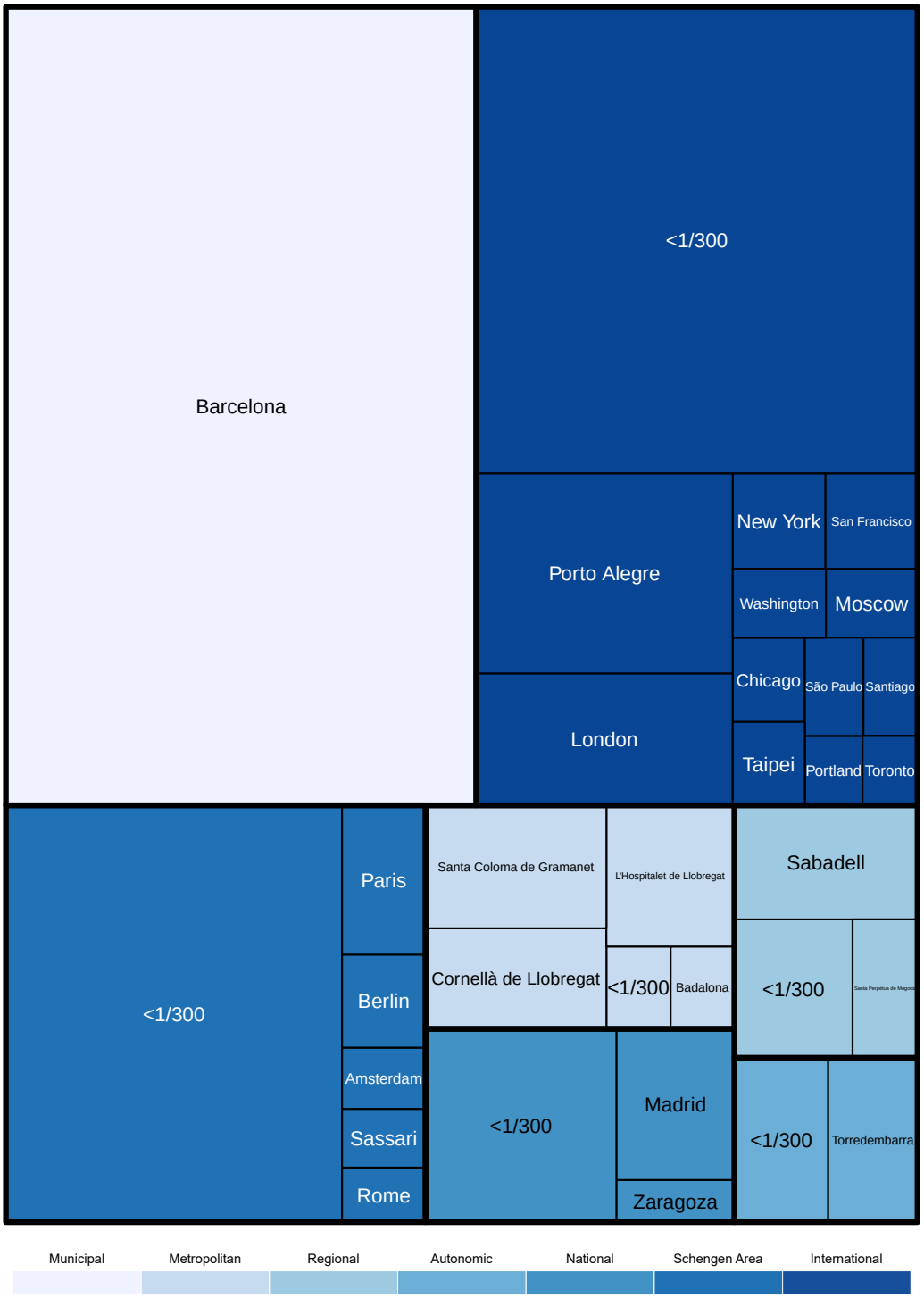


Figure 5.7: Treemap of the number of geotagged pictures of Barcelona posted on Flickr, according to the city of residence of their author. The cities are grouped according to their scope as discussed in section 4.5. In each region, cities with less than 1/300 of total pictures are collapsed into a single tile.

axis. Both axes used a base-10 logarithmic scale, to accommodate the wide range of values (Table 5.1). The median pictures per user of each city was displayed proportionally as the area of the point.

In addition, a log/log linear regression estimation (white) with a 0.95 confidence level (shaded area) was included. Of the cities with 20 users or more, the 20 cities with the highest and lowest ratios —defined as the ratio of logs instead of the log of ratios— were labeled and colored accordingly (blue for highest, red for lowest).

The figure shows that most of the cities with many users —including Barcelona itself— have a low ratio and a low median, while the highest ratios correspond to cities —generally closer to Barcelona— with fewer users and not exceptionally high medians, indicating the presence of a small number of users that took many pictures.

Finally, it can be observed that the cities with fewer than 10 users —and in particular the cities with less than 5 users—, the variance is much higher across the number of pictures and the median pictures per user than in the rest of the cities with larger counts.

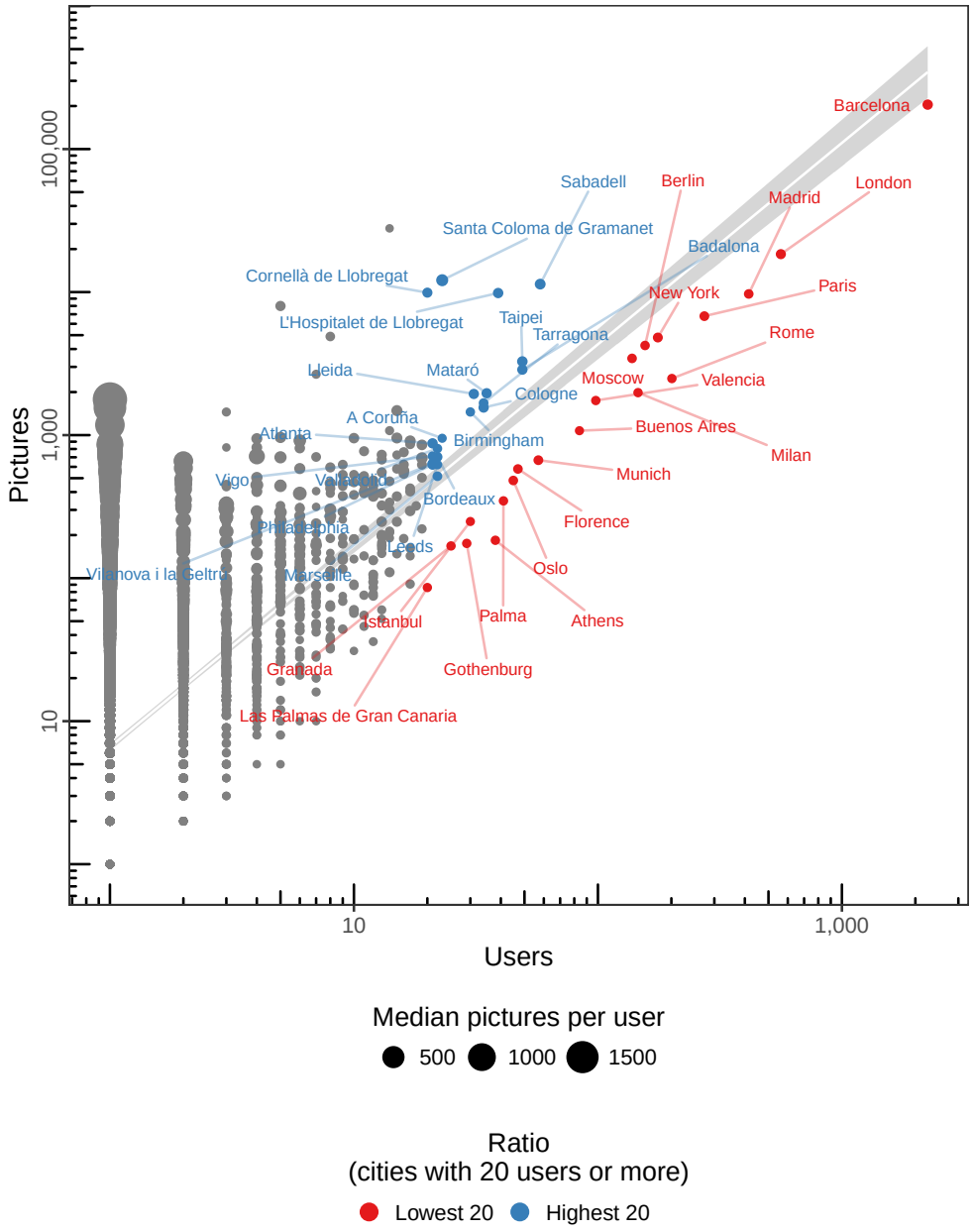


Figure 5.8: Scatter plot of the number of pictures per user, for each city of residence of users who posted a geotagged picture of Barcelona on Flickr. Point area is proportional to the median number of pictures of all users sharing the same location. The 20 locations with the largest or smallest pictures/user ratio are labeled (excluding cities with less than 20 users) and colored according to which tail in the distribution they belong to. Axes use a base-10 logarithmic scale.

5.4 Countries of Origin

5.4.1 Source Data

Beyond cities (discussed in section 5.3), the second level of aggregation (defined in section 4.6) was the country where the locality was located. The base cartography was the Natural Earth 1:110 million scale dataset, which offered cleaner versions of the polygons —with generalized borders—, adequate for the intended printed resolution. However, in spatial join with the country polygons the 1:10 million scale cartography was used instead for improved accuracy.

The cartography included a total of 255 countries, further divided into 295 map units to accommodate special cases such as overseas territories in multiple continents. Each polygon included the following data:

- Data about the geographic context of the polygon boundary:
  - Name of the map unit and its unique code.
  - Name of the country and country code (ISO 3166-1 alpha-3).
  - World region according to the World Bank classification.
- Official statistics about the map unit:
  - Estimated population (discussed in section 5.5).
  - Estimated GDP (discussed in section 5.6).
- Summary statistics of the collected Flickr data for each polygon boundary:
  - Number of pictures.
  - Number of users.
  - Mean pictures per user.
  - Median pictures per user.

5.4.2 Geographic Distribution per Countries

To visualize the global distribution of users, the aggregated counts of users (Fig. 5.9) and pictures (Fig. 5.10) were aggregated by map unit and displayed on a world map, using the equal-area Eckert IV projection (discussed in section 5.2.2).

Because of the wide range of values that had to be displayed (Table 5.1), both choropleth maps used a base-10 logarithmic scale for the color breaks. This transformation allowed the distribution of the mapped values to become quasi-normal, as shown in the histogram placed below the color legend, where each interval was colored to match the corresponding legend entry.

In the maps, European countries are the origin of the majority of the users, probably because of their shorter travel time and fewer difficulties (e.g. same currency, less border controls) when visiting Barcelona, but many users are also

located in the United States, whose users do not enjoy these benefits (a possible explanation being that the user base is biased towards the US).

Although the aggregation into larger units reduced the variance in the data, compared to the aggregation into cities (discussed in section 5.3), there were still some outliers. One of these outliers was Spain, which was to be expected because of the geographical and cultural proximity to Barcelona, but others were the result of the presence of users who took many pictures of the city (this distortion will be further explored in section 5.4.5).

5.4.3 Top Countries per User and Picture Count

While while the maps were very valuable to visualize the variation of spatial distributions, they were limited to show detailed information about the entities they represented (discussed in section 5.2.3), as the countries where the majority of the users resided (Table 5.2) and the amount or pictures their residents took of Barcelona (Table 5.3), were difficult to identify in the map for multiple reasons:

- Some features could be too small to be easily recognizable.
- The color graduation could be difficult to distinguish.
- It was not possible to appreciate the variance within entities clumped inside a color category.

These issues were aggravated because of the skewness of their distributions; therefore, to easily identify the most frequent countries of origin of users (Fig. 5.11) and their pictures (Fig. 5.12) the corresponding dot plots [160] were produced, with the subset of countries whose contribution was at least 0.5% of the total[20].

Both dot plots showed that users from Spain were the majority —as seen in the maps—, but its magnitude compared to the others could be assessed more clearly. The sharp decrease in the contribution of the less represented countries was also easily detected, a fact that was obscured in the maps.

5.4.4 Regional Distribution of Countries

With a similar approach used for the cities (discussed in section 5.3.3), the treemaps for the proportion of users (Fig. 5.13) and pictures (Fig. 5.14) from each country were produced, classified according to the seven world regions defined by the World Bank[21] (discussed in section 4.6). By design, the areas of each region and its distribution were the same as in their counterparts of users (figures 5.4 and 5.13) and pictures (figures 5.5 and 5.14), but partitioned into countries instead of cities.

[20]The cutoff value choice was arbitrary.

[21]Using the same color scheme as figure 4.4 on page 113.

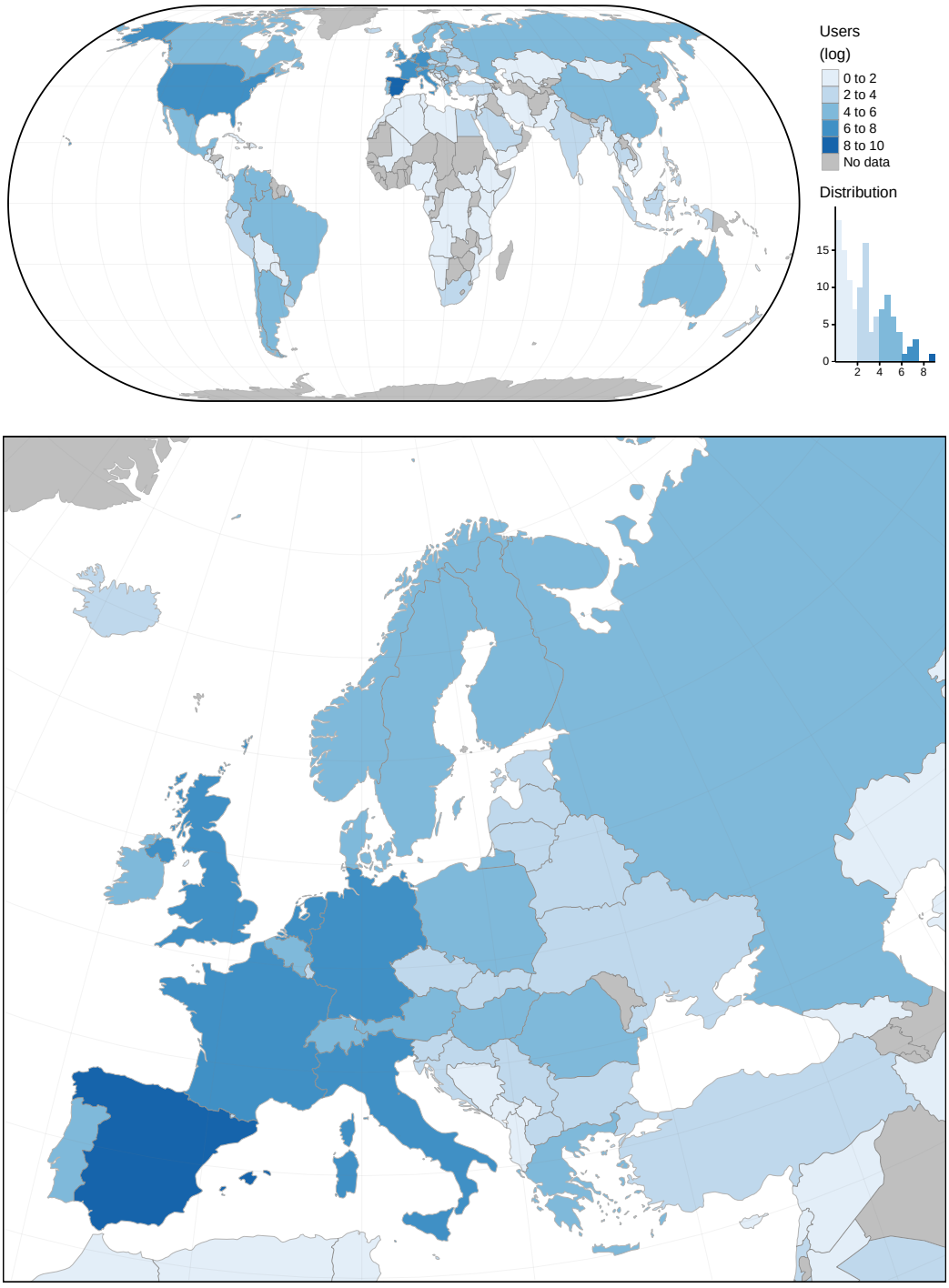


Figure 5.9: Choropleth maps of the number of Flickr users with geotagged pictures of Barcelona per country of origin. Color breaks follow a base-10 logarithmic scale. Some countries with overseas territories are broken into map units. World map uses the equal-area World Eckert IV projection (ESRI:54012), Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on cartography from Natural Earth, in the public domain.

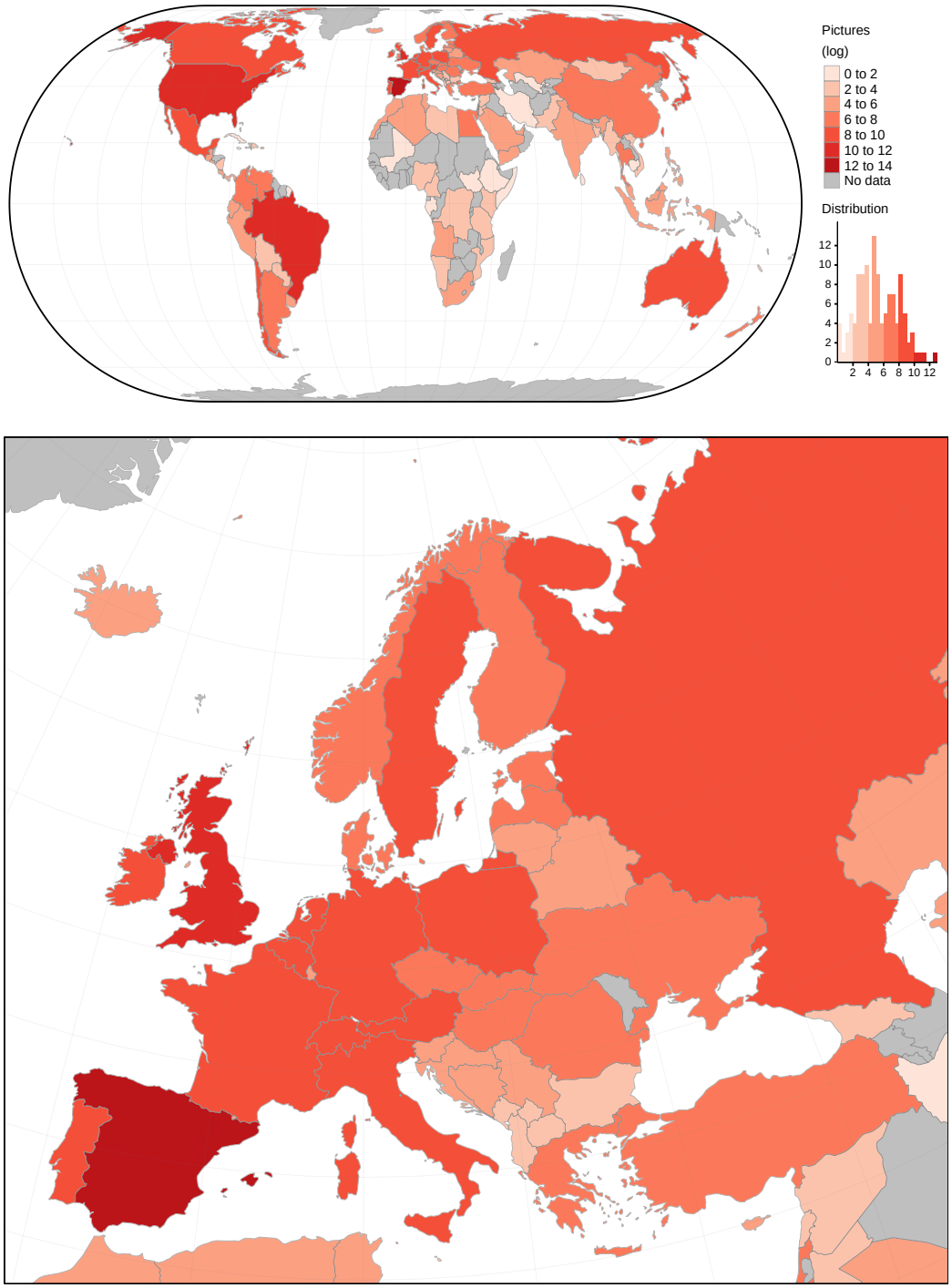


Figure 5.10: Choropleth maps of the number of geotagged pictures of Barcelona posted on Flickr per country of origin. Color breaks follow a base-10 logarithmic scale. Some countries with overseas territories are broken into map units. World map uses the equal-area World Eckert IV projection (ESRI:54012), Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on cartography from Natural Earth, in the public domain.

Table 5.2: Top ten countries according to the number of Flickr users who posted a geotagged picture of Barcelona.

Country	Region (World Bank)	Users	Percent
Spain	Europe & Central Asia	5,229	30.50 %
United Kingdom	Europe & Central Asia	1,944	11.34 %
United States	North America	1,803	10.52 %
Italy	Europe & Central Asia	1,436	8.38 %
France	Europe & Central Asia	874	5.10 %
Germany	Europe & Central Asia	759	4.43 %
Netherlands	Europe & Central Asia	470	2.74 %
Canada	North America	369	2.15 %
Brazil	Latin America & Caribbean	364	2.12 %
Russia	Europe & Central Asia	258	1.50 %

Table 5.3: Top ten countries according to the number of geotagged pictures of Barcelona posted by Flickr users.

Country	Region (World Bank)	Pictures	Percent
Spain	Europe & Central Asia	355,715	51.55 %
United States	North America	61,940	8.98 %
United Kingdom	Europe & Central Asia	58,598	8.49 %
Brazil	Latin America & Caribbean	36,183	5.24 %
France	Europe & Central Asia	21,708	3.15 %
Germany	Europe & Central Asia	21,513	3.12 %
Italy	Europe & Central Asia	21,155	3.07 %
Netherlands	Europe & Central Asia	11,104	1.61 %
Canada	North America	8,655	1.25 %
Switzerland	Europe & Central Asia	7,313	1.06 %

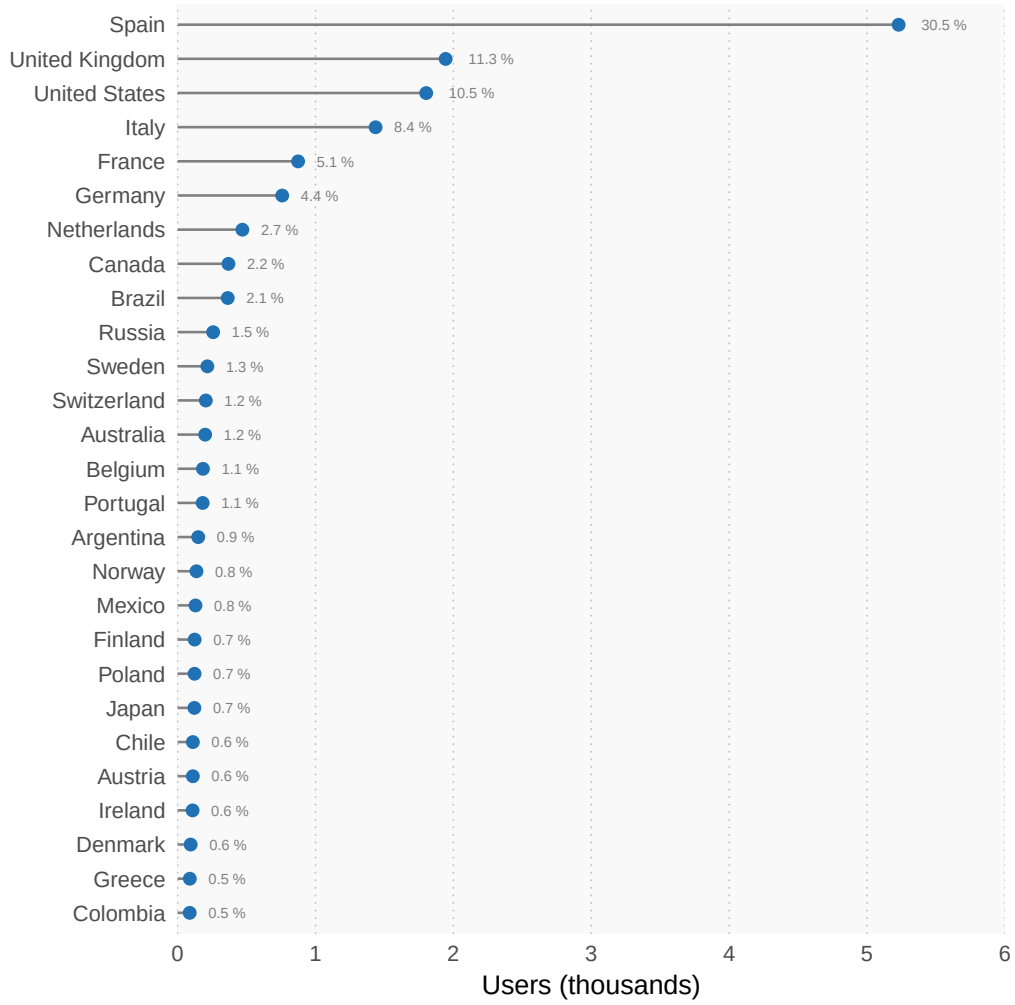


Figure 5.11: Dot plot of the countries with more users who took a geotagged picture of Barcelona. Only countries with more than 0.5% of the total users are depicted.

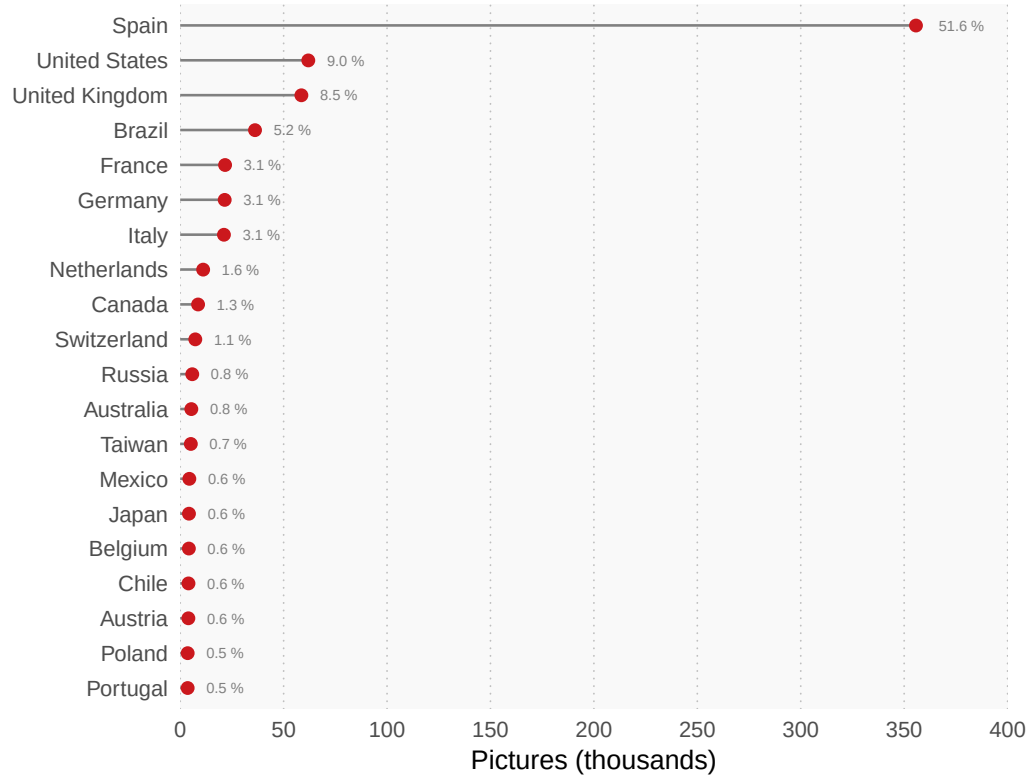


Figure 5.12: Dot plot of the countries whose users took more geotagged pictures of Barcelona. Only countries with more than 0.5% of the total pictures are depicted.

The threshold for aggregating small countries in each region into a single tile was the same (1/300 of total) but in this case the units were larger and the diagram was therefore broken into a larger number of tiles, as more units passed the threshold. For the same reason, the effect of "super-users" was more diluted than in the case of cities, because countries had more users aggregated together to compensate their influence.

The diagrams show that the majority of users are Spaniards, who take proportionally more pictures of Barcelona, arguably because they are able to visit more frequently or stay longer (and many of them are residents). The majority of the visitors and their pictures are from European countries, with a significant contribution of the most populated countries in America and Asia.

5.4.5 Ratio Distribution among Countries

To visualize the differences in the amount of pictures taken by users of each country, maps of the average pictures per user aggregated by country were produced (Fig. 5.15). The distribution —because of the presence of country outliers— made the range of values to be mapped large enough to justify breaking the color scale using a base-10 logarithm. The resulting transformed densities had a roughly normal distribution, as shown in the histogram below the color guide beside the maps.

However, this maps were sensitive to outliers, as can be observed in Angola and Yemen, each with a single user who took 166 and 157 pictures of Barcelona, respectively. It was therefore necessary to find a more robust method less sensitive to outliers.

To reduce the influence of outliers, the selected approach focused on the median, which is more robust to the presence of outliers than the mean. The median calculation was restricted to the countries with more than 15 users, a figure that was considered large enough to provide a sample less likely to be biased by a single user.

Unlike the maps focusing on the mean, the ranges of the resulting maps (Fig. 5.16) did not require transformation, and the corresponding distribution (shown in the histogram below the color guide) was roughly normal. However, the choropleth map format was still unsuccessful in identifying and sorting the countries with the highest median (Table 5.4).

Finally, to identify the countries with the highest ratio they were visualized in a scatter plot (Fig. 5.17), with the users as the independent variable (horizontal axis), the number of pictures as the dependent variable (vertical axis) and the size of the point proportional to median pictures per user. Both axes used a base-10 logarithmic scale, to accommodate the wide range of values (Table 5.1).

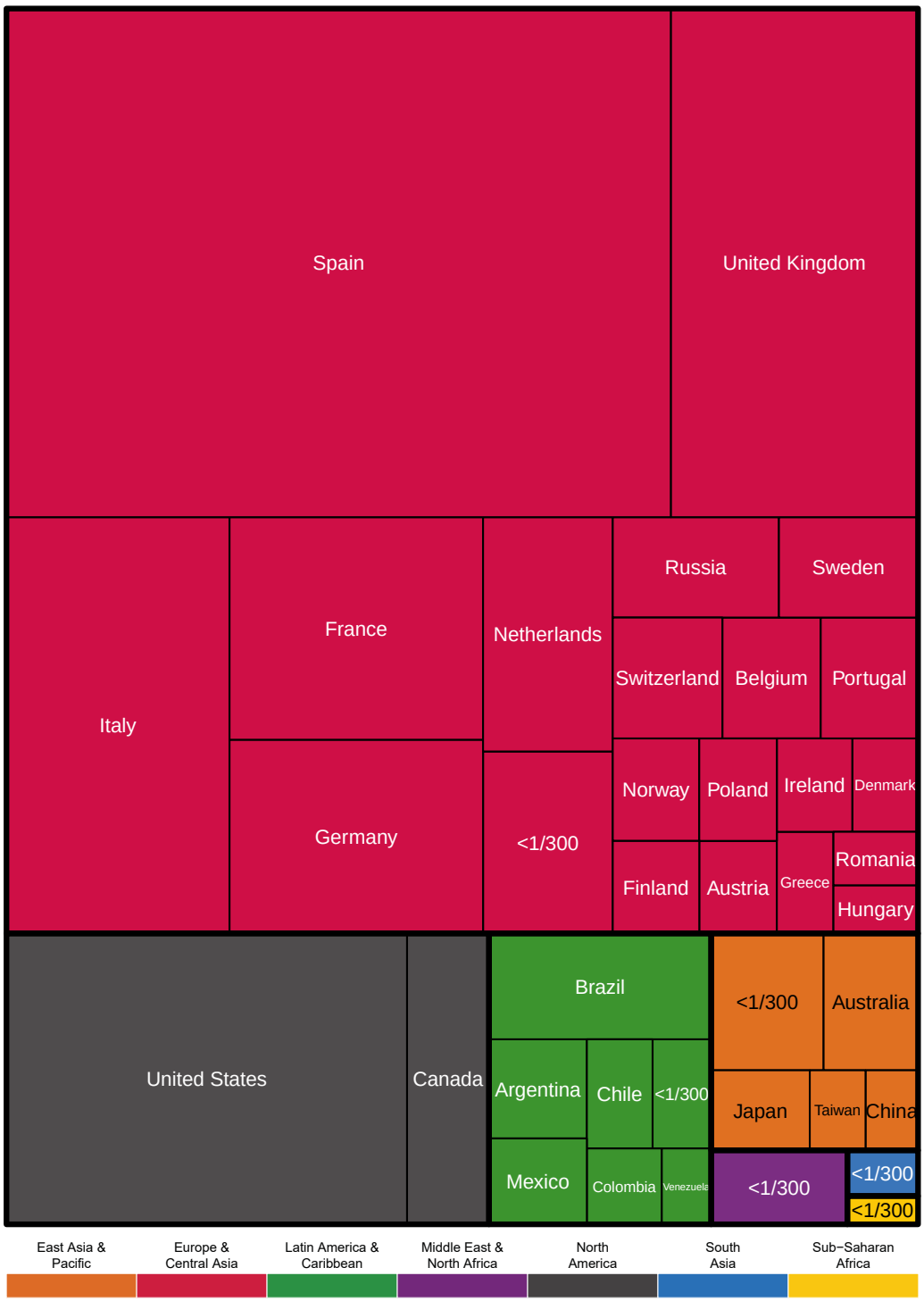


Figure 5.13: Treemap of the number of Flickr users with geotagged pictures of Barcelona, classified according to their country of origin. The countries are grouped according to their region following the World Bank classification, and colored with the same scheme as figure 4.4 on page 113. In each region, countries with less than 1/300 of total users are collapsed into a single tile.

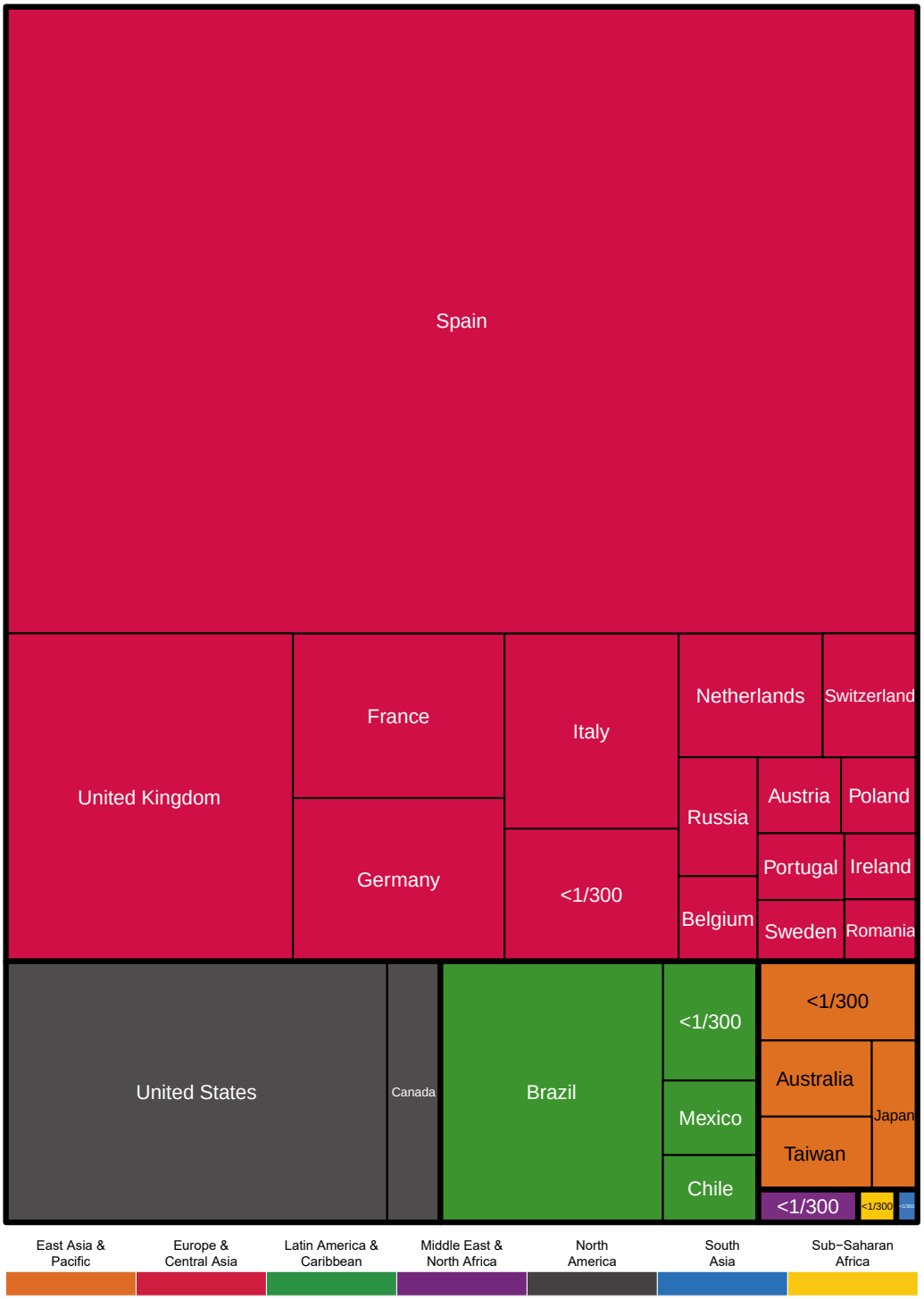


Figure 5.14: Treemap of the number of geotagged pictures of Barcelona posted on Flickr, classified according to the country of origin of the author. The countries are grouped according to their region following the World Bank classification, and colored with the same scheme as figure 4.4 on page 113. In each region, countries with less than 1/300 of total pictures are collapsed into a single tile.

Table 5.4: Countries with the highest median geotagged pictures of Barcelona posed on Flickr per user, excluding countries with 15 users or less.

Country	Users	Median
Taiwan	69	12
Wales	46	10
South Korea	21	9
Israel	44	8
Japan	122	7
Spain	5229	6
United States	1803	6
England	1543	6
Switzerland	205	6
Finland	124	6
China	65	6
Venezuela	58	6

In addition, a log/log linear regression line (white) with a 0.95 confidence level (shaded area) was included. Of the countries with 20 users or more, and the ones[22] with the highest and lowest ratios were labeled and colored accordingly (blue for highest, red for lowest), with their ratio defined as the quotient of the logarithms of pictures and users.

The Pearson correlation between the number of users and their pictures was very high (0.95). The majority of the countries with a large number of pictures —among them Spain— had a low ratio according to this metric, as their users collectively produced many pictures, but individually produced fewer than the average.

[22]Top and bottom 20.

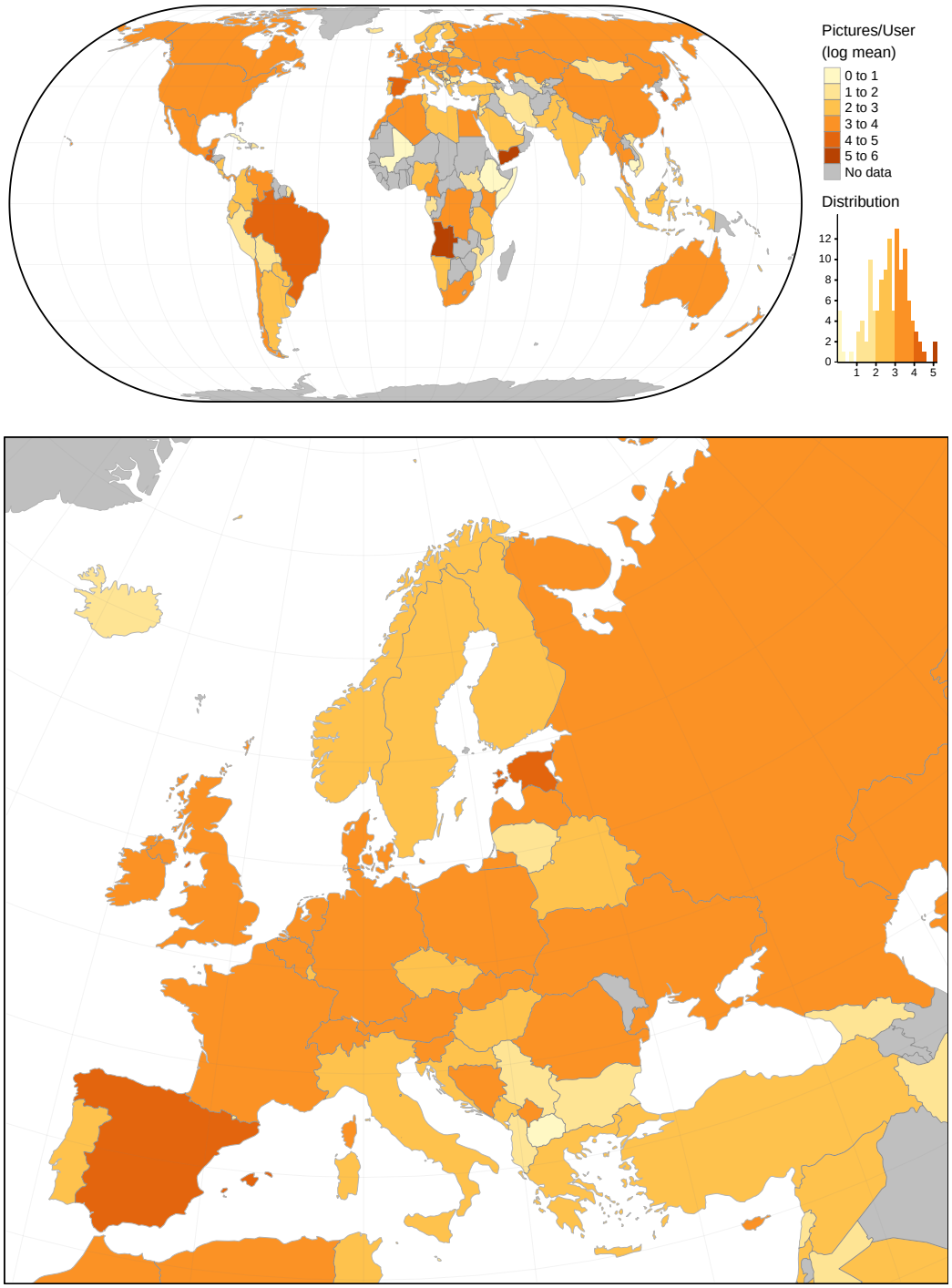


Figure 5.15: Choropleth maps of the average number of geotagged pictures of Barcelona posted by Flickr users per country. Color breaks follow a base-10 logarithmic scale. Some countries with overseas territories are broken into map units. World map uses the equal-area World Eckert IV projection (ESRI:54012), Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on cartography from Natural Earth, in the public domain.

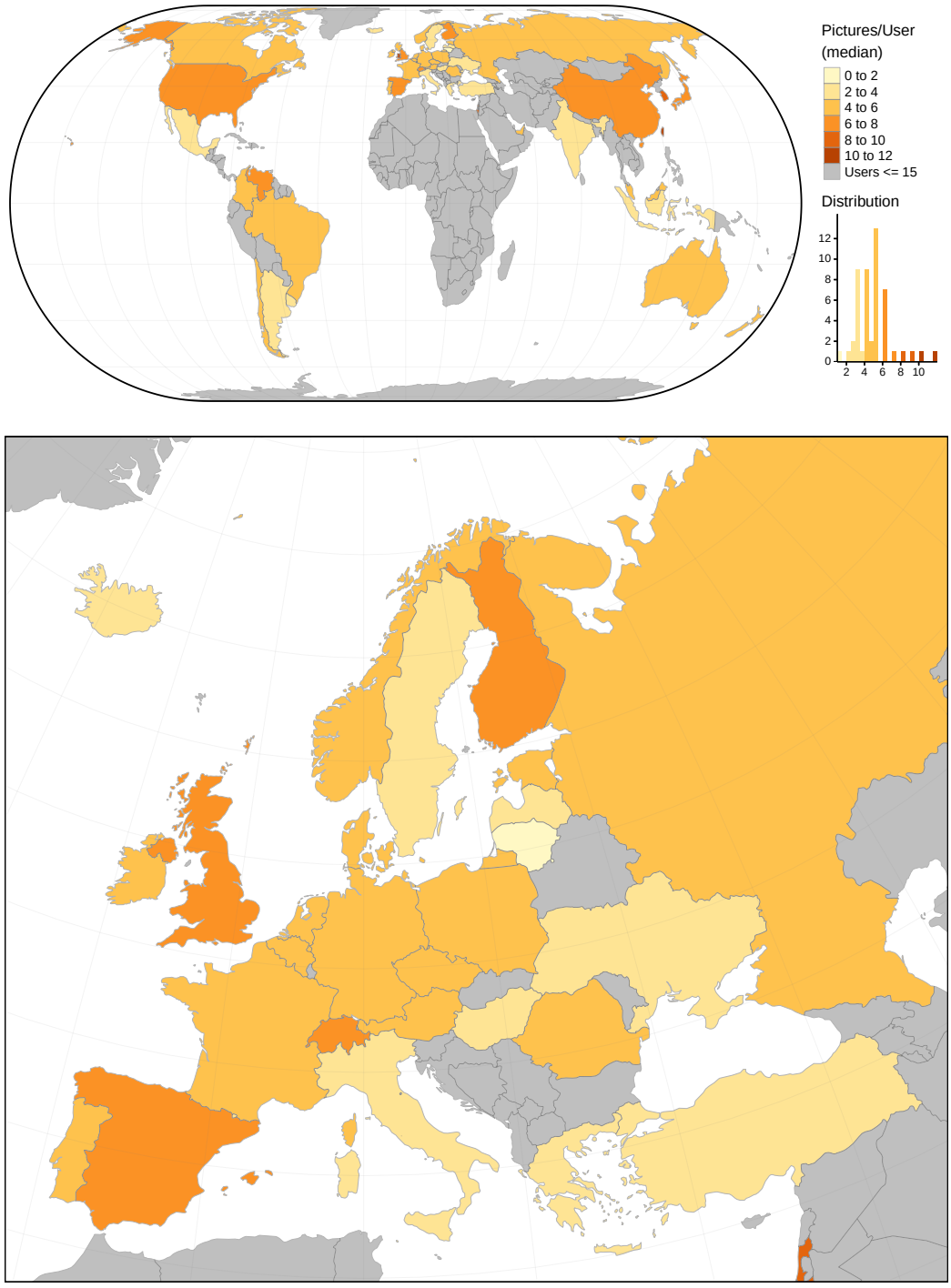


Figure 5.16: Choropleth maps of the median number of geotagged pictures of Barcelona posted by Flickr users per country. Countries with 15 users or less are excluded. Some countries with overseas territories are broken into map units.World map uses the equal-area World Eckert IV projection (ESRI:54012), Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on cartography from Natural Earth, in the public domain.

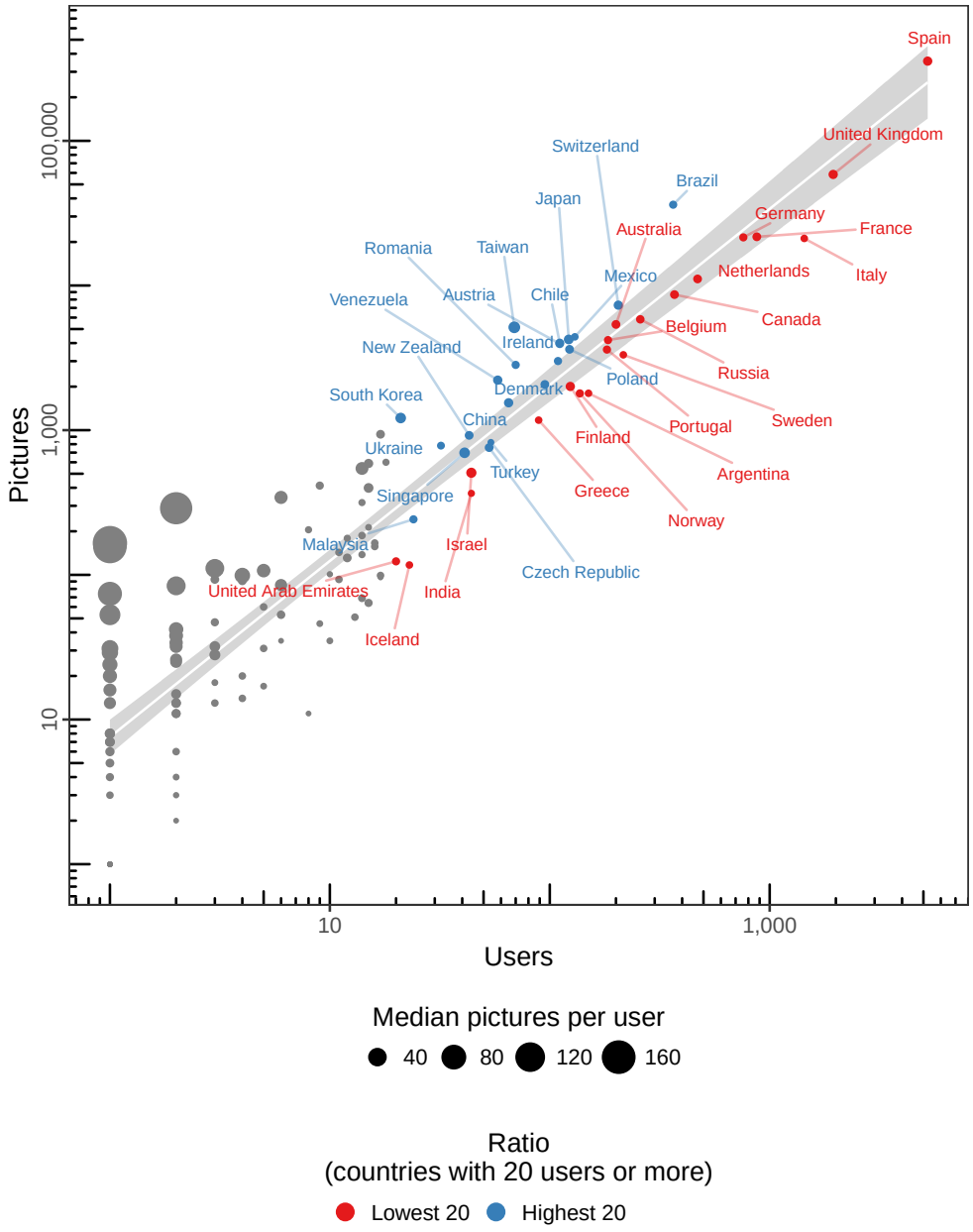


Figure 5.17: Scatter plot of the number of pictures per user, for each country of residence of users who posted a geotagged picture of Barcelona on Flickr. Point area is proportional to the median number of pictures of all users sharing the same country. The 20 countries with 20 users of more with the largest or smallest pictures/user ratio are colored and labeled. Axes use a base-10 logarithmic scale.

5.5 Influence of the Country Population

5.5.1 Geographic Distribution per Population

The analysis discussed earlier in section 5.4 does not take into account potential confounding variables, and treats all countries equally. However, more populated countries would be expected to be the origin of a a larger number of users —all other factors being equal— under the assumption that the decision to take a picture of Barcelona is independent of the country of origin.

It would be expected that when adjusted by population the countries would yield a similar user/population or picture/population ratios, otherwise indicating that there were additional factors involved, such as:

- Countries were users take proportionally more pictures (discussed in section 5.4)
- GDP of the country, which influences disposable income of the population and therefore their capacity to travel or access the Internet (discussed in section 5.6)
- Differences in the popularity of the Flickr platform across countries
- Geographic proximity to Barcelona (discussed in section 5.3)

To verify this assumption, using the population estimates of the Natural Earth data, two maps were produced for the ratio of users (Fig. 5.18) and pictures (Fig. 5.19) relative to the population of each country, showing that it did not hold true.

The color breaks in the maps correspond to the ratio of the base-10 logarithms of the users divided by the population (Fig. 5.18) and the pictures divided by the population (Fig. 5.19) respectively, per country. This transformation makes the distribution of this ratio roughly normal, as shown below the color legends in the corresponding histograms, colored to match the same intervals.

5.5.2 Number of Users and Population

Although the number of pictures and users per country were fairly proportional (Fig. 5.17), the number of pictures was more sensitive to outliers than the number of users. Therefore, to explore the connection between the country population and the collected Flickr data (users and pictures), the scatter plot (Fig. 5.20) —further broken down per world region (Fig. 5.21)—, focused on the user counts.

For each country, the population was placed in the horizontal axis, and the number of users in the vertical axis, while the number of pictures was proportional to the point size. Both axes used a base-10 logarithmic scale. The points were colored

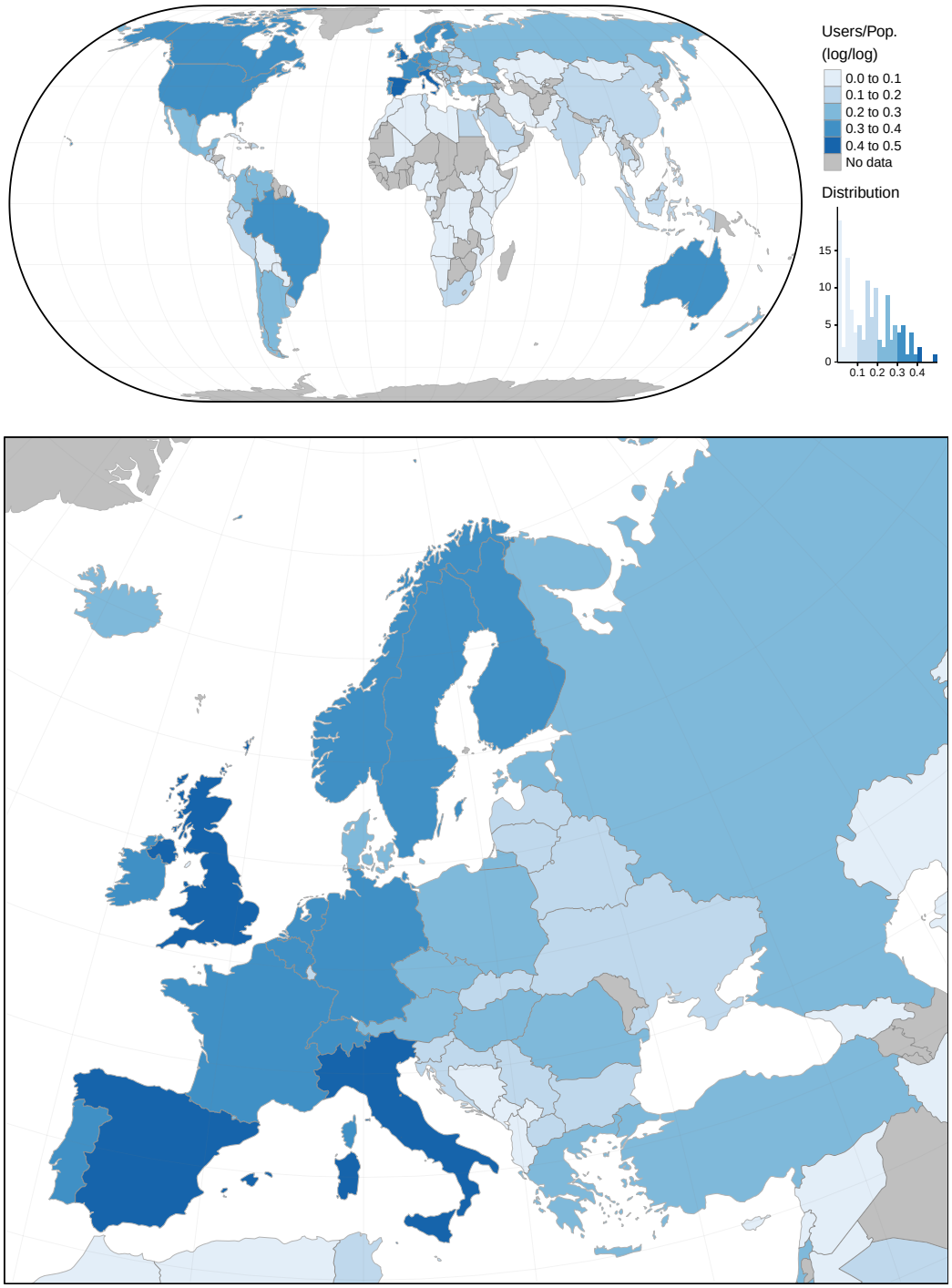


Figure 5.18: Choropleth maps of the number of Flickr users with geotagged pictures of Barcelona per population of their country of origin. Color breaks follow a base-10 logarithmic scale. Some countries with overseas territories are broken into map units. World map uses the equal-area World Eckert IV projection (ESRI:54012), Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on data and cartography from Natural Earth, in the public domain.

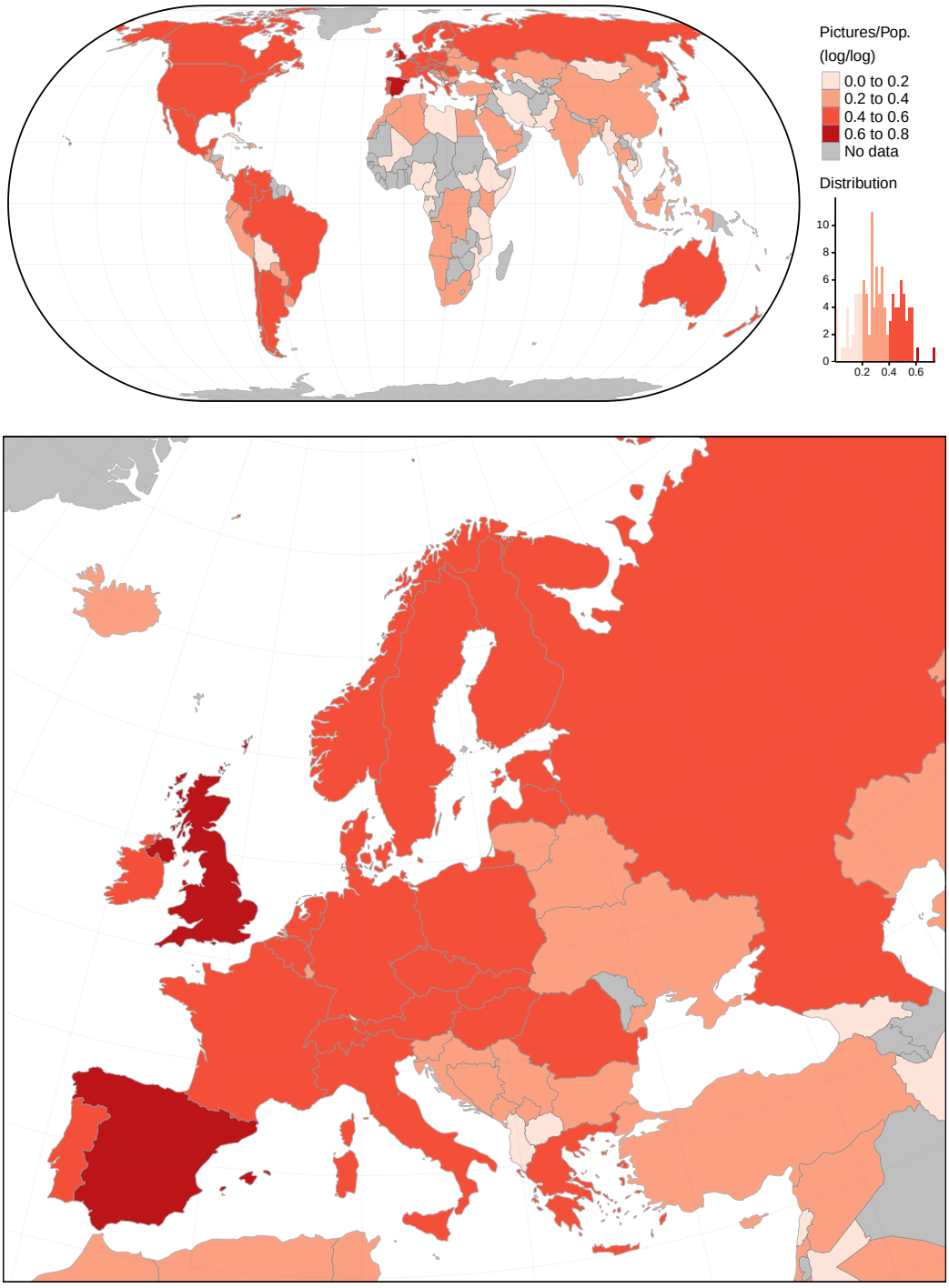


Figure 5.19: Choropleth maps of the number of geotagged pictures of Barcelona posted on Flickr per population of the country of origin. Color breaks follow a base-10 logarithmic scale. Some countries with overseas territories are broken into map units. World map uses the equal-area World Eckert IV projection (ESRI:54012), Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on data and cartography from Natural Earth, in the public domain.

according to the corresponding world region[23]. The outstanding countries were labeled in both plots:

- The top 20% countries with the highest ratio, using a base-10 logarithm (Fig. 5.20).
- The 5 countries with more users on each region[24] (Fig. 5.21), with the Americas placed on the same facet.

As seen in the graphs, the countries with the highest ratios were in the European region and had a proportionally large number of pictures. In both figures the relationship between population and users had an upward trend —except Middle East & North Africa— which was statistically significant for all regions except two in the African continent (Middle East & North Africa and Sub-Saharan Africa). However the Pearson correlation of population and number of Flickr users was almost zero (0.06).

[23]Using the same color scheme as figure 4.4 on page 113.
[24]Except North America, who contained only two countries (United States and Canada).

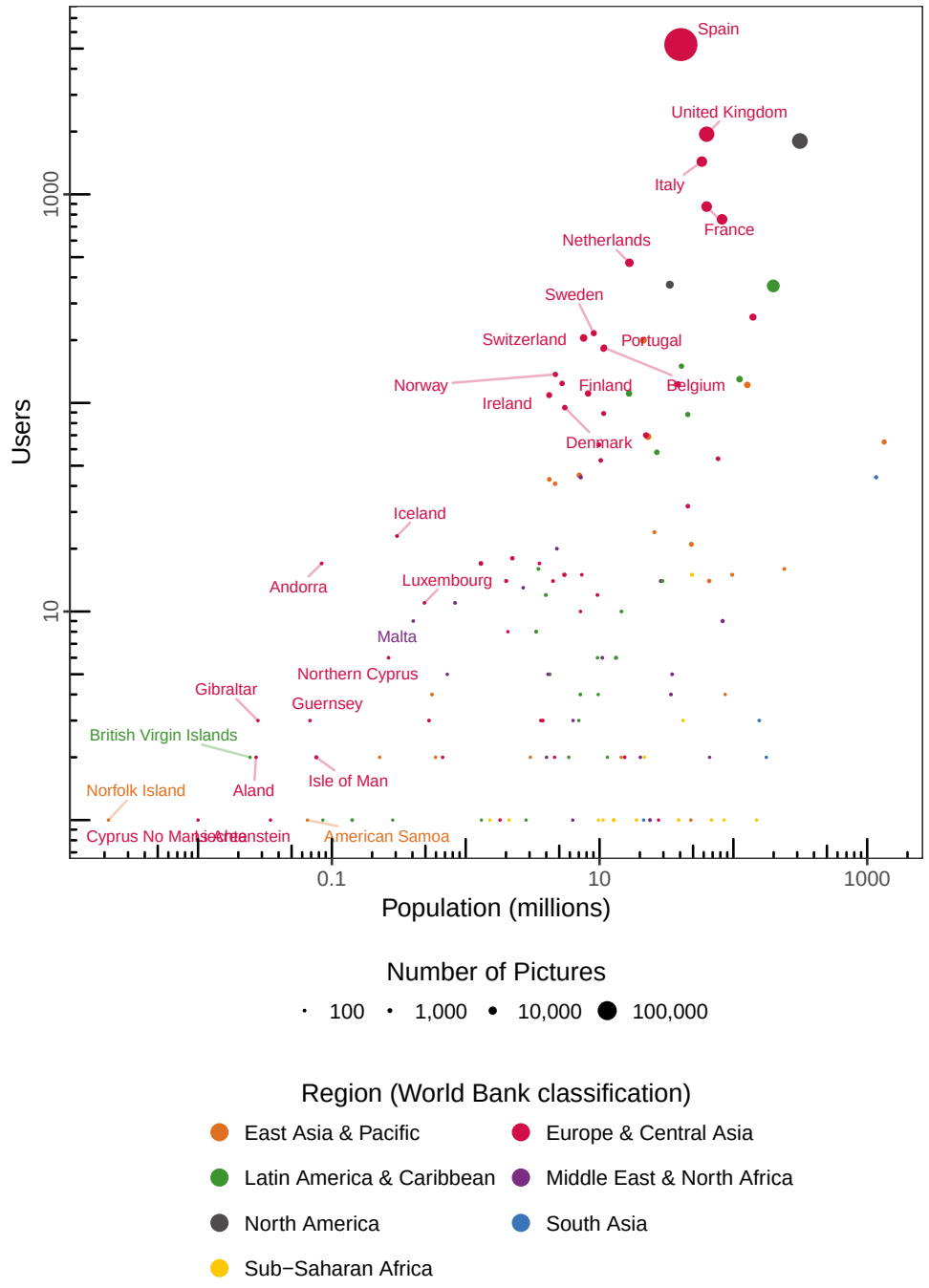


Figure 5.20: Relationship between country population and number of Flickr users who posted a geotagged picture of Barcelona. Axes use base-10 logarithmic scales. Only the top 20% countries with the highest user/population ratio are labeled. Colors of labels and points match the regions in figure 4.4 on page 113. Size of the point proportional to the number of pictures taken by users from the country. Source: Own work.

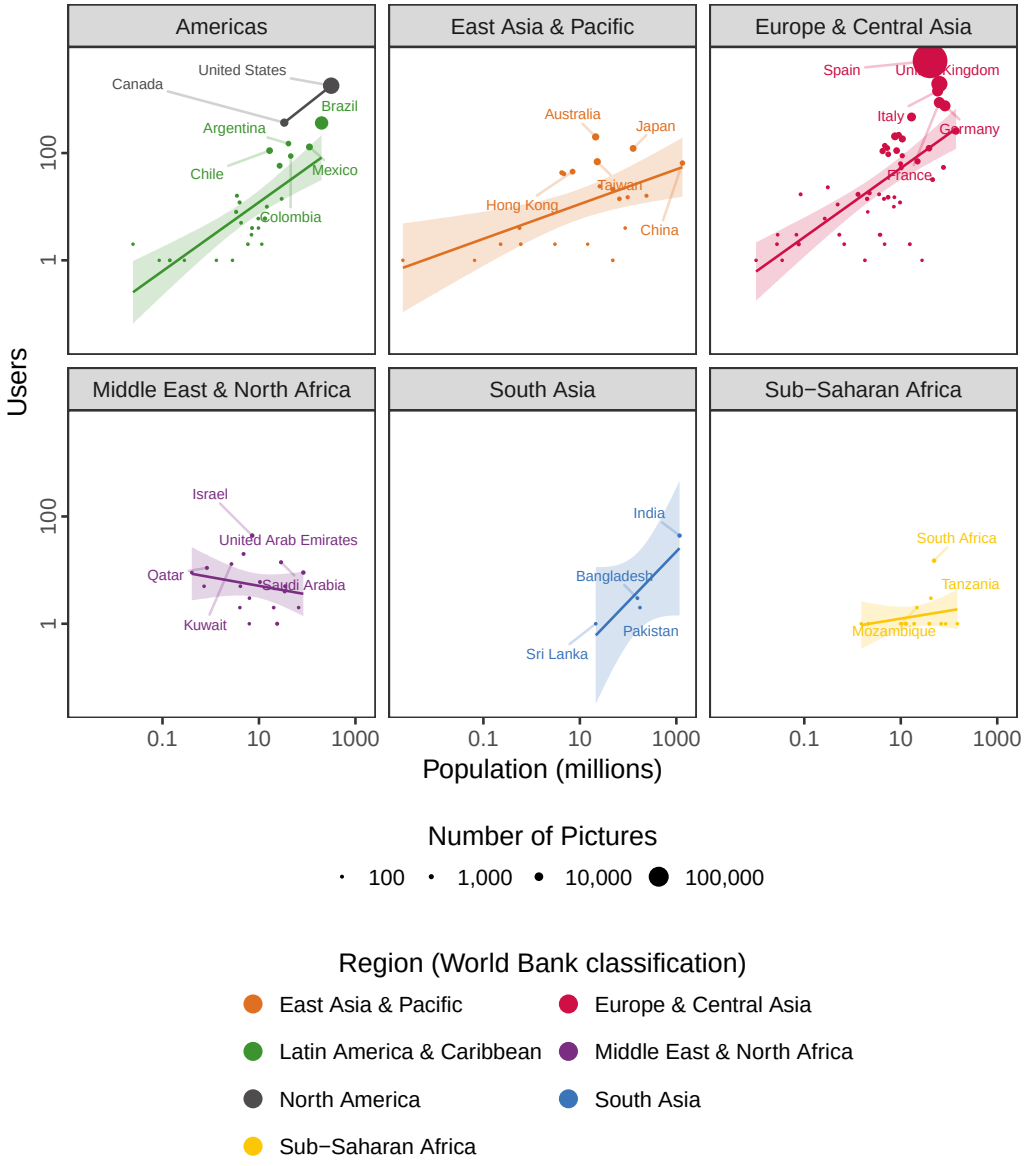


Figure 5.21: Relationship between country population and number of Flickr users who posted a geotagged picture of Barcelona. Axes use base-10 logarithmic scales. Regions are plotted separately, except the Americas. The 5 countries of each region with more users are labeled. Colors of labels and points match the regions in figure 4.4 on page 113. Size of the point proportional to the number of pictures taken by users from the country. Source: Own work.

5.6 Influence of the Country GDP

5.6.1 Geographic Distribution per GDP

Beyond population (discussed in section 5.5), another factor that could influence the amount of users was the economic output of the country, as more affluent users could afford to travel to Barcelona, all other factors being equal.

Under this assumption, the number of users would be proportional to the GDP of the country, provided their distance (discussed in section 5.3), the population (discussed in section 5.5), the average pictures per user (discussed in section 5.4) and the popularity of the service offered by Flickr were equal across the countries.

Using the GDP estimates of the Natural Earth data, two maps were produced for the ratio of users (Fig. 5.22) and pictures (Fig. 5.23) relative to the estimated GDP of each country[25], similarly to the method used for the population.

The color breaks of the maps used a base-10 logarithmic scale applied to the ratio of number of users or pictures (numerator) and GDP (denominator). The transformed distribution was roughly normal, as seen in the histogram below the legend, with colored bands matching the corresponding legend entries.

5.6.2 Number of Users and GDP

The influence of the GDP of the country on the number of users and the pictures taken by them was explored using a scatter plot (Fig. 5.24), as well as the influence of the world region (Fig. 5.25).

In the plots the GDP was on the horizontal axis and the number of users in the vertical axis, while the number of pictures was proportional to the size of the point. Both axes used a base-10 logarithmic scale. Each point color corresponded to the world the region of the country[26]. In both plots only the outstanding countries were labeled:

- The top 20% countries with the highest ratio, using a base-10 logarithm (Fig. 5.24).
- The 5 countries with more users on each region[27] (Fig. 5.25), with the Americas sharing a single facet.

In both figures, the highest ratios corresponded to European countries. As in the case of the population, an upward trend was visible, but using GDP as an explanatory variable showed a more pronounced connection than using the

[25] Measured in millions USD.
[26] Following the color scheme of figure 4.4 on page 113.
[27] If applicable.

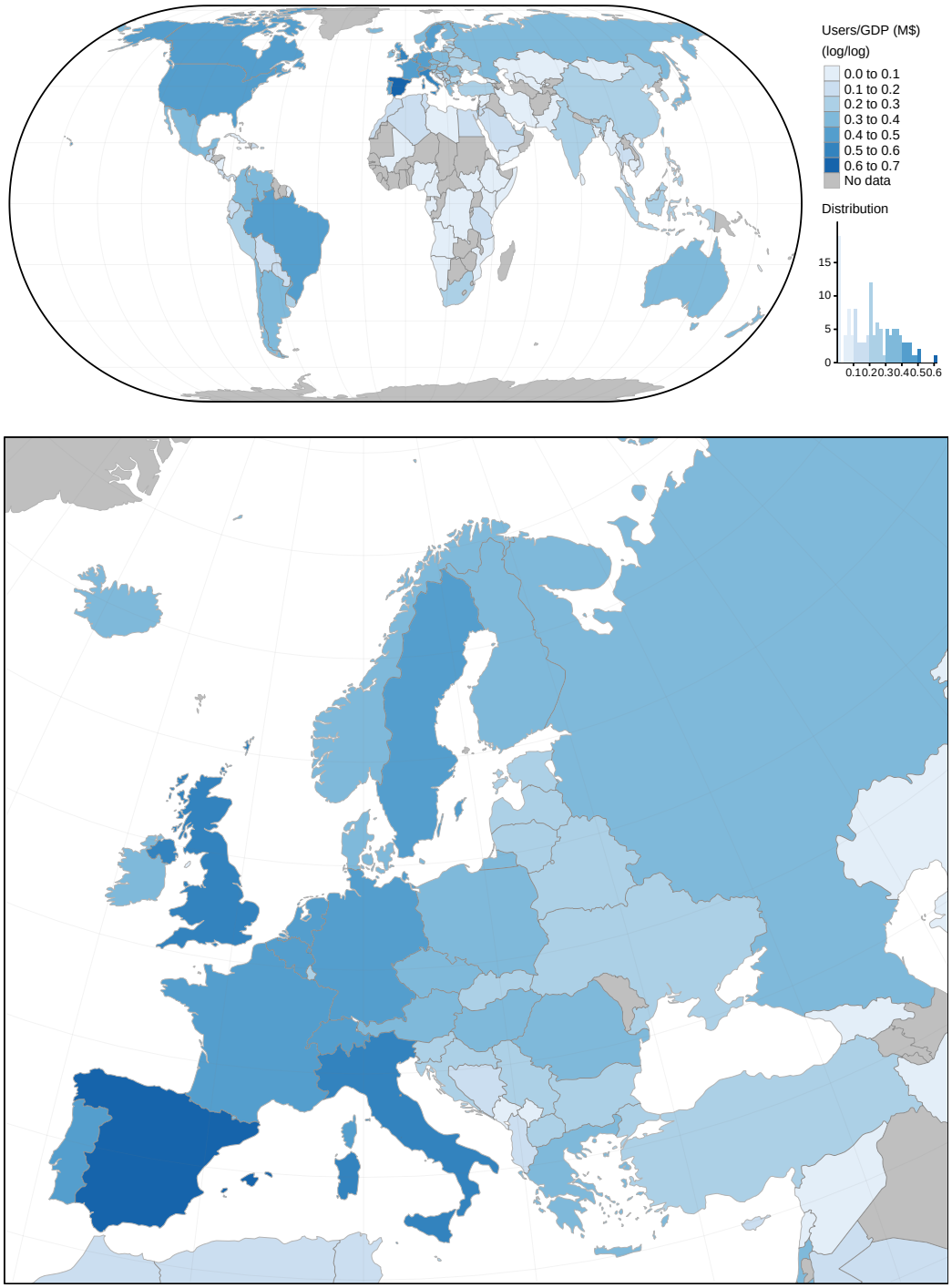


Figure 5.22: Choropleth maps of the number of Flickr users with geotagged pictures of Barcelona per GDP of their country of origin. Color breaks follow a base-10 logarithmic scale. Some countries with overseas territories are broken into map units. World map uses the equal-area World Eckert IV projection (ESRI:54012), Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on data and cartography from Natural Earth, in the public domain.

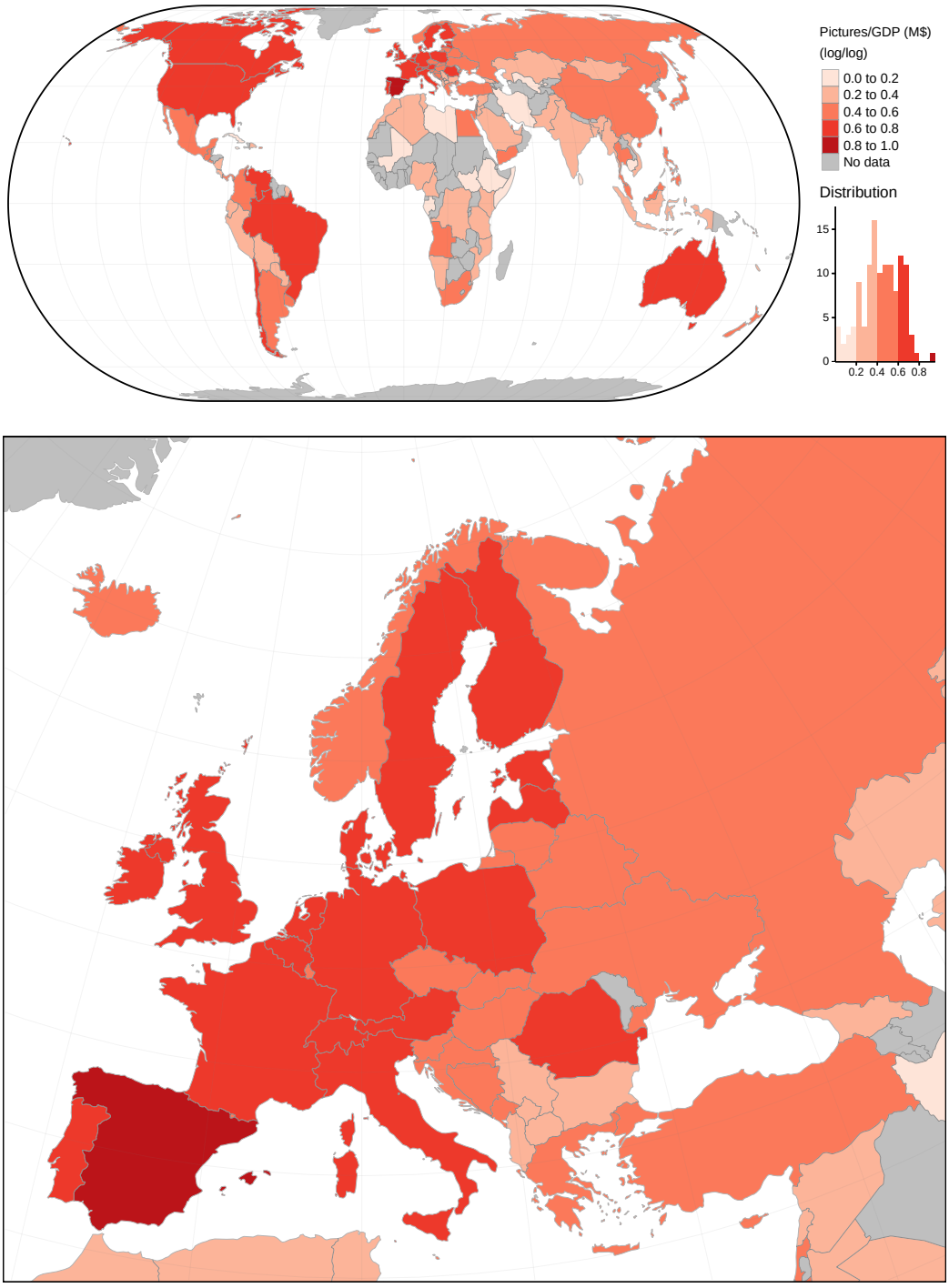


Figure 5.23: Choropleth maps of the number of geotagged pictures of Barcelona posted on Flickr per GDP of the country of origin. Color breaks follow a base-10 logarithmic scale. Some countries with overseas territories are broken into map units. World map uses the equal-area World Eckert IV projection (ESRI:54012), Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on data and cartography from Natural Earth, in the public domain.

population (Fig. 5.20), suggesting that GDP could be a better predictor of the number of users, although the Pearson correlation of GDP and number of users was only 0.37.

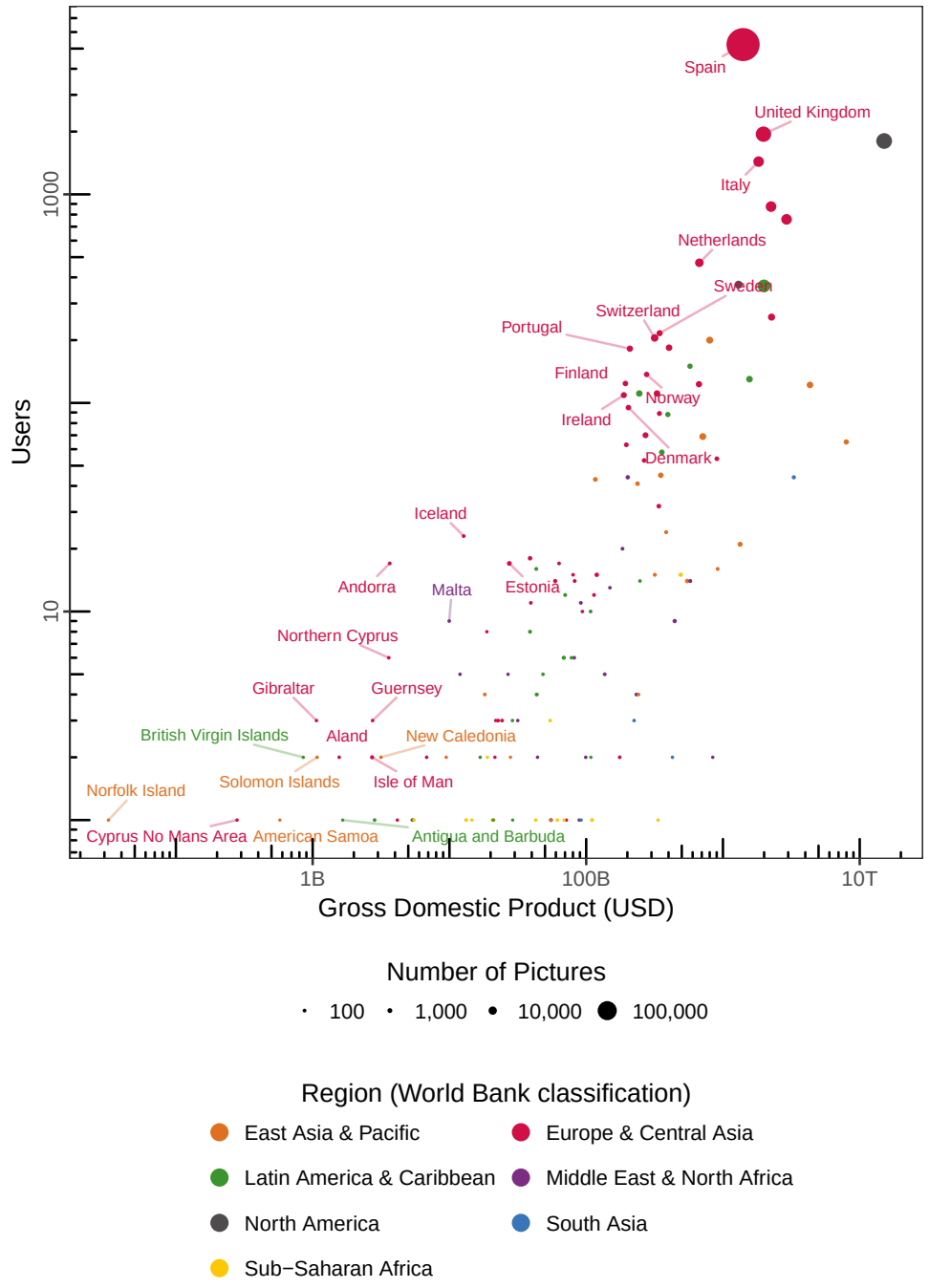


Figure 5.24: Relationship between country GDP and number of Flickr users who posted a geotagged picture of Barcelona. Axes use base-10 logarithmic scales. Only the top 20% countries with the highest user/GDP ratio are labeled. Colors of labels and points match the regions in figure 4.4 on page 113. Size of the point proportional to the number of pictures taken by users from the country. Source: Own work.

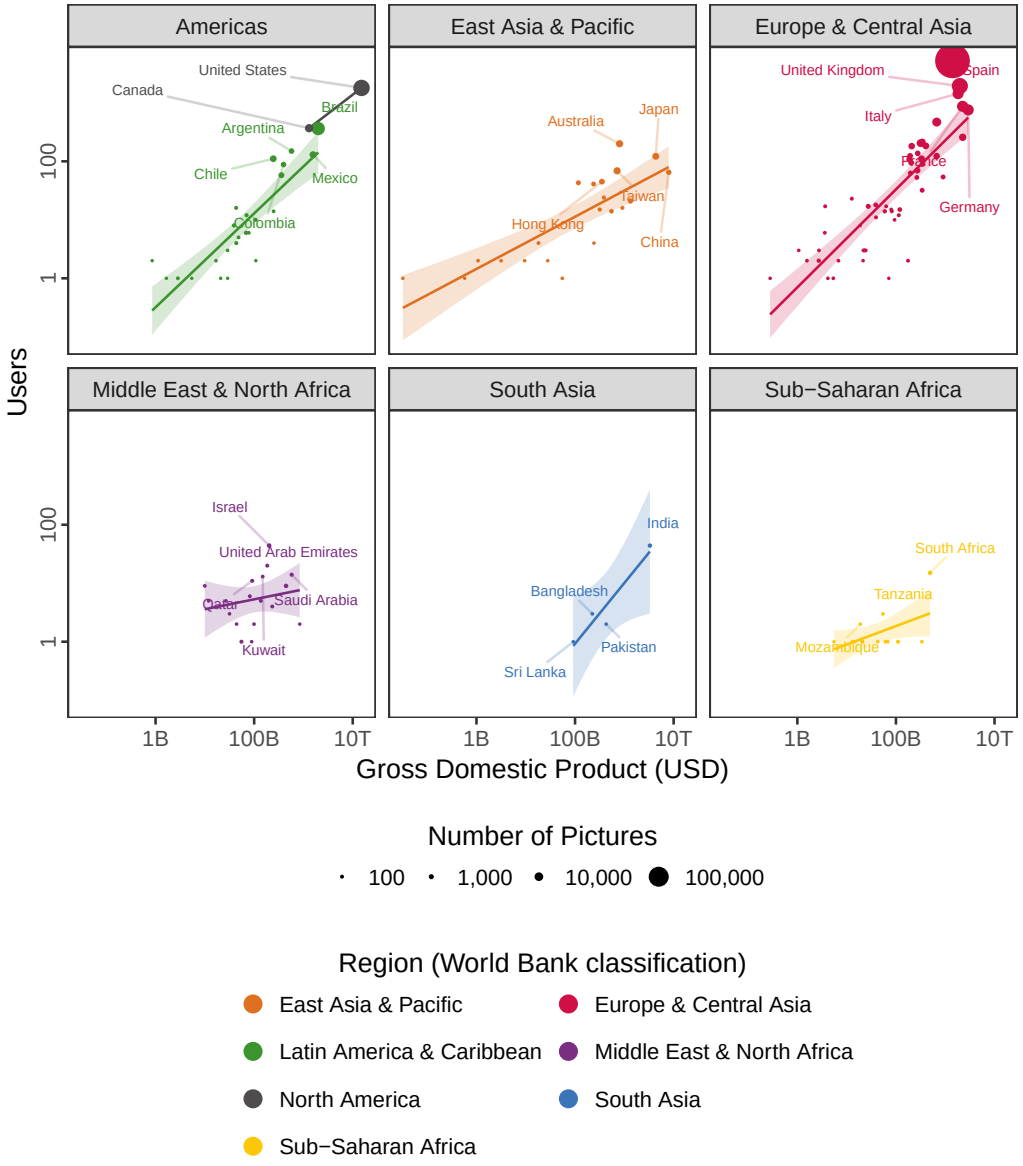


Figure 5.25: Relationship between country GDP and number of Flickr users who posted a geotagged picture of Barcelona. Axes use base-10 logarithmic scales. Regions are plotted separately, except the Americas. The 5 countries of each region with more users are labeled. Colors of labels and points match the regions in figure 4.4 on page 113. Size of the point proportional to the number of pictures taken by users from the country. Source: Own work.

5.7 Expected Income Estimation

5.7.1 Expected Income

As an example of a practical application, this section proposes a method based on the retrieved user data to determine the expected contribution of visitors to the economic growth of their destination according to their country of origin, under the assumption that the expected income from any country would be proportional to the number of Flickr users who visited —and took and uploaded a picture of— the city. The results can only provide a ballpark estimate, because of the impossibility to access all available data, but otherwise should be applicable in similar cases.

Using both the population data and the GDP of a country, a derived metric —the GDP *per capita*, roughly equivalent to the purchasing power parity (PPP)— was defined as the quotient between the GDP of the country and its population. Both predictors were weakly correlated, with a Pearson correlation of 0.58.

Therefore, it was possible to compute the the fraction of users in relation to the total for any country. This fraction was weighted using the corresponding PPP value calculated previously as the product of both values.

The proportion of the result of this calculation, divided by the sum of all products computed for each country was the expected income of the country. The expected income proportion I for country i could therefore be expressed as:

$$I_i = \frac{u_i \frac{g_i}{p_i}}{\sum_{j \in U} u_j \frac{g_j}{p_j}}$$

Being the following items in the formula above:

I as the expected income expressed as a probability (between 0 and 1).
i as the country for which the calculation was being performed.
u as the count of Flickr users.
U as the set of all counts of Flickr users of each country.
g as the Gross domestic product.
p as the population.

And therefore, the sum of all probabilities of all countries considered in the calculation would be one:

$$\sum_{j \in U} I_j = 1$$

Table 5.5: Top ten countries with the highest expected income from visitors, according to the number of Flickr users with geotagged pictures of Barcelona in each country and its GDP *per capita*.

Country	Region (World Bank)	Percent
Spain	Europe & Central Asia	31.48 %
United States	North America	15.07 %
United Kingdom	Europe & Central Asia	10.59 %
Italy	Europe & Central Asia	7.83 %
France	Europe & Central Asia	5.41 %
Germany	Europe & Central Asia	4.68 %
Netherlands	Europe & Central Asia	3.29 %
Canada	North America	2.49 %
Switzerland	Europe & Central Asia	1.48 %
Sweden	Europe & Central Asia	1.43 %

5.7.2 Geographic Distribution of the Expected Income

The computed expected incomes per country were visualized on a map (Fig. 5.26), where the countries were colored according to their corresponding percentage. As seen in the distribution of the percentages, displayed in the histogram below the color legend, the distribution is positively (right) skewed.

The skewness of the distribution (7.17) placed the majority of the values in the less than 1% category and could result in difficulties distinguishing any country except the ones placed in the two top categories (Spain and United States).

5.7.3 Top Countries per Expected Income

In the detail map of Europe (Fig. 5.26), some countries appear to have a larger proportion of expected income. However, since the quantities are small, the colors of the map can be misleading because very different values can be clumped together in the same category, such as Germany and Sweden (Table 5.5).

In this case, it can be more adequate to use a dot plot (Fig. 5.27); while not using a map disconnects data from their location, in the case of countries the readers can be expected to be familiar with their location, making this issue almost moot while allowing a more explanatory representation of the data.

The positions of the dots were accompanied by their corresponding expected proportions, showing that a third of the income could be expected to be domestic, with significant contributions from United States and United Kingdom, followed

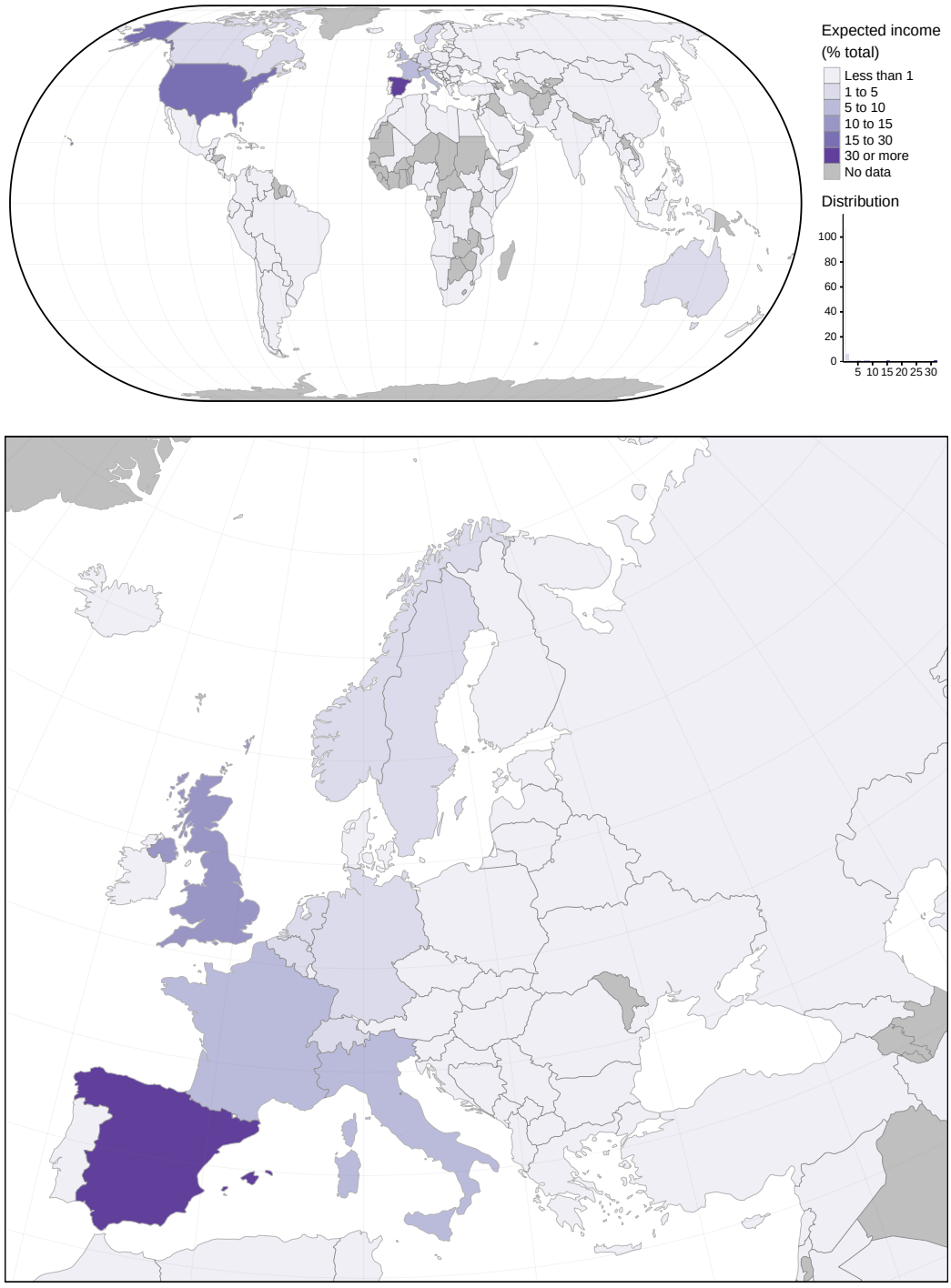


Figure 5.26: Choropleth maps of the fraction of the expected income from visitors per country, according to the number of Flickr users with geotagged pictures of Barcelona and its GDP *per capita*. Some countries with overseas territories are broken into map units. World map uses the equal-area World Eckert IV projection (ESRI:54012), Europe map uses the ETRS Lambert Azimuthal Equal-Area projection (EPSG:3035). Source: own work based on data and cartography from Natural Earth, in the public domain.

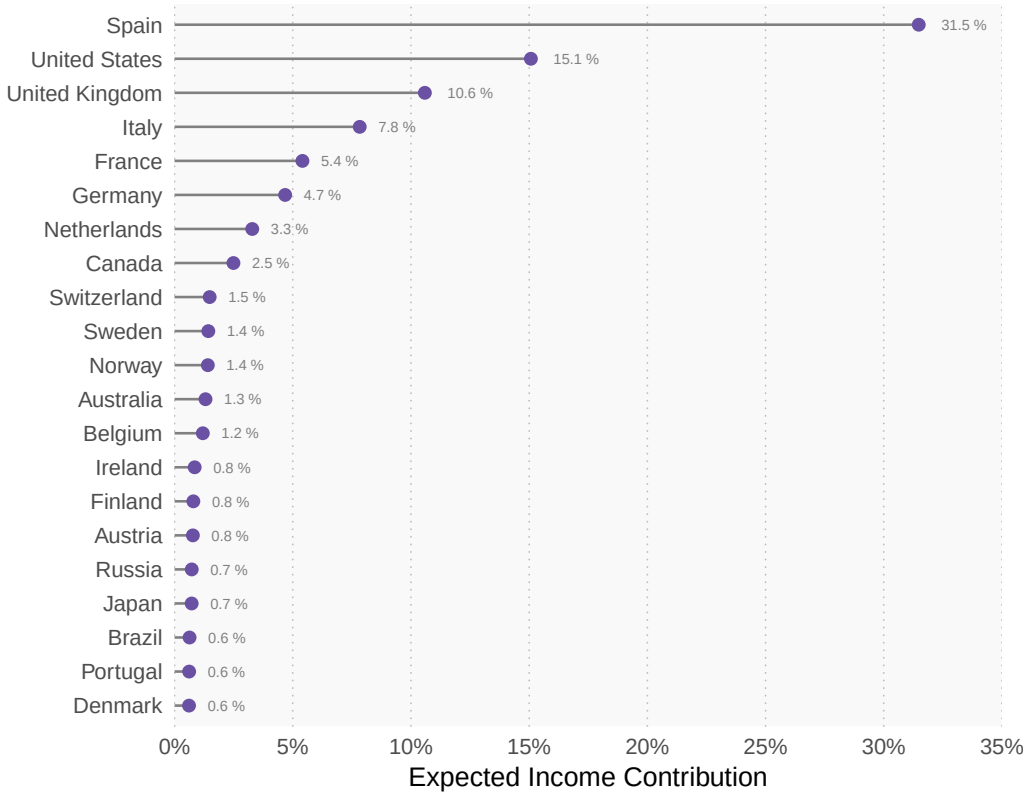


Figure 5.27: Dot plot of the proportion of the expected income from visitors per country, according to the number of Flickr users with geotagged pictures of Barcelona and its GDP *per capita*. Only countries with more than 0.5% of the total expected income are depicted.

chiefly by European countries before quickly reaching the 1% mark. Countries below the 0.5% cutoff value are not shown.

5.7.4 Regional Distribution of the Expected Income

Since the expected income was measured in proportions of a whole, a treemap was considered a good compromise, losing the geographic context of the map and the precision of the dot plot, but allowing comparing more countries in less space, while retaining some geographic context (grouping and coloring the tiles according to their region[28]).

Countries whose expected income were less than 1/300 of total were grouped

[28]The tiles used the same color scheme as figure 4.4 on page 113.

in a special category, separately for each region. This approach allowed the comparison of countries as well as regions, and even cross-comparisons such as the proportion expected from the UK compared to the whole South American region (Fig. 5.28).

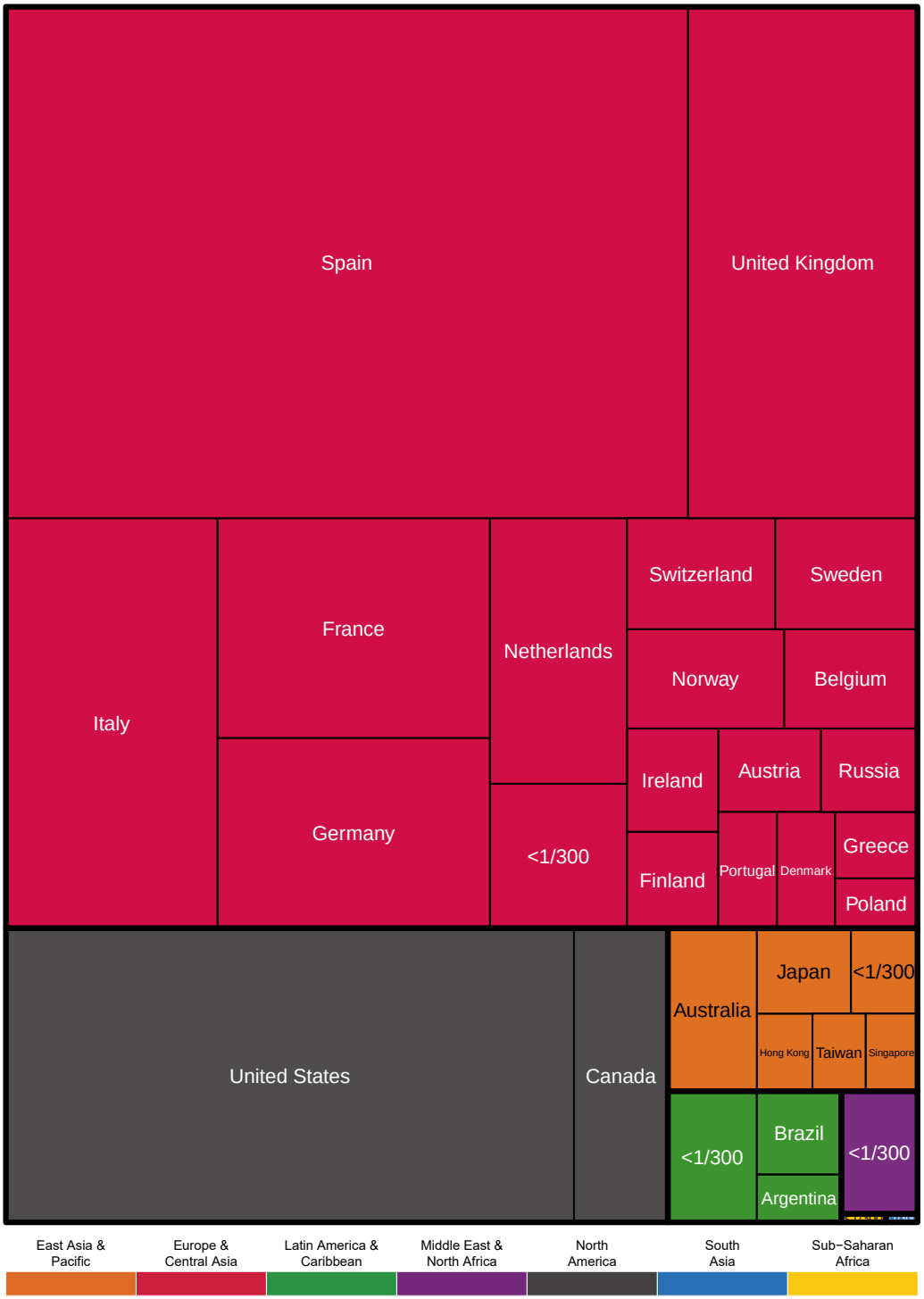


Figure 5.28: Treemap of the expected income from visitors, proportional to the product of the number of Flickr users with geotagged pictures of Barcelona in each country and its GDP *per capita*. The countries are grouped according to their region following the World Bank classification, and colored with the same scheme as figure 4.4 on page 113. In each region, countries with less than 1/300 of total are collapsed into a single tile.

Chapter 6

The City Scale

> "This City is what it is because our citizens are what they are"
>
> Plato

6.1 Introduction

6.1.1 Background

This chapter focuses on the urban scale, which for the purposes of this study will be defined as a smaller scope than the metropolitan scale —where main subject of interest is focused in its functional aspects as a complex system—, but larger than the neighborhood scale (characterized by a relatively homogeneous morphology and functional mix), which will be discussed in chapter 7.

The urban scale is the natural frame of reference of urban planning policies, which must deal with the articulation of the city within its metropolitan (and sometimes global) context, but also the dynamics of the different realities of its constituent neighborhoods.

The big issue is therefore the definition of the limits of the city, whose boundaries will inevitably be in the opinion of the author very fuzzy and subjective, and as such will always be subject of endless debate[1]. When faced with this conundrum, the author took the most dispassionate route adopting the official city limits as the area of study, seeking a consensus which will probably neither please nor be completely rejected by anybody.

[1] And a without doubt the subject of another research.

Multiple methodologies are discussed in the chapter, to be able to extract knowledge from the spatial distribution of events (pictures, messages) and locations, with the objective of identifying structures within the urban fabric, providing responses while generating new questions.

6.1.2 Urban Data

Comprehensive data acquisition it is expensive to collect and keep up-to-date. This capacity is only within the reach of governments or large corporations, which are capable of providing long-term funding to these endeavors. As a result, in the case of data collected by public entities, some of these data are released free of charge.

Therefore, data-driven approaches to urban-scale analysis tend to rely on officially-sourced information such as cadastral data [89], a country-level comprehensive register of real-estate properties. This approach is motivated because of practical reasons, as cadastral data is generally the only viable source of consistent urban data for municipal or supra-municipal scales.

However, the objectives of urban research do not always align perfectly with the goals of the data collection, which in the case of the Spanish Cadastre[2] are the valuation of real-estate properties for taxation or expropriation purposes.

From the knowledge acquired researching the suitability of cadastral databases [6] for urban research, which despite their obvious advantages are limited in scope by the very nature of the objectives of the collected data, this research focuses in analyzing urban phenomena from data retrieved from social networks. While it is true that this approach has its own limitations, it should complement —but not replace— data from official sources and provide a richer context for urban analysis.

Another limitation of cadastral data is that it is parcel-based, in the sense that for all intents and purposes the smallest individually addressable spatial unit is the parcel. As a consequence, all available data must be transferred to a parcel to be spatially analyzed, a process that requires summarizing any available data per parcel (e.g. counting, adding or computing a proportion).

6.1.3 Visualizing Urban-Scale Data

This chapter discusses three approaches developed to visualize in a principled way the spatial distribution of the point patterns corresponding to the four collected data sources: the picture locations of Barcelona retrieved from Panoramio and

[2]The Spanish Cadastre website is available at *http://www.catastro.meh.es/* at the time of writing.

Flickr, the status messages retrieved from Twitter over a period of one year, and the "places" retrieved from Instagram.

In this case, in contrast with cadastral data, where parcels drive the scale of the spatial analysis, there is not a predefined aggregation unit. Therefore, the choice of this unit, and as a consequence the scale of the spatial analysis itself, must be evaluated carefully (a complementary approach, using the road network as the topological space is discussed in chapter 8).

The sources were very different in the nature of the events collected, the number of the points registered and also their accuracy. To study their distinct spatial distributions, they were used as a workbench to develop methodologies applicable to a wide range of urban visualization research problems, and distill a gigantic pile of data into informative and visually engaging maps, using three *complementary* approaches:

- Rasterization (section 6.2).
- Local Measures of Spatial Autocorrelation (section 6.3).
- Smoothing Estimation of Intensity (section 6.4).

Special care was taken that the parameters of the discussed methodologies were robust and derived from the properties of data when possible [301], with the objective of avoiding biases that could hinder the interpretation and comparison of the results.

6.2 Rasterization

6.2.1 Representation Challenges

When visualizing large spatial data sets consisting in point features —either with or without attributes—, what is being plotted is essentially the scatter plot of the corresponding longitude (x) and latitude (y) coordinates of every single point in the dataset[3].

As the size of these data sets becomes larger (Table 3.5), the maximum resolution of the output device becomes increasingly incapable of resolving the necessary detail (Table 6.1), because of the limited availability of individually addressable elements in the output device (total pixel or dot count), and capacity of discriminating individual coordinate pairs (computed as the optimal case of a regular grid of points separated by exactly one empty pixel in either direction). These issues become even more challenging when the points are clustered instead of dispersed,

[3]This "naive" approach was used in the results overview in chapter 3, in figures 3.5, 3.7, 3.12 and 3.9.

Table 6.1: Addressable and effective resolutions of two common outputs (paper and screen).

Output	Addressable resolution	Effective resolution
A4 Paper (300 dpi)	8,700,632	2,175,158
Full HD Monitor	2,138,400	534,600

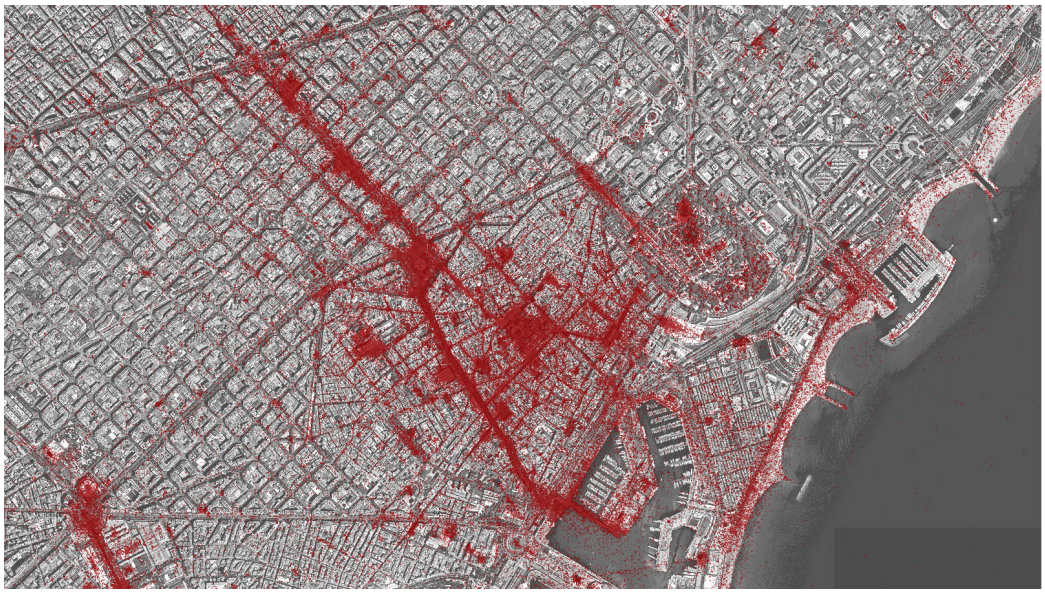
as generally occur in the case of geospatial data such as the ones discussed in this research.

In this context, the main challenge when dealing with large data sets composed of spatial points is overplotting (Fig. 6.1a), which is the occlusion of multiple points that are plotted on top of each other (overdrawing), an issue that is especially critical when the plotted points belong to multiple categories and need to be distinguished by their color, shape or size.

When transparency is used to reduce the effects of overplotting —the most common approach—, oversaturation issues can arise depending on the chosen alpha (transparency) parameter, as the accumulation of points can "saturate" the alpha value (if the number of overdrawn points exceed 1/alpha) while rendering points in areas with low concentration barely distinguishable (Fig. 6.1b). Therefore, the result of this approach depends on the choice of the transparency setting but also on the quantity and spatial distribution of the points and their plotted size (Fig. 6.1).

Another popular choice is binning the points in a regular grid (rasterization or pixellation) producing a heatmap, the 2D counterpart of the 1D binning in histograms. In this context, rasterizing the picture locations to convert them to an image could be described using the terms employed in the process of capturing a digital photography:

- The picture locations would be analogous to the individual photons hitting the image sensor.
- The image sensor would need an sufficient number of "photodiodes" (pixels) to resolve the detail, but the pixels should not be too small to avoid introducing noise (discussed in section 6.2.2).
- The "exposure" and "sensitivity" should be the adequate to the amount of available dynamic range to avoid overexposed or underexposed pictures (discussed in section 6.2.3), or after the capture, using tone mapping.



(a) Collected picture locations with fully opaque point features.

(b) Collected picture locations with point features drawn with 80 % transparency.

Figure 6.1: Overview of the geotagged pictures of Barcelona retrieved from Flick overlaid (with 20% layer transparency) on an orthophoto. Each point is drawn as a circle with a diameter of 5 meters (printed size is 0.1 millimeters). Maps use the WGS 84 / Pseudo-Mercator – Spherical Mercator (EPSG:3857) projected coordinate system. Base cartography by Cartogràfic i Geològic de Catalunya (ICGC), under CC BY 4.0.

The *datashader*[4] Python library was developed[5] with the objective of addressing these challenges[6] [302], but since the development platform used in the research was based on the R programming language, the *raster* v.2.6-7 R package [303] was used for the analysis and the *ggplot2* v.2.2.1 R package [304] was used for visualization, a combination that provided a comparable solution within the R ecosystem, despite sacrificing the real-time capabilities offered by *datashader*. The raster modules of GIS sofware were also discarded because of their limited performance and flexibility.

6.2.2 Pixel Size Selection Criteria

In spatial data analysis, at least two *a priori* considerations that can have a profound impact on the results must be given proper consideration:

- The delimitation of the area of study or the analysis window (e.g. a neighborhood).
- The choice of the analysis unit (e.g. parcels).

In the case of this research, the delimitation of the area of study was coincident with the administrative limits of the city of Barcelona. The nature of this boundary is in some of its sections a "hard" physical barrier —such as along the waterfront facing the Mediterranean sea—, while in others is the result of the administrative division of the territory into municipalities, and therefore more arbitrarily defined.

The choice of analysis unit is oftentimes determined by the source data, using the the smallest spatial unit these data is aggregated into (e.g. parcels, census tracts, neighborhoods). However, this approach is known to introduce biases as a consequence of the Modifiable Area Unit Problem (MAUP) [305, 286], further discussed in chapter 7.

While the MAUP is an ill-defined problem, the tessellation of the space into a regular grid of polygons can solve some of its issues by partitioning the space in a systematic way. However, this approach is sensitive to the definition the chosen grid, in particular:

- The size of the aggregation units, compared to the size of the area of study or the spatial distribution of events.
- The shape of the polygons (in the Euclidean plane, either regular triangles, rectangles or hexagons).

[4]The datashader documentation is available at *http://datashader.readthedocs.io/* at the time of writing.

[5]A video presentation at the SciPy 2016 conference titled "Datashader: Revealing the Structure of Genuinely Big Data" discussing the issues is available at *http://youtu.be/6m3CFbKmK_c* at the time of writing.

[6]Further discussed in https://anaconda.org/jbednar/plotting_pitfalls/notebook

- The orientation of the grid.
- The grid offset relative to the boundary.

Nonetheless, as the grid becomes smaller, the distortions introduced by the chosen grid shape, orientation and offset become less important. Therefore, a square grid with the same orientation as the coordinate reference system commonly used in Barcelona (EPSG:25831) was used in the rasterization process, because its enormous computational advantages as the pixels are spatially indexed by design.

Under this premise, and considering the pixel size sufficiently small compared to the size of the area of study, the first objective became choice of the optimal pixel size. To estimate this optimal pixel size from the data itself [306], the following experiment was designed:

1. A collection 500 pixel grids were produced, with pixel resolutions from 1 m to 500 m, in 1 m increments.
2. For each of the grids, the number of events (coordinate pairs corresponding to a picture or tweet) were counted per pixel.
3. The number of pixels totally or partially covered by the area of study was also counted for each of the pixel grids (effective pixels).
4. The ratio between the number of pixels with at least one event divided by the corresponding effective pixels was computed for each resolution.

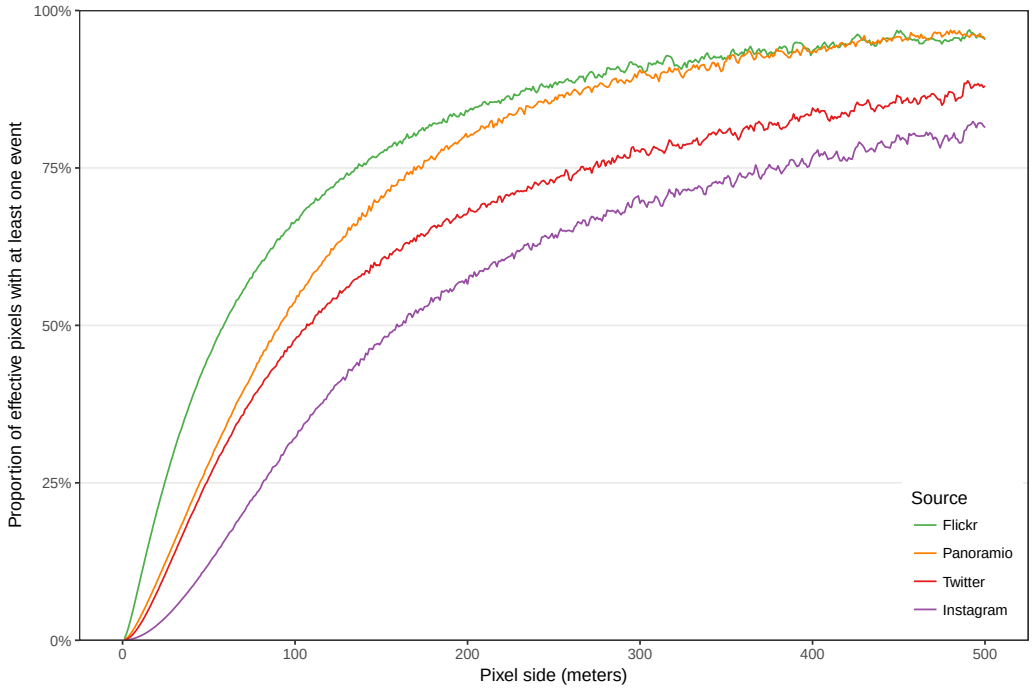
This procedure was repeated for each of the four sources (Fig. 6.2), to empirically determine the optimal pixel size in two geographic scopes:

- The complete area of study (Fig. 6.2a).
- An area with a high concentration of points arbitrarily defined as a 1x1 km square with its sides facing the four cardinal directions, centered on the intersection[7] between Gran Via de les Corts Catalanes and Passeig de Gràcia (Fig. 6.2b).

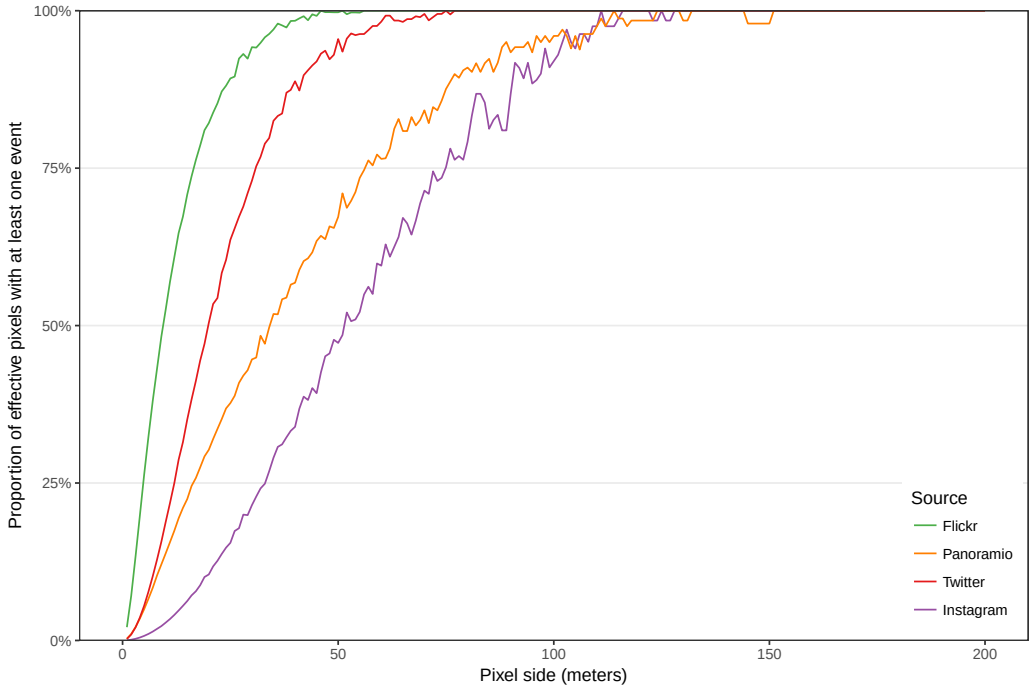
From the results of these experiments (Table 6.2), the selection criterion was defined as the size threshold where the number of effective pixels within the area of study with at least one event became more abundant than the number of pixels without any event. Using this criterion, the number of effective pixels —within the city limits— in the resulting maps that contained at least one event were at least equal than the pixels where no events were recorded (Fig. 6.3).

Finally, using the results of the experiment as a guide, a pixel size of 25 m was arbitrarily chosen as a compromise value to generate maps comparable across sources (Figures 6.4, 6.5 and 6.6), sharing the same pixel size but ensuring that the resulting grids contained pixels with enough data to preserve sufficient detail (Table 6.3).

[7] At the coordinates 41.3894258N, 2.1681192E (EPSG:4326).



(a) The majority of pixels have at least one event with a pixel size of around 100 m, considering the whole area of study.



(b) Focusing on an area with a high concentration of points, the majority of pixels have at least one event with a pixel size under 50 m.

Figure 6.2: Fraction effective pixels recording at least one event as a function of the pixel size for the four sources analyzed. Values were determined empirically for sizes between 1 and 500 meters in steps of one meter, but the bottom figure shows a narrower range.

Table 6.2: Grid sizes where the number of pixels recording at least one event become more abundant that the pixels without any events, within the set of all the pixels necessary to cover the corresponding observation window entirely. The values are obtained using linear interpolation from the empirical functions in figure 6.2.

Source	Full scope	High concentration
Flickr	58.62 m	9.40 m
Panoramio	91.12 m	34.29 m
Twitter	107.95 m	19.87 m
Instagram	161.57 m	52.37 m

Table 6.3: Ratio between the number of pixels with at least one event and the number of pixels necessary to cover the corresponding observation window entirely, for the compromise resolution of a 25x25 m pixel grid.

Source	Full scope	High concentration
Flickr	25.5 %	89.2 %
Panoramio	12.5 %	37.7 %
Twitter	10.6 %	63.6 %
Instagram	3.6 %	15.5 %

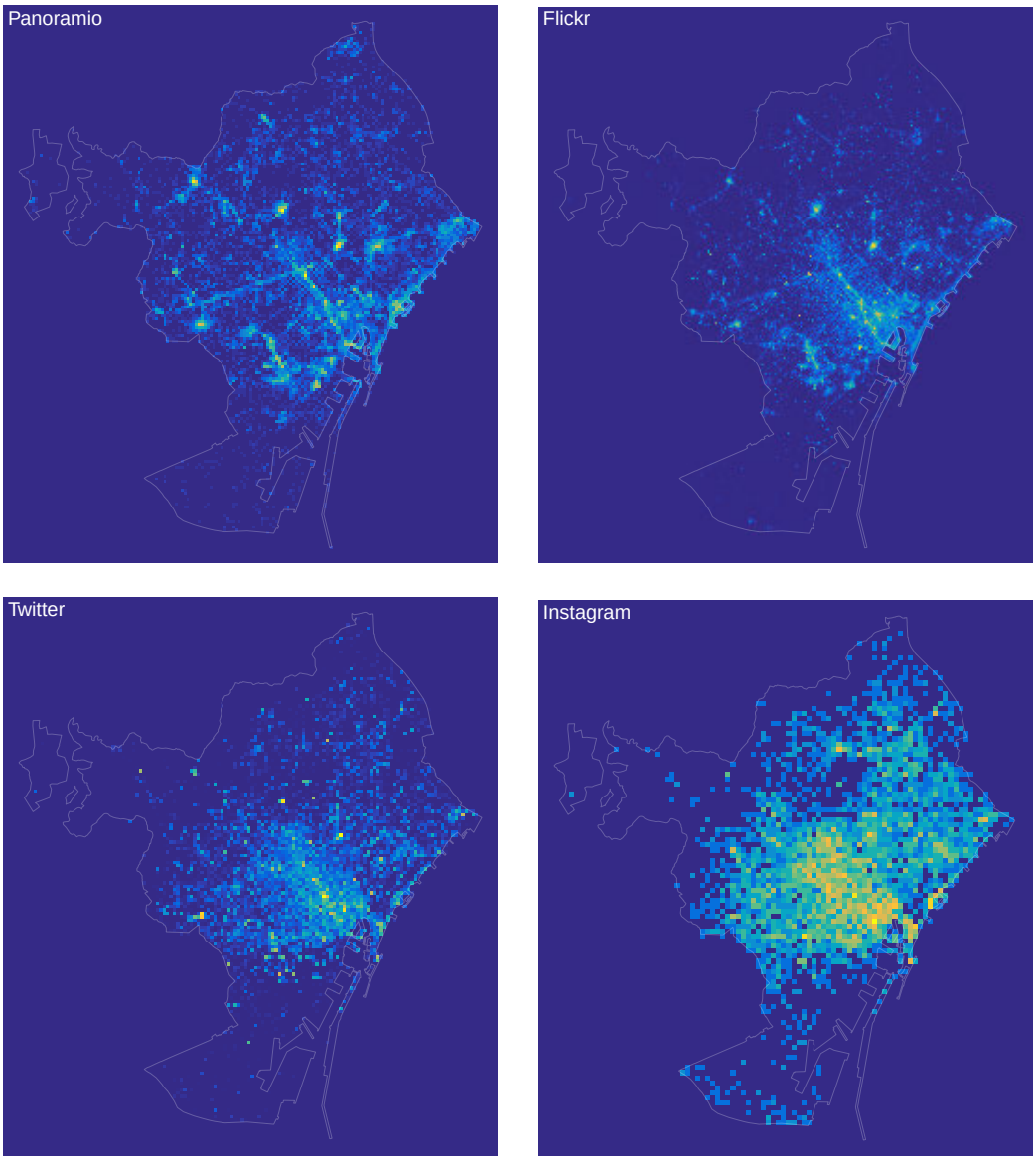


Figure 6.3: Maps of the retrieved sources, binned in a grid of square pixels with sizes determined as the threshold where the number of pixels with at least one recorded event exceeds or matches the number of pixels with none, within the city limits. Sizes detailed in table 6.2. City limits cartography from CartoBCN, under CC BY 3.0.

Table 6.4: Untransformed and transformed (log1p) statistics of the distribution of the number of events per pixel (maximum value, mean, standard deviation and skewness) of the four retrieved sources.

Source	Trans.	Max	Mean	Std. Dev.	Skewness	Kurtosis
Panoramio	none	577	0.17	2.77	53.04 >> 0	4,297 >> 3
	log1p	6.36	0.05	0.27	4.29 > 0	27 > 3
Flickr	none	18,827	1.85	58.69	104.37 >> 0	14,962 >> 3
	log1p	9.84	0.16	0.61	2.80 > 0	12 > 3
Twitter	none	10,569	0.80	30.74	112.83 >> 0	17,935 >> 3
	log1p	9.27	0.07	0.41	4.64 > 0	29 > 3
Instagram	none	4	0.01	0.13	6.28 > 0	51 > 3
	log1p	1.61	0.01	0.09	5.34 > 0	32 > 3

6.2.3 Transformation

Once determined the optimal pixel size from the number and spatial distribution of the collected data, the points were binned into the pixels of the grid, resulting in an integer array of spatial pixels. However, the distribution of the values of these pixels was very skewed (Table 6.4), because the spatial distribution of the points was heavily clustered in some locations an sparse in others.

However, if the results were transformed using the natural logarithm applied to each of the values plus one —using the *log1p* function, which is accurate for values much smaller than one—, the resulting distribution was much less skewed, despite still being very different from a normal distribution (Table 6.4).

6.2.4 Issues with Count Data

After rasterization, count data —which was effectively a density, considering that all pixels were the same size— was dependent on factors that were difficult to control to be useful for urban analysis across different cities, time spans or data sources.

Because the analyzed sources consisted in data sets of very different sizes (Table 3.5), the comparison between the different results was challenging, as sources with a higher number of samples were much more likely to contain pixels with a large number of events and *vice-versa*.

Furthermore, grids of smaller pixels (higher resolution) were more likely to produce pixels with few event counts, while discretization into larger pixels

(lower resolution) was conversely less likely to produce pixels with event counts near zero (Fig. 6.2).

In addition, the event counts per pixel could be misleading, especially when comparing different sources; even when dealing with a single source, as data is accumulating continuously, any analysis involving time should be considered as a sliding window, and therefore the size of the sample should not be considered fixed.

6.2.5 Relative Attractiveness

The issues with count data —even after transformation— needed to be addressed to minimize the influence of the factors discussed above. The proposed solution was to transform the simple counts obtained after the rasterization process, to produce a synthetic measure of "attractiveness" for each source.

To interpret this attractiveness value, we could imagine an initial situation where all the pixels inside the analysis boundary (observation window) contain a uniform quantity —corresponding for example to a single grain of sand, or iron filings—. Then, an imaginary force would act on these units, redistributing them without adding or destroying any grain (wind) or filings (magnetic field).

Therefore, measuring the values after the action of the force would provide a measure of the attractiveness of the pixels that had collected a larger amount of material, compared to the initial uniform condition[8]. Using this analogy, to compute the corresponding attractiveness value mathematically for all pixels in the grid (Table 6.5) the following steps were necessary:

1. The values for each pixel were divided by the sum of all values of all pixels, effectively obtaining the proportion of the total number of events per pixel, normalized between zero and one.
2. These proportions were multiplied by the total number of effective pixels (pixels inside the city limits), obtaining the attractiveness as defined earlier.

6.2.6 Mapping Values to a Color Scale

As the final step to convey the resulting values (the computed transformed event counts) stored in each pixel to the reader, their numeric quantities had to be transformed into a color scale. However, the appearance of color depends on multiple factors such as the output device (e.g. computer monitor, printed paper, black and white photocopy), the type of graphic (e.g. filled polygons, lines, points) or the reader (e.g. different forms of colorblindness, cultural differences).

[8]Which would be the null hypothesis in the case of null hypothesis significance testing.

Table 6.5: Untransformed and transformed (log1p) statistics of the distribution of the "attractiveness" per pixel (maximum value, mean, standard deviation and skewness) of the four retrieved sources.

Source	Trans.	Max	Mean	Std. Dev.	Skewness	Kurtosis
Panoramio	none	1,254	0.36	6.02	53.04 >> 0	4,297 >> 3
	log1p	7.13	0.08	0.38	3.50 > 0	17 > 3
Flickr	none	37,024	0.36	11.54	104.37 >> 0	14,962 >> 3
	log1p	10.52	0.07	0.34	4.37 > 0	28 > 3
Twitter	none	4,823	0.36	14.03	112.83 >> 0	17,935 >> 3
	log1p	8.48	0.05	0.31	5.52 > 0	41 > 3
Instagram	none	98	0.36	0.31	6.28 > 0	51 > 3
	log1p	4.60	0.04	0.38	5.00 > 0	26 > 3

To convert these values into colors for visualization, the R package pals 1.4 [307] was used. The package conveniently includes a comprehensive catalog of selected color palettes suitable for any of the three most common visualization situations:

- Categorical data, where hues must be distinguishable from each other.
- Color ramps for continuous quantitative data (either single or multi hue).
- Diverging color ramps for variables with a "neutral" center between two extremes (e.g. positive/negative, high/low, good/bad).

In the case of the raster maps described in this section, the values to be represented belonged to the second category as as they consisted in magnitudes without negative values. These corresponding color ramps should be preferably perceptually uniform, avoiding Mach banding across their entire range (particularly in their luminance), while being distinguishable by a color-blind person (deuteranomaly, protanomaly or tritanomaly) or when displayed in grayscale (e.g. photocopied) unambiguously.

The considered color ramps were the default in the matplotlib Python library named viridis [308], the default palette in the newest versions of the Matlab software named parula[9], the classic palettes[10] by Harrower and Brewer [289], the cubehelix palette developed for astronomy [309], a recent palette collection developed for oceanography visualization [310], and the palettes developed by Kovesi [311], and Tol [312].

After a series of tests, the final two contenders were viridis and parula, because

[9]Parula is discussed in the colormap section of the MatLab documentation available at *http://www.mathworks.com/help/matlab/ref/colormap.html* at the time of writing.

[10]The interactive ColorBrewer website is available at *http://colorbrewer2.org/* at the time of writing.

of their more efficient use of the available spectrum. Finally, parula was considered better overall against the four default variations offered by viridis[11], and subjectively more pleasing to the eye[12].

6.2.7 Results

The results showed a very detailed spatial distribution of events because of the chosen 25 m pixel size (discussed in section 6.2.2), which was however too small for the areas with a low occurrence of events. The proposed approach showed some advantages when focusing (zooming in and therefore enlarging) a smaller area of the map:

Adaptive pixel size as the pixel size on screen was not determined by the zoom level but by the spatial distribution —clustering or dispersion— of events in the observation window, and the pixel size changed when focusing in a region with a high or low concentration of events.

Dynamic range was not wasted because the color scale could adapt to the specific range of values displayed on the screen, stretching the color ramp to match the distribution in the area of interest, avoiding underutilized dynamic range as well as clipping artifacts[13].

Transformation as the defined attractiveness measure allowed the comparison of very different sources, with very different distributions and population sizes, showing exceptional robustness across the researched services, without being sensitive to the choice of pixel size.

These advantages can be observed when comparing the resulting maps of the spatial distribution of the geotagged pictures collected from Panoramio (Fig. 6.4) or Flickr (Fig. 6.5), and the geolocated messages retrieved from Twitter (Fig. 6.6) during a one-year period.

The two sources using picture-based data (Panoramio and Flickr) provided much more detail than Twitter, allowing the identification of many individual streets. Of these two data sets, Panoramio data appeared "sharper" because of the focus of the service on pictures of landmarks, but the Flickr data set was and order of magnitude larger, included higher diversity of picture themes and had the advantage of providing richer metadata that allowed researching other aspects

[11] Viridis offers four color ramps, two with black as their darkest color ("magma" and "inferno") and another two ("plasma" and "viridis") starting with a brighter hue.

[12] The "inferno" palette was also considered, but only after discarding the bottom 10 % range because it was difficult to distinguish the detail in the darkest (almost black) colors, an issue that did not affect parula.

[13] Following metaphor explained earlier in section 6.2.1, avoiding underexposure and overexposure.

such as the origins of visitors (chapter 5) and temporal patterns (chapter 9) beyond picture locations.

Despite the maps of Panoramio and Flickr locations showing a much more focused map of the city than their Twitter counterpart —which provided a coarser accuracy in the location of its users—, the later was probably more representative of the behavior of the inhabitants because it reflected a broader range of activities beyond picture taking. Another advantage of the Twitter dataset was that it allowed matching semantic and temporal content to the location data, and some findings using this approach have been recently published [143].

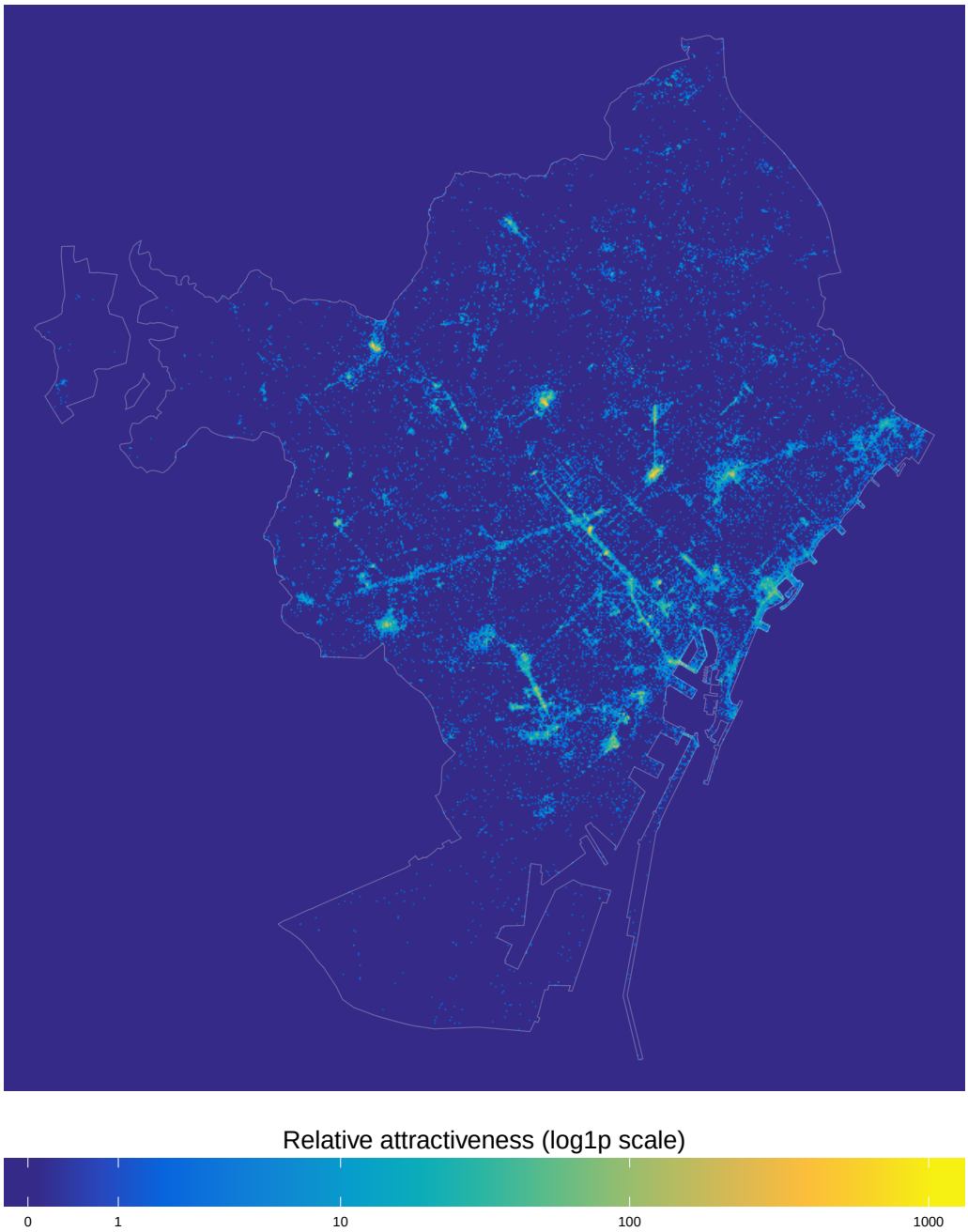


Figure 6.4: Spatial distribution of the attractiveness of urban spaces using the geotagged pictures of Barcelona collected from Panoramio, counted in a grid of 25x25 meter cells. Color scale is perceptually uniform and magnitudes are log-transformed. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

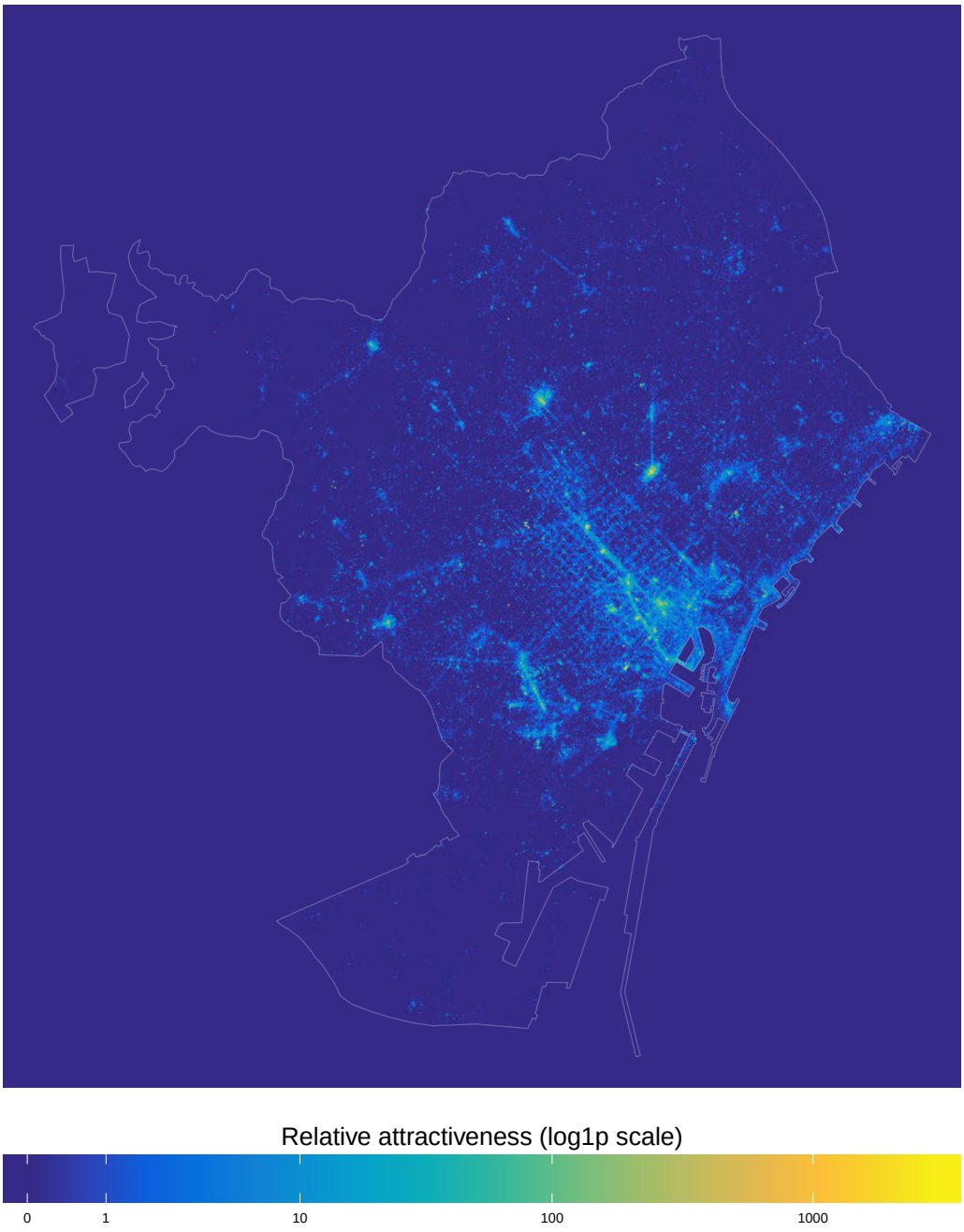


Figure 6.5: Spatial distribution of the attractiveness of urban spaces using the geotagged pictures of Barcelona collected from Flickr, counted in a grid of 25x25 meter cells. Color scale is perceptually uniform and magnitudes are log-transformed. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

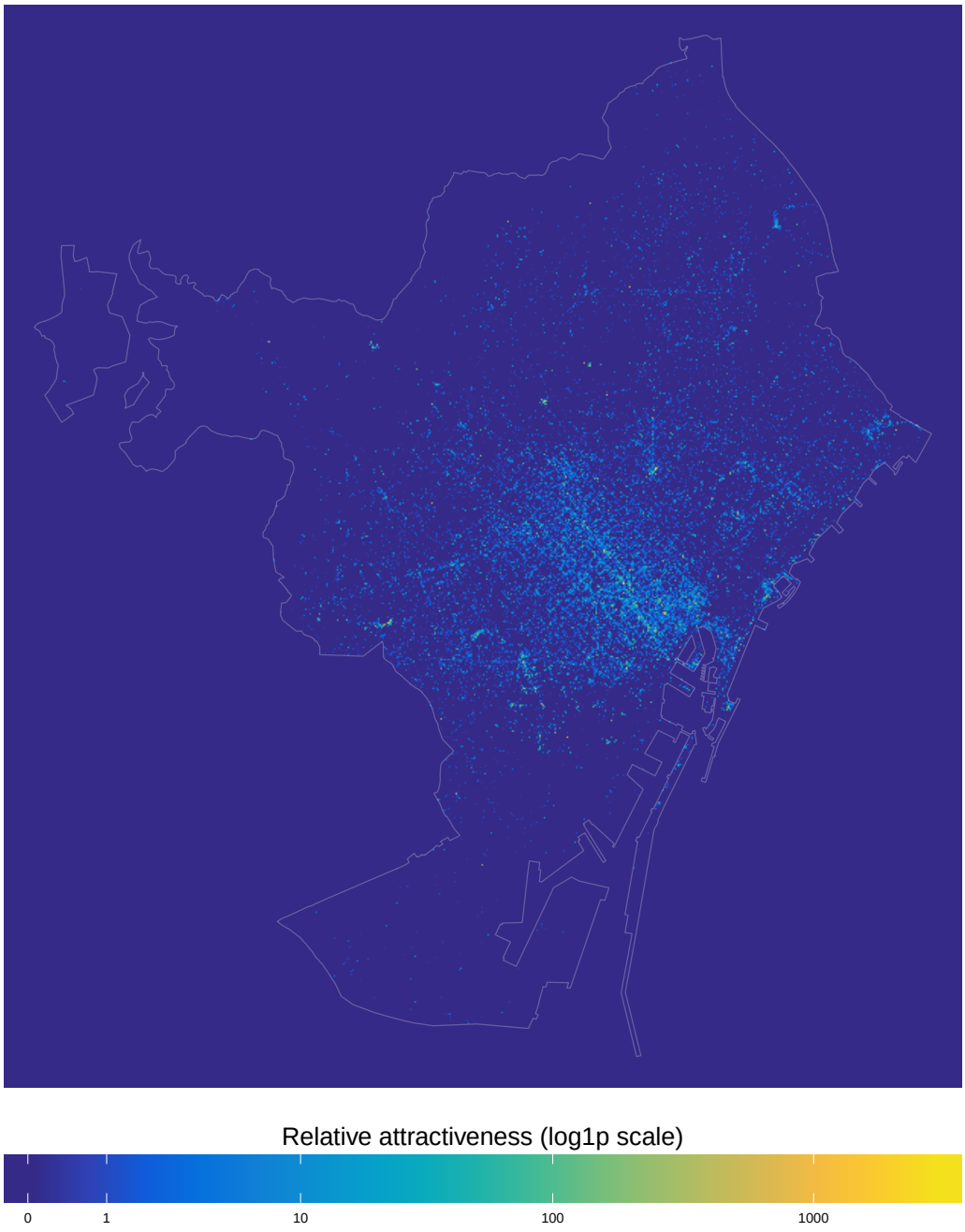


Figure 6.6: Spatial distribution of the attractiveness of urban spaces using the geotagged status messages collected from Twitter, counted in a grid of 25x25 meter cells. Color scale is perceptually uniform and magnitudes are log-transformed. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

6.3 Local Measures of Spatial Autocorrelation

6.3.1 Limitations of the Rasterization Approach

The rasterization process described in section 6.2 resulted very successful in visualizing the areas with a high concentration of events (bright spots) and areas with near-zero event counts (darker areas), providing a faithful picture of the spatial distribution of events.

However, despite the transformation of event counts into an attractiveness measure —which homogenized the results across services—, and the aggregation of the individual points into pixels —avoiding overplotting issues[14]—, the resulting maps were still sensitive to the features of the studied data set (e.g. spatial distribution, number of events) and the representation choices (e.g. choice of color scale, transformation).

Therefore, with the objective of producing more informative maps and reducing the arbitrariness in the subjective choice of the rendering scheme, it was necessary a complementary technique to the density-based rasterization methodology, using a more principled approach based on spatial statistics (local autocorrelation), backed by the robust framework of statistical significance testing.

6.3.2 Local Indicators of Spatial Association

Tobler's first law of geography [313] states that "everything is related to everything else, but near things are more related than distant things". This statement, based on the concept of the "friction of distance", is the foundation of the concepts of spatial dependence, spatial autocorrelation and spatial association.

Spatial autocorrelation (SA) measures the degree of clustering or similarity (positive SA) or dispersion or dissimilarity (negative SA) of the distribution of spatial features and/or their associated values, compared to what would be expected in a pattern produced by a random process, which would not exhibit neither positive nor negative SA.

In the context of an area of study, global autocorrelation analyses whether the features exhibit clustering or dispersion considering the entire map pattern, while local autocorrelation focuses on the identification of clusters (hot or cold spots) within the area of study. The present research focuses on two of the most widely used Local Indicators of Spatial Association (LISAs):

- Local Moran's I (discussed in section 6.3.7).
- Local Getis and Ord's G and G* (discussed in section 6.3.8).

[14]Specially biases introduced by the size of the dots or the use of transparency.

- Geary's C[15] [314], recently extended to a multivariate context [315], was not pursued because it did not provide information on the *type* of association.

Because the development platform was based on the R programming language, some algorithms in the state-of-the-art Python Spatial Analysis Library[16] (PySAL) [316, 317, 318] were adapted and partially re-implemented using the R package spdep 0.7-4 [319, 320]. However, in later stages the R package reticulate 1.7 [321] was used to access the PySAL library within the R environment. This approach allowed a tighter integration of the analysis pipeline, affording greater flexibility to programmatically tweak some of the parameters when necessary, an option not provided by current GIS packages such as ArcGIS or QGIS.

The maps were produced with the tmap 1.11 R package [296] using the color palettes provided by the RColorBrewer[17] 1.1-2 package [322].

6.3.3 Regular Tiling

The first step for the calculation of the LISAs was the tessellation of the area of study into a set of contiguous polygons, to obtain the event counts per tile (quadrat counts). This step was also performed implicitly during the rasterization process, where each pixel corresponded to a small square-shaped polygon.

There are only three possible regular tessellations of the two-dimensional plane —denominated Euclidean tilings— on which to base the geometry of the regular polygons used for event aggregation:

- Triangles (deltille)
- Squares (quadrille)
- Hexagons (hextille)

While both the square and hexagonal tiling can be defined as the Voronoi (or Dirichlet) tessellation of a plane where the points are arranged in a regular lattice, the hexagonal tiles enclose a larger area using the same perimeter, as they are closer in shape to a circle.

In this case, the size of the polygons (2 hectares, discussed in section 6.3.4) was considerably larger than the pixel size discussed in section 6.2 (32 times larger) and therefore the shape and orientation of the polygons could have an impact on the results of the analysis and had to be taken into account. The square tiling was discarded because it favored the two orthogonal orientations corresponding to either sides, while triangles and hexagons favored three orientations instead.

[15] Geary's C is also known as Geary's contiguity ratio.
[16] The PySAL project website is available at *http://pysal.org/* at the time of writing.
[17] The ColorBrewer palettes are included in QGIS and recently in ArcGIS Pro also.

In addition, the type of the neighborhood (discussed in section 6.3.5) of a square tile through its contiguity —excluding itself— depends heavily on the definition of contiguity: edges only (four neighbors) or edges and vertices (eight neighbors). Similarly, the triangular tiling had the same issues with the definition of its neighborhood, which the chosen hexagonal tiling did not have, as all neighbors of an hexagon shared only edges and never single vertices.

6.3.4 Optimal Tile Size Selection

Unlike collections of numbers, which can be summarized into a single value (e.g. mean, standard deviation), the complexity of planar point patterns requires a different approach. Beyond purely graphical methods such as the Fry plots [323] (originally developed by Patterson [324]), the Morisita index plots [325] or the Stienen diagrams, the majority of these summaries take the form of a *function* of some metric of the point pattern over a radius [326], generally plotted against the expectation of a Poisson process of the same intensity under the null hypothesis of Complete Spatial Randomness (CSR):

K(r) function is the cumulative average number of points within a distance r of a typical point. In order to make the comparison of point pattern possible, the estimation is corrected for edge effects (to accommodate different observation windows) and divided by the intensity (to compare patterns with different number of points).

L(r) function is a common transformation of the K(r) function which transforms the theoretical reference Poisson K-function into a straight line. This transformation is sometimes used instead of the true K function, as in the case of ArcGIS where this transformation is used by default.

g(r) function is the "pair correlation function" which is similar to the K function but considers only the interpoint distance *equal* to r instead of *equal or less* than r, and therefore is not a cumulative function. Its calculation is computationally intensive but can be estimated from the K function using smoothing splines to approximate its derivative.

F(r) function is the "empty space" function, sometimes called "spherical first contact distribution" or the "point-to-nearest-event distribution". It provides the probabilities (in the 0 to 1 range) of finding a point within a radius r of any fixed reference location.

G(r) function is the "nearest-neighbor distance distribution" function, and provides the cumulative distribution function of the nearest neighbor distance at a typical point. Like the F function provides a probability in the 0 to 1 range.

J(r) function compares the 1-F(r) and 1-G(r) functions as a dimensionless ratio.

> For J(r) > 1 there is evidence of a regular pattern and for J(r) < 1 evidence of clustering, at scales less than or equal to r. The empirical J-function is insensitive to edge effects because the effects of the F and G functions cancel out.

Three of these functions were estimated with the spatstat 1.54-0 R package [327] for each of the four analyzed sources of point locations and plotted in separate panels corresponding to each of the three functions (Fig. 6.7). The computed functions for each source were assigned a color and plotted at the same scale; however, their corresponding expected reference functions under CSR were not plotted because their definition depends on the intensity of the point process to be compared against, which was different across the data sets.

The K function shows that the point pattern corresponding to Flickr —followed by Twitter— is more clustered than the Panoramio, although the later is more clustered than Twitter in the range of distances between 75 and 325 meters. The Instagram point pattern is the less clustered until around 1 km, where becomes more clustered than Panoramio, probably because Panoramio data focuses on landmarks while Instagram records business locations, and there are many areas devoid of landmarks but with a significant presence of businesses.

The G function shows a similar pattern, with the Twitter dataset exhibiting very short distances with small radii, suggesting that many points overlap because of truncation or approximation artifacts, followed by Flickr data, which suggest the same problem but less pronounced. The Panoramio G function suggests a dataset with more accurate locations (the differences and similarities of the collected locations according to their source are discussed in chapter 3). Finally, the Instagram point pattern shows a behavior much closer to a random process than the rest of the sources, with a much larger median of the distribution of nearest-neighbor distances of about 30 meters (but far than the 60 meters of its expected median under CSR).

The F function (with results very similar than the obtained experimentally in section 6.2.2) was useful to determine the optimal size of the tiles from the radii —and their corresponding circle areas— that satisfied a probability threshold. Two criteria that were deemed reasonable were tested (Table 6.6):

1. Distances where the probability of finding some point was equal or larger than finding none (p = 1/2).
2. Distance where the probability of finding some point was twice as large or higher than finding none (p = 2/3)

Using the previous results as a guide to determine the optimal size, four candidates were selected with sizes of one to four hectares in increments of one hectare, along with their corresponding radii. These radii were used to estimate for each source

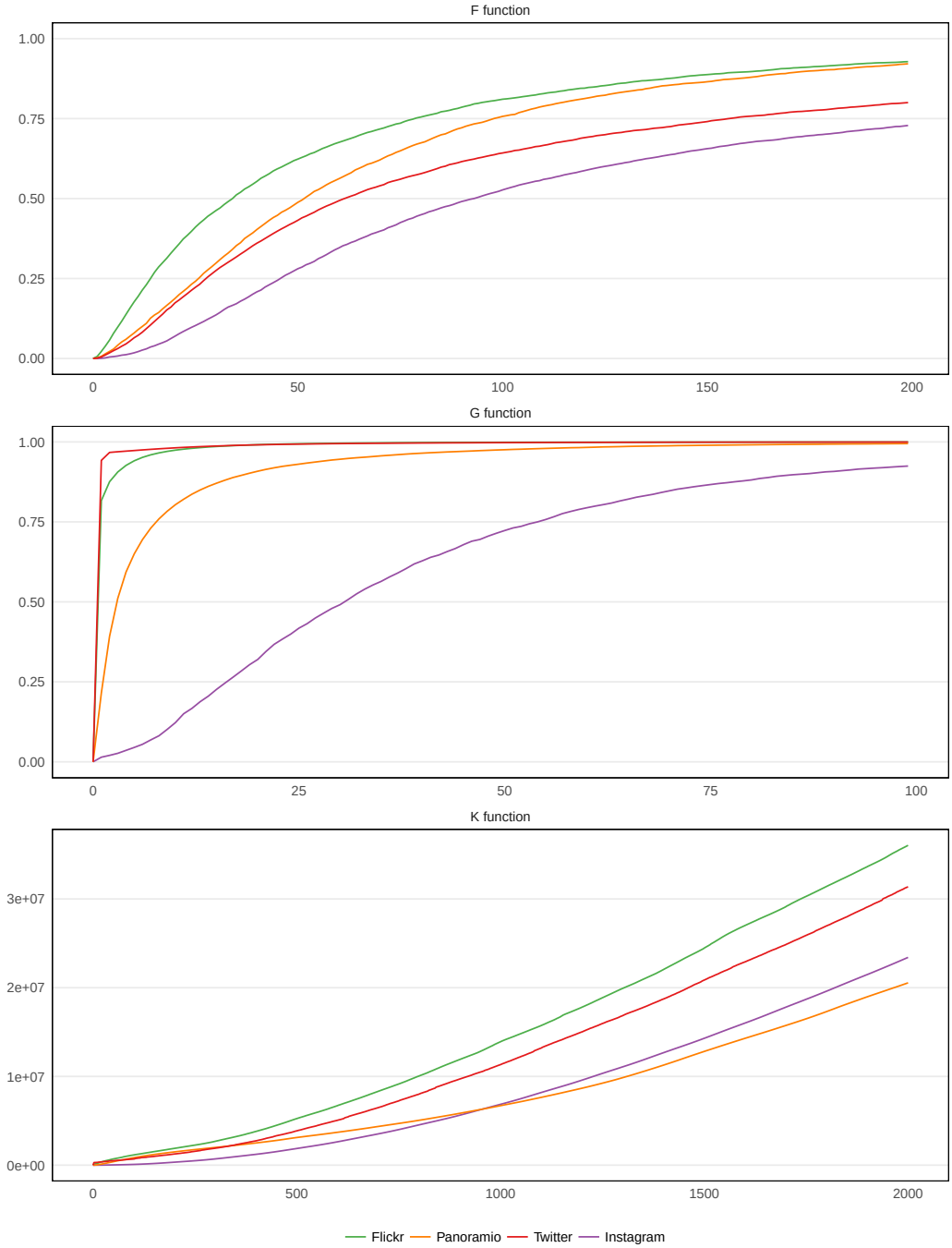


Figure 6.7: Estimation of the F(r), G(r) and K(r) functions corresponding to the point patterns of the four studied sources. Radii are on the x axis (with different ranges). In the y axis, values of the F and G functions are probabilities while values of the K function are standardized picture counts.

Table 6.6: Radius and area of its corresponding circle where the probability of finding a point is 1/2 or 2/3, for all retrieved sources. Values derived from the estimate of the empty space F(r) function using linear interpolation between computed values in one meter intervals.

	$p(\exists) \geq p(\nexists)$		$p(\exists) \geq 2p(\nexists)$	
	Radius (m)	Area (ha)	Radius (m)	Area (ha)
Panoramio	51.53	0.83	78.34	1.93
Flickr	33.97	0.36	58.05	1.06
Twitter	61.17	1.18	109.98	3.80
Instagram	92.97	2.72	155.35	7.58

Table 6.7: Four candidate tile sizes and the corresponding probabilities of finding a point within a circle of the same area. Values derived from the estimate of the empty space F(r) function using linear interpolation between computed values in one meter intervals.

Area	Circle Radius	Panoramio	Flickr	Twitter	Instagram
1 ha	56.42 m	54 %	66 %	47 %	32 %
2 ha	79.79 m	67 %	75 %	58 %	45 %
3 ha	97.72 m	74 %	80 %	64 %	52 %
4 ha	112.84 m	80 %	83 %	67 %	57 %

the probability of finding at least one point from the computed the F function using linear interpolation (Table 6.7).

From this table, a consensus value of two hectares was selected, as it provided the desired minimum requirement of 50 % probability of finding at least one point for all sources —except Instagram which was slightly below with 45 %—, and estimated a probability of over 2/3 for the most accurate sources (Panoramio and Flickr), which translated to having at least twice the probability of finding at least one point than finding none.

6.3.5 Conceptualization of Spatial Relationships

Once the shape of the tiles of the regular tessellation (hexagons) and the size of the analysis unit (two hectares) were chosen, is was necessary to formalize the concept of neighborhood between tiles, using either of the two conventional approaches [317] widely used to define the spatial relationship between features

in a collection:

Distance-Based criteria define the neighborhood relation as a function of the Euclidean distance between the features, up to a maximum distance threshold and/or a maximum number of nearest neighbors.

Contiguity-Based neighborhood considers whether the features share a common border, vertex or either, resulting in the corresponding rook (Von Neumann neighborhood, with a Manhattan distance of 1), bishop (seldom used) and queen (Moore neighborhood, with a Chebyshev distance of 1) contiguities.

Distance-based relationships are suitable for point features whereas contiguity-based relationships are appropriate for polygon features. However, point features can be used to construct Thiessen polygons and polygon features can be collapsed into their centroids, and therefore this distinction is not always applicable.

The neighborhood relationships are stored in a spatial weights matrix, which for spatial data is generally a binary a sparse matrix that captures the structure of a network where features are nodes and links correspond to the presence of a neighbor relation. This matrix was stored in the GAL file format, introduced in the 1980s by the Geometric Algorithms Lab at Nottingham University and included in the 1990s in the SpaceStat[18] software package [328].

The R package spdep [329] was used for the computation of the spatial weights matrix from the geometries and their topological relationships, using the rook contiguity (although for the case of hexagons all three possible contiguity criteria produce the same results). The computation produced a matrix with 63,198 nonzero links for 10,862 regions. The average number of links per feature was 5.82 (instead of 6, because of edge effects) and the sparsity of the matrix was 99.95 %.

For each feature, the spatial weights were standardized dividing the binary weights by the total number of neighbors (row standardization), and therefore all row-standardized weights summed one for all hexagonal tiles.

6.3.6 Statistical Significance Testing

The determination of the hot and cold spots using local spatial statistics (Moran's I and Getis-Ord's G*) allows the production of maps where the representation is backed by a robust theoretical framework in contrast with the standard choropleth maps approach, which is subject to biases (intentional or not) in the classification and representation choices.

[18] SpaceStat is developed by BioMedware, whose website is available at *http://www.biomedware.com/* at the time of writing.

The computation of both statistics for the event counts lying inside each hexagonal tile used two functions included in the R package spdep, which returned the corresponding z-value for each of the polygons (*localmoran* in the case of Moran's I and *localG* in the case of Getis-Ord's G*).

Since a z-score is the difference between the observed and expected mean, divided by the standard deviation, it could be used for statistical inference converting the z-scores to p-values, but this approach violated the assumption of normality (Table 6.4), even after transforming the values.

These distributions are commonly found in urban studies, and render many common statistical inference techniques unsuitable. To tackle this problem, the approach used was the computation of *pseudo* p-values instead of the usual *analytical* p-values, inspired by the algorithms implemented in the open-source PySAL Python library.

Unfortunately, the spdep package did not implement (at the time of writing) the required functionality, which had to be developed independently by the author. This approach however, allowed tweaking some parameters that would otherwise depend on external implementation choices, without leaving the R environment.

To compute the pseudo p-values for the local Moran's I and the local Getis-Ord G*, these statistics were initially computed with the event counts corresponding to each hexagon, with the neighborhood defined in section 6.3.5. Next, the following procedure was repeated a large (n) number of times (999 permutations in the case of the maps shown) to obtain a reference empirical distribution of the statistics under the null hypothesis of spatial randomness:

1. The event counts were randomly reassigned to the hexagons.
2. The statistics were computed with the shuffled values.
3. For each of the hexagons, the results of each permutation were compared to the original result to check if the simulated value was more extreme (one-sided test).

For each hexagon, the number of times (m) were counted when either the true value was positive and the simulated value was larger or the true value was negative and the simulated value was smaller. The pseudo-significance (p') was computed as:

$$p' = \frac{m+1}{n+1}$$

6.3.7 Cluster and Outlier Detection (Local Moran's I)

The global spatial autocorrelation measures the clustering of values of a variable inside a defined area, as the existence of areas of higher of lower values that

would be expected by chance alone (the null hypotheses is that values exhibit a random distribution in space).

In spatial statistics, the global Moran's I [330] is widely used as a measure of spatial autocorrelation, and characterizes the distribution of values in space in either dispersed (negative values) or clustered (positive values) usually in the range of -1 to 1, but generally transformed into z-scores. Its computation is defined as follows:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Where the items in the formula correspond to:

N is the total number of features.
i, j are the indices of the features.
x is the studied variable.
w is the spatial weights matrix, with zeroes in the diagonal.
W is the sum of all elements of *w*.

The results are highly dependent on the choice of the —generally sparse— spatial weights matrix (*w*), which can be binary and contain only ones and zeroes[19] (e.g. contiguity, k-nearest neighbors) or consist on continuous values, as the result of a distance decay function where weights are assigned according to a defined measure of distance.

While the *global* Moran's I assumes homogeneity and only characterizes the overall dispersion or clustering of the area of study in a single statistic, it is also possible to find local clusters analyzing *local* spatial autocorrelation.

The Local Indicators of Spatial Association (LISA) [331], evaluate clustering of each individual spatial unit (*i*), computing a local Moran's I for each of them, exploiting the fact that Moran's I is the sum of individual cross-products:

$$I_i = \frac{(x_i - \bar{x})}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_j - \bar{x})$$

where the denominator expands to:

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n} (x_j - \bar{x})^2}{N - 1}$$

[19] Or logical, containing *true* and *false* values.

The sign of the computed local Moran's I indicates whether the value of the feature is *similar* to its neighbors (positive local I values) and therefore a belongs to a cluster of high or low values, or conversely if it is an outlier with *dissimilar* values compared to its neighbors (negative local I values).

In contrast with the computation the global Moran's I, which provides a single result for the complete area of study, the interpretation of the local Moran's I of a *feature* depends on three combined factors that must be interpreted together:

- The sign of the standardized value of interest.
- The sign of the standardized spatially lagged variable.
- The statistical significance of the result.

To evaluate whether the values of the variable of interest are higher or lower than the mean, the variable is standardized, with positive values indicating a relatively high value and negative values indicating a relatively low value, in relation to the mean of the distribution.

The same procedure is applied to the spatially lagged variable, with high (positive) standardized values corresponding to high values of the neighboring features, and conversely low (negative) standardized values corresponding to low values in the neighboring features.

Assuming a normal distribution, the statistical significance can be determined analytically from the z-scores; otherwise —as in the case of most urban data— permutation-based pseudo p-values can be computed instead, as discussed in section 6.3.6. Features with pseudo p-values above a threshold (in this research the standard 0.05 was chosen) are deemed statistically non significant and not considered in any category.

The results can be inspected using a Moran scatter plot of the variable of interest (Fig. 6.8), with standardized values in the x axis and the standardized spatially lagged values in the y axis[20]:

High-High (HH) Local spatial cluster of high values (top right quadrant, dark red): high value features surrounded by other high value features in their neighborhood.

Low-Low (LL) Local spatial cluster of low values (bottom left quadrant, dark blue): low value features surrounded by other low value features in their neighborhood.

High-Low (HL) Local high spatial outlier (bottom right quadrant, light red): high value features surrounded by low value features in their neighborhood.

Low-High (LH) Local low spatial outlier (top left quadrant, light blue): low value features surrounded by high value features in their neighborhood.

[20]By convention the values in both axes are standardized and are therefore z-scores.

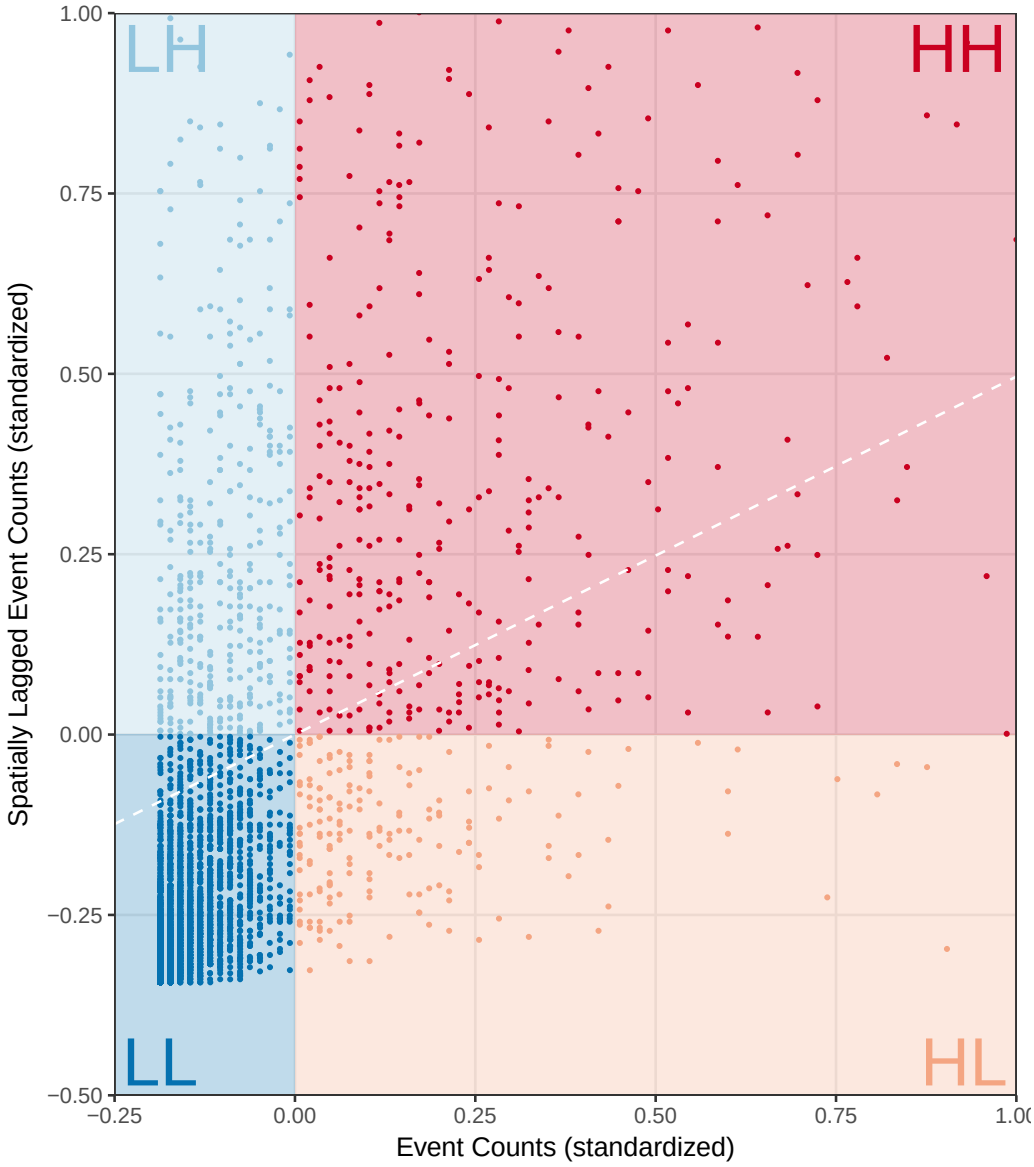


Figure 6.8: Example Moran scatter plot of the event counts retrieved from Panoramio, plotted against the spatially lagged event counts (both variables are standardized). Ranges are limited between -0.25 and 1 standard units in both axes. The four quadrants and points are colored with the same color scheme as figures 6.9, 6.10, 6.11 and 6.12, according to the four categories (HH, LL, HL and LH).

In the resulting maps for the point data retrieved from Panoramio (Fig. 6.9), Flickr (Fig. 6.10), Twitter (Fig. 6.11) and Instagram (Fig. 6.12), and compared to the rasterization technique discussed in section 6.2, the local Moran's I show more clearly the areas where the event counts are more concentrated, and the areas with few events, at the expense of a reduced spatial resolution because of the larger size of the analysis units.

6.3.8 Hot and Cold Spots (Local Getis-Ord's G*)

Local G statistics of local spatial autocorrelation were introduced [332] and developed [333] by Getis and Ord from the logic of point pattern analysis. Their approach to measuring spatial autocorrelation allowed a more general definition of spatial weights, respect the work of Moran [330] and Geary [314].

When applied point patterns, G-statistics for a point feature are defined as the ratio between the number of observations within a distance of a point divided by total point count. When generalized to areal units, they focus on associations among (nonnegative) attributes of a feature and the attributes of features in its neighborhood (defined by the spatial weights matrix).

There are two versions of the G-statistic, which solely differ in whether just the neighboring features are taken into account (Getis-Ord G) or the value of the feature is also included (Getis-Ord G*) when comparing the value of the feature to its spatial context.

The G statistic for an attribute (x) of a feature (i) is defined as the ratio of the weighted average of the values in the neighboring locations of i, to the sum of all values in the area of interest, excluding the value of feature i.

$$G_i = \frac{\sum_{j \neq i} w_{ij} x_j}{\sum_{j \neq i} x_j}$$

Similarly, the G* statistic for feature i is also defined as the same ratio, but including the value of feature i in both the numerator and denominator (which is constant across the calculations for all features).

$$G_i^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}$$

In both cases the neighborhood is defined by the spatial weights matrix (w), which for spatial features is a sparse matrix which consists mainly of zeroes. The difference between both matrices is in the diagonal, which in the case of the G statistic is composed only of zeroes. After the computation of the local Getis-Ord

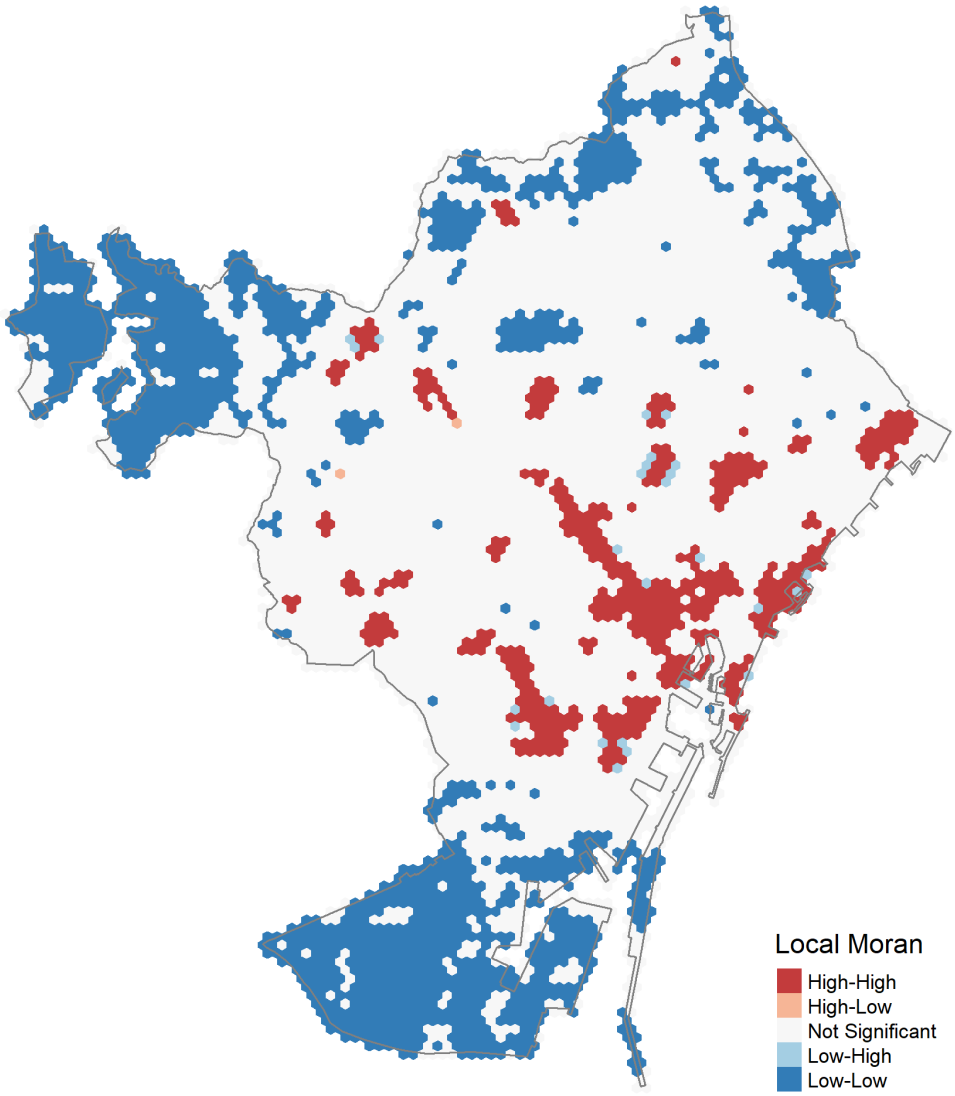


Figure 6.9: Local Moran's I corresponding to the intensity of the geotagged pictures of Barcelona collected from Panoramio, aggregated in a grid of 2-hectare regular hexagons. Map shows local spatial clusters of high intensity (red) and low intensity (blue), as well of high (light blue) and low (light red) spatial outliers. Statistical significance determined using permutation-based pseudo p-values (999 permutations), with a threshold of p' < 0.05. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

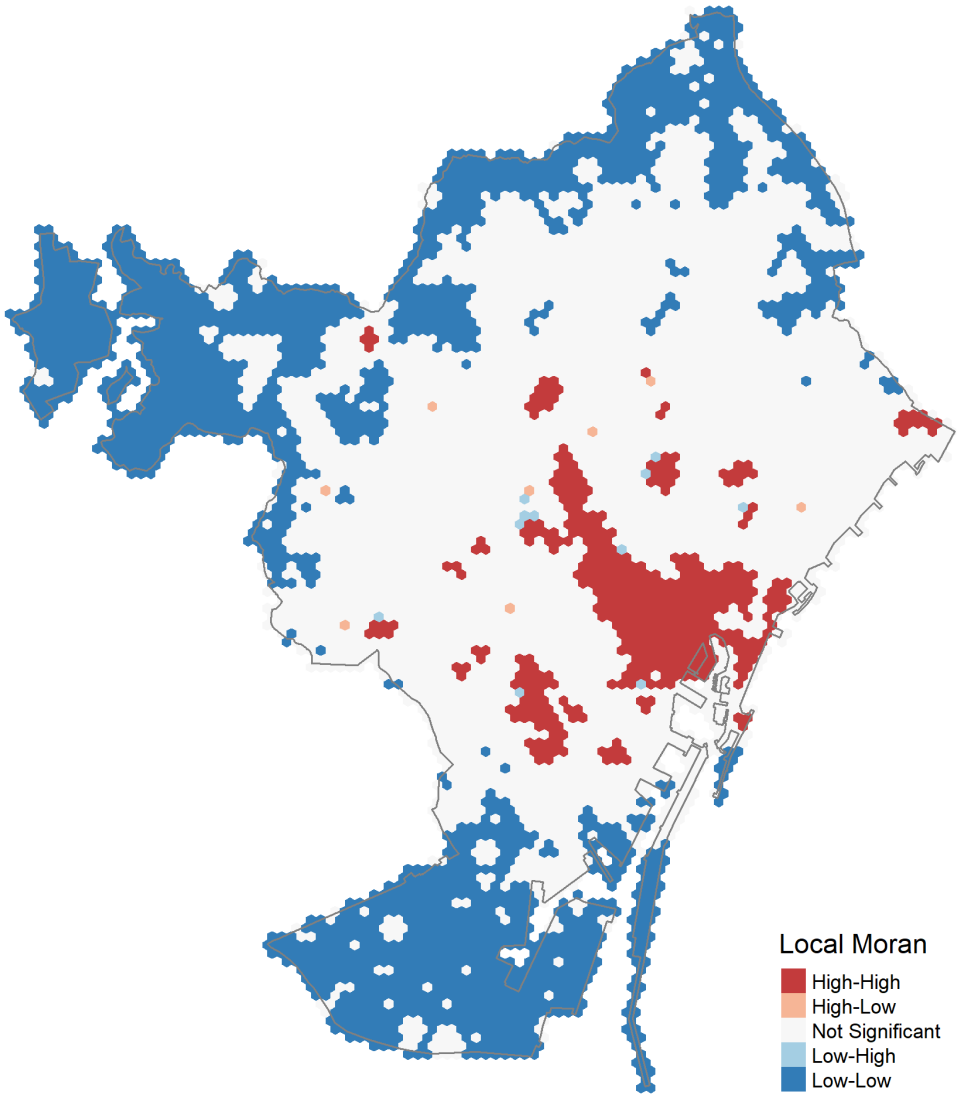


Figure 6.10: Local Moran's I corresponding to the intensity of the geotagged pictures of Barcelona collected from Flickr, aggregated in a grid of 2-hectare regular hexagons. Map shows local spatial clusters of high intensity (red) and low intensity (blue), as well of high (light blue) and low (light red) spatial outliers. Statistical significance determined using permutation-based pseudo p-values (999 permutations), with a threshold of p' < 0.05. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

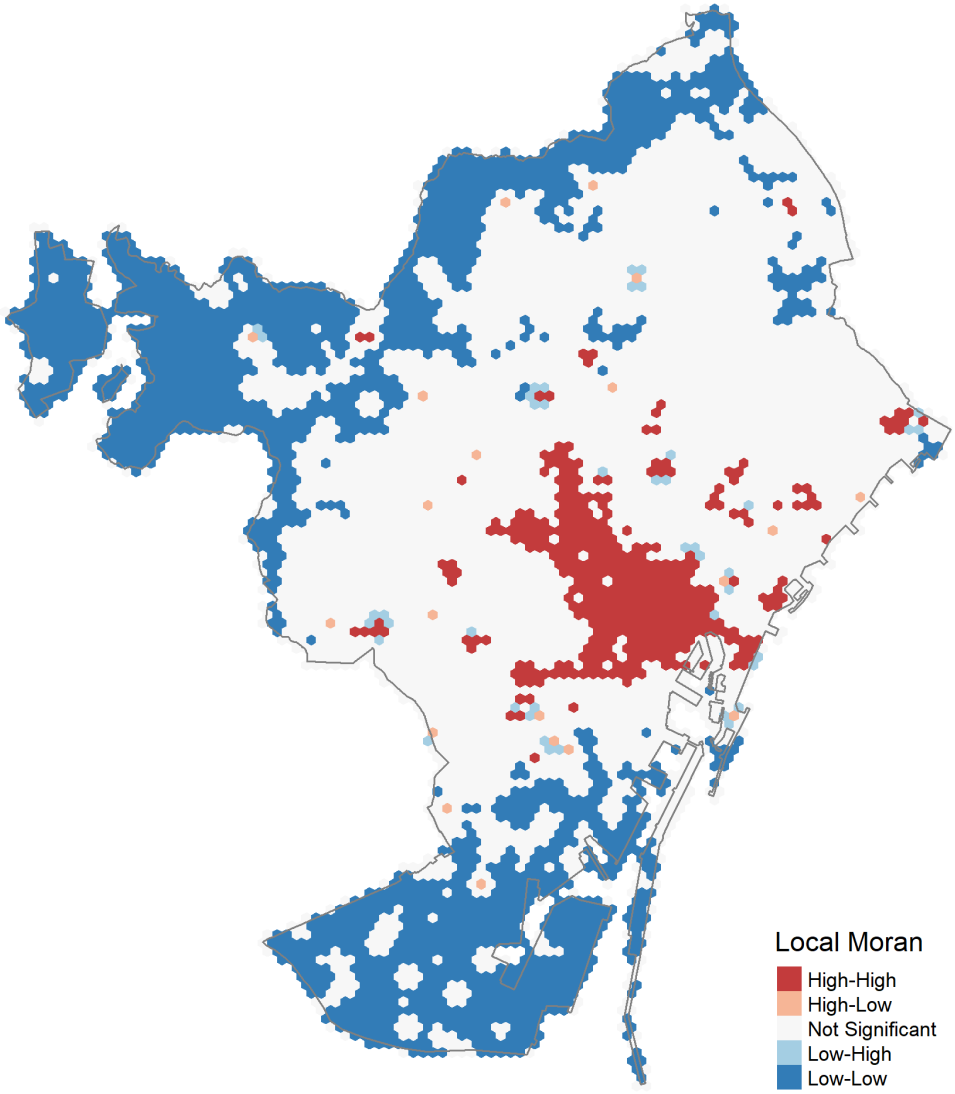


Figure 6.11: Local Moran's I corresponding to the intensity of the geotagged status messages collected from Twitter, aggregated in a grid of 2-hectare regular hexagons. Map shows local spatial clusters of high intensity (red) and low intensity (blue), as well of high (light blue) and low (light red) spatial outliers. Statistical significance determined using permutation-based pseudo p-values (999 permutations), with a threshold of p' < 0.05. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

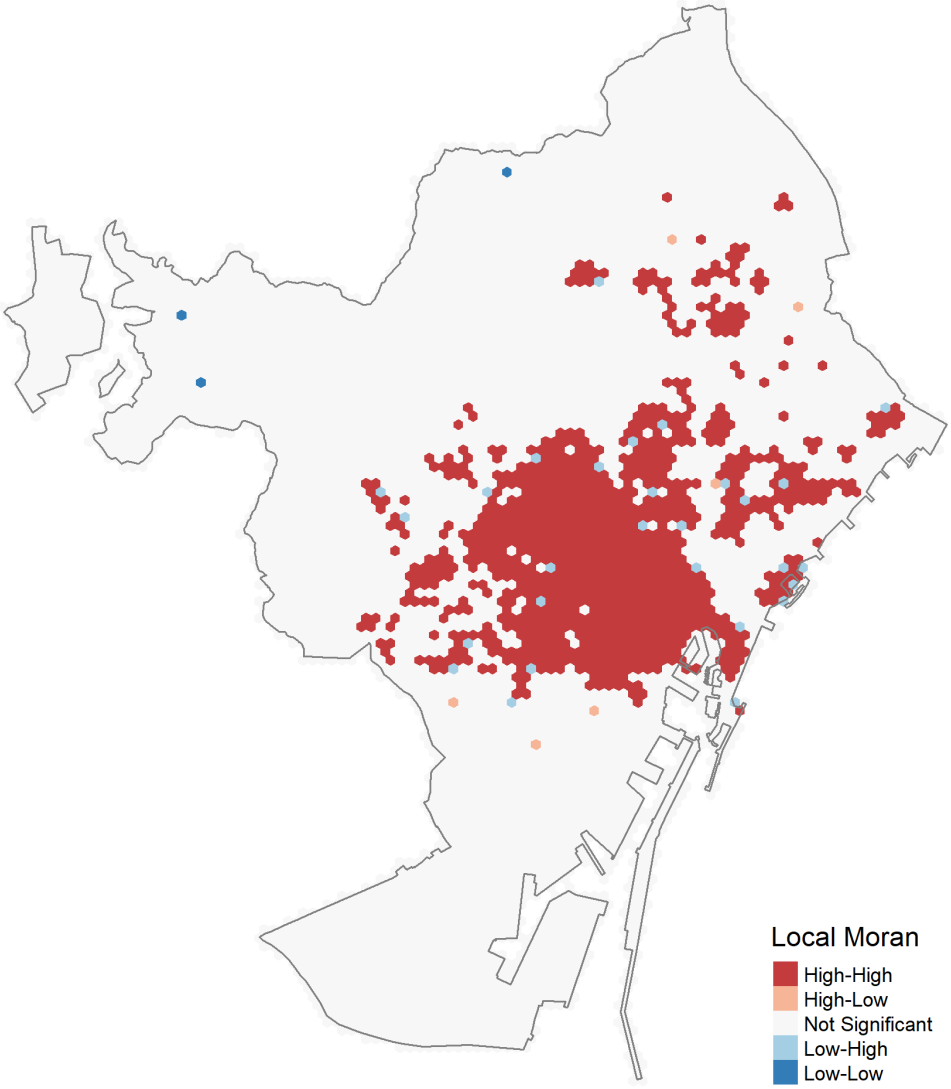


Figure 6.12: Local Moran's I corresponding to the intensity of the unique places in Barcelona collected from Instagram, aggregated in a grid of 2-hectare regular hexagons. Map shows local spatial clusters of high intensity (red) and low intensity (blue), as well of high (light blue) and low (light red) spatial outliers. Statistical significance determined using permutation-based pseudo p-values (999 permutations), with a threshold of p' < 0.05. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

G-statistics, its interpretation for any given feature depends on the combination of two factors: the sign of the statistic and the statistical significance of the result.

Positive values of the statistic indicate a hot spot (high-high cluster) of features with high values, while negative values indicate a cold spot (low-low cluster) of features with low values. This interpretation is the same as in the Moran's I statistics —discussed in section 6.3.7—, without the consideration of spatial outliers (HL or LH).

As in the case of the local Moran's I, statistical inference can be determined analytically from the z-scores of the results (assuming a normal distribution). However, because the skewed distribution of event counts that violated this assumption, the results were based on pseudo p-values computed using conditional permutation, as discussed in section 6.3.6.

Features with pseudo p-values above a threshold (the conventional value of 0.05) were deemed statistically non significant and not considered in any category. Statistical significance was further categorized into the values under the thresholds of 0.05 (statistically significant), 0.01 (highly statistically significant) and 0.001 (very highly statistically significant).

Similarly to local Moran's I statistics, local Getis-Ord G statistics can detect spatial clusters of high values (hot spots) and low values (cold spots). However, they cannot be used to identify local spatial outliers (features with values very different than values in their neighborhood). Therefore G-statistics are more useful when negative spatial correlation is in the area of study is not very frequent.

In particular, the spatial distribution of the events collected from the four studied services was clustered (Fig. 6.7), and therefore the density values exhibited positive spatial autocorrelation. This was further observed in the aggregated event counts of the Moran's I maps in figures 6.9, 6.10, 6.11 and 6.12, where the number of identified spatial outliers were rare.

In the resulting maps for the point data collected from Panoramio (Fig. 6.13), Flickr (Fig. 6.14), Twitter (Fig. 6.15) and Instagram (Fig. 6.16), the local Getis-Ord G* statistic was capable of isolating clusters of hot and cold spots, which otherwise were easily obscured by the overall randomness of data in the rasterization approach discussed in section 6.2.

However, similarly to the the local Moran's I maps, local Getis-Ord G* maps lost spatial resolution because of the aggregation of the events in 2-hectare units. Nonetheless, in contrast these maps, their Getis-Ord G* counterparts provided a more detailed measure of statistical significance, visualized with a diverging color scale with darker and more saturated hues corresponding to the higher significance levels at either end of the classification into hot and cold spots.

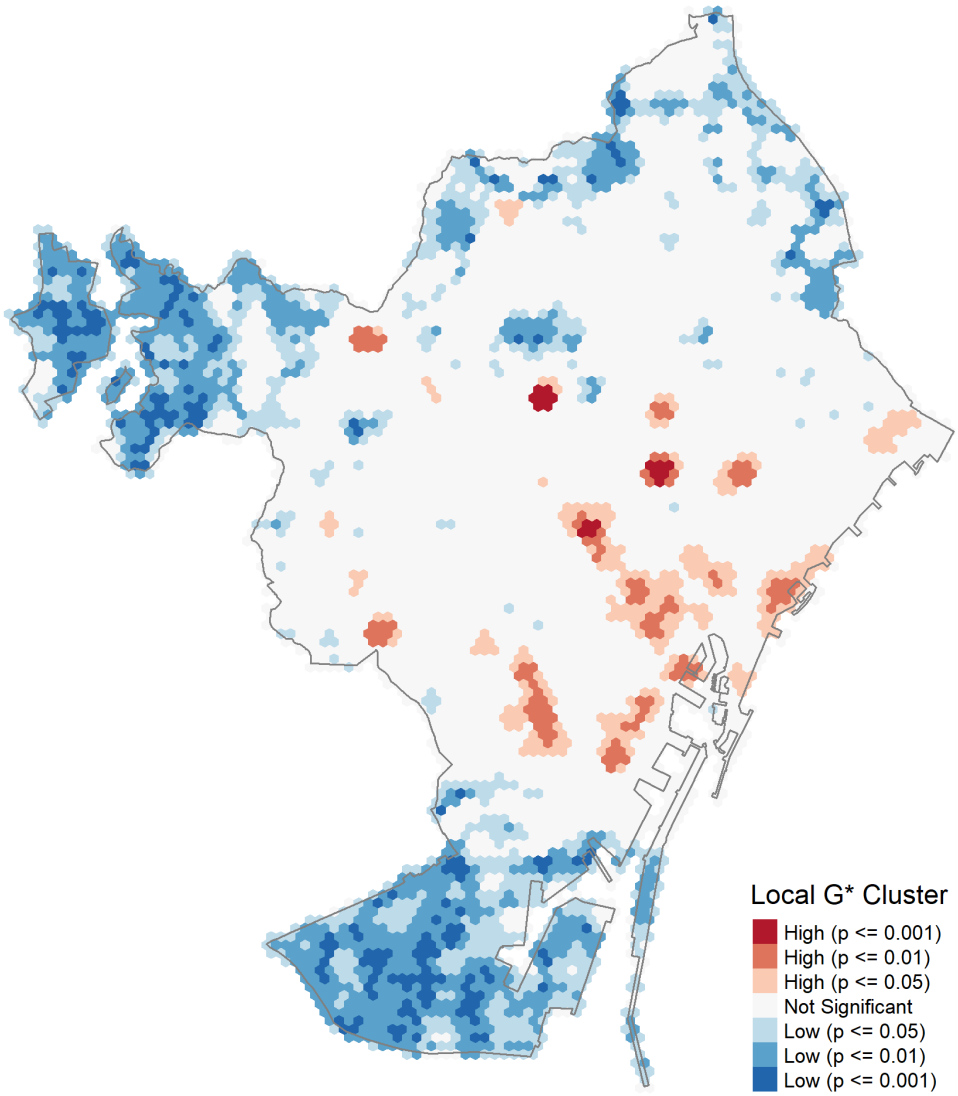


Figure 6.13: Local Getis-Ord G* corresponding to the intensity of the geotagged pictures of Barcelona collected from Panoramio, aggregated in a grid of 2-hectare regular hexagons. Map shows local spatial clusters of high intensity (red) and low intensity (blue) with darker and more saturated hues corresponding to the higher statistical significance levels. Statistical significance determined using permutation-based pseudo p-values (999 permutations), with a threshold of p' < 0.05. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

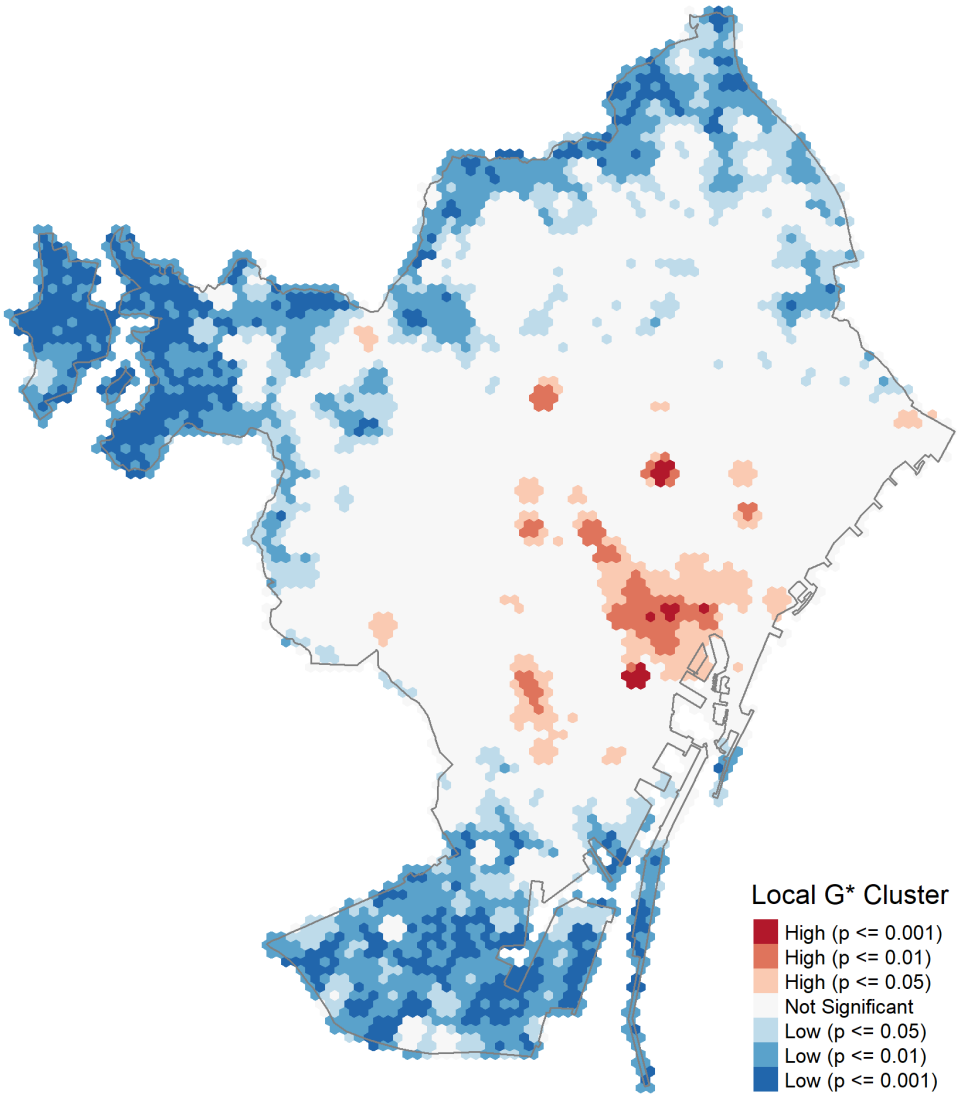


Figure 6.14: Local Getis-Ord G* corresponding to the intensity of the geotagged pictures of Barcelona collected from Flickr, aggregated in a grid of 2-hectare regular hexagons. Map shows local spatial clusters of high intensity (red) and low intensity (blue) with darker and more saturated hues corresponding to the higher statistical significance levels. Statistical significance determined using permutation-based pseudo p-values (999 permutations), with a threshold of p' < 0.05. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

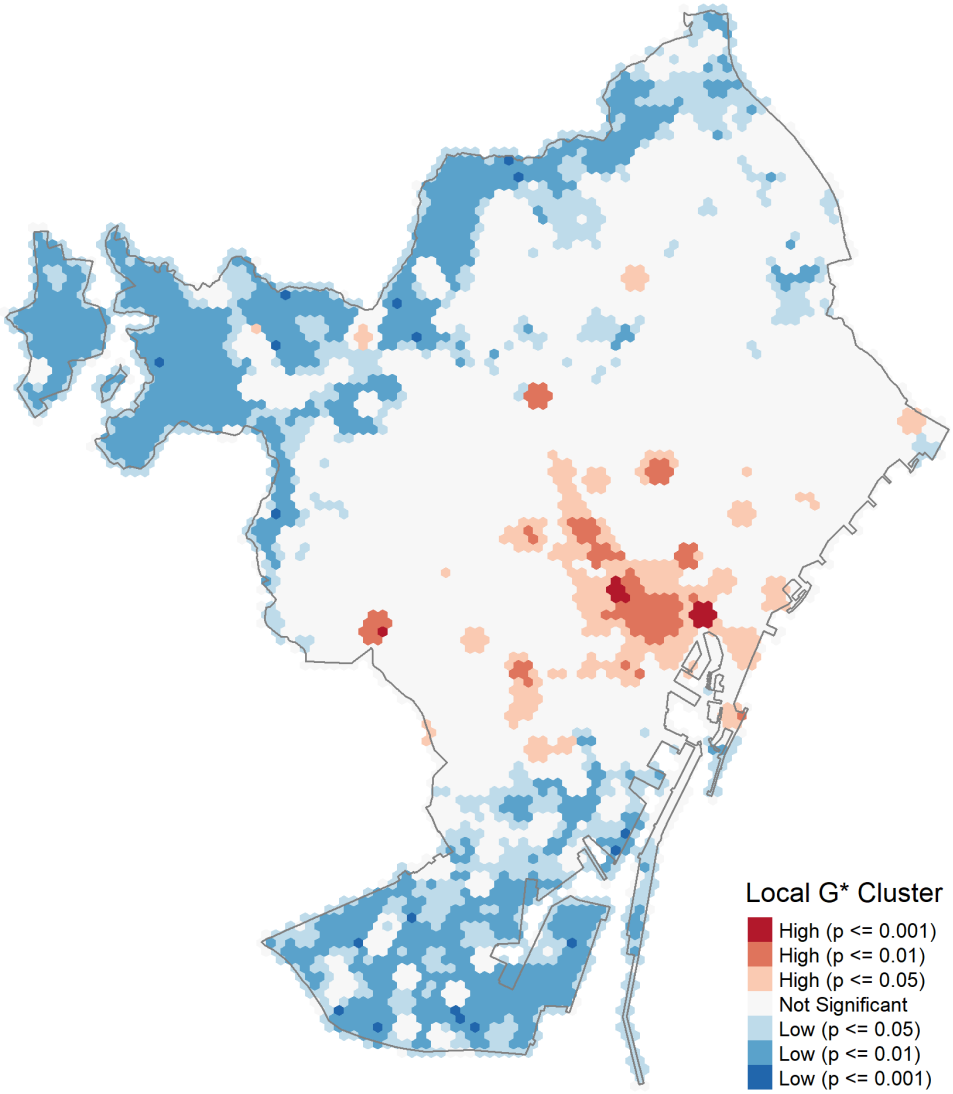


Figure 6.15: Local Getis-Ord G* corresponding to the intensity of the geotagged status messages collected from Twitter, aggregated in a grid of 2-hectare regular hexagons. Map shows local spatial clusters of high intensity (red) and low intensity (blue) with darker and more saturated hues corresponding to the higher statistical significance levels. Statistical significance determined using permutation-based pseudo p-values (999 permutations), with a threshold of p' < 0.05. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

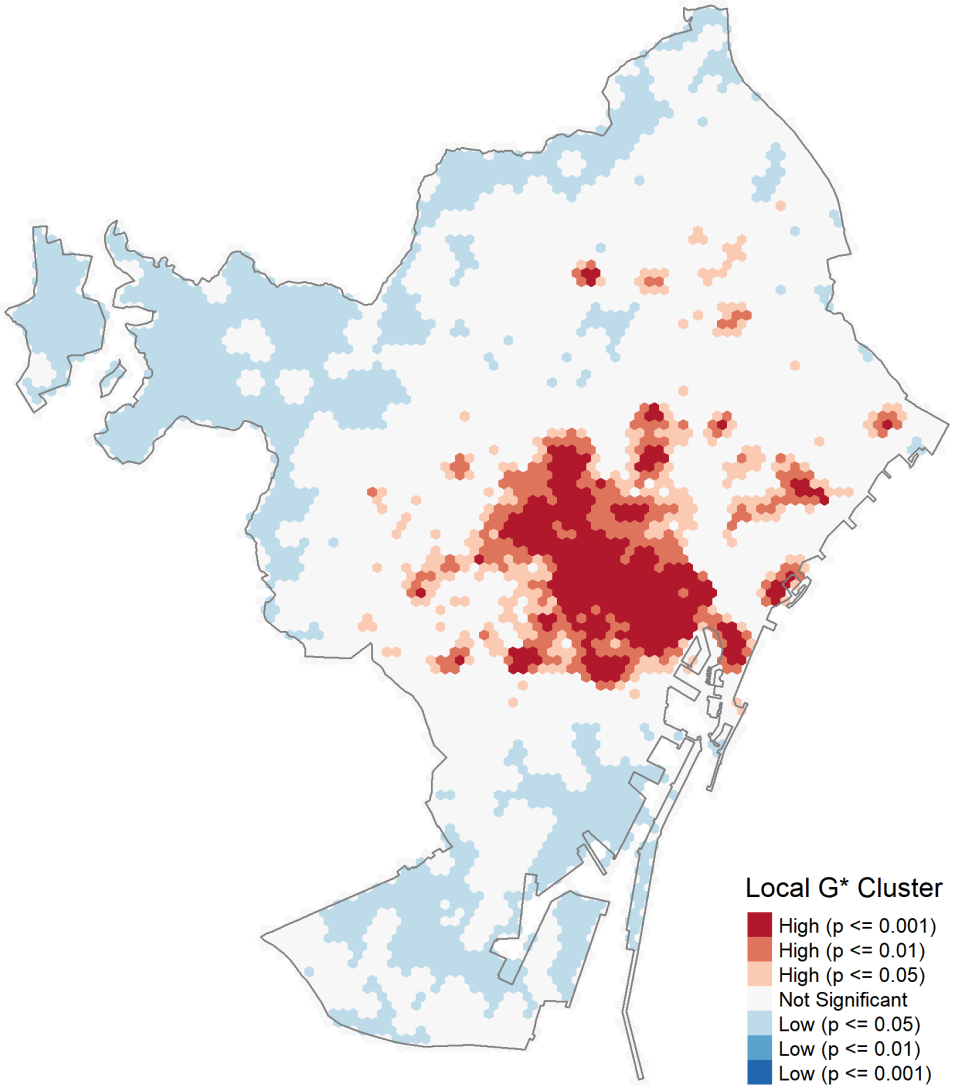


Figure 6.16: Local Getis-Ord G* corresponding to the intensity of the unique places of Barcelona collected from Instagram, aggregated in a grid of 2-hectare regular hexagons. Map shows local spatial clusters of high intensity (red) and low intensity (blue) with darker and more saturated hues corresponding to the higher statistical significance levels. Statistical significance determined using permutation-based pseudo p-values (999 permutations), with a threshold of p' < 0.05. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

6.4 Smoothing Estimation of Intensity

6.4.1 Kernel Density Estimation

The rasterization approach (discussed in section 6.2) counted the number of events inside small-sized square pixels, while the computation of the local measures of spatial autocorrelation (discussed in section 6.3) aggregated the events into relatively larger-sized hexagons. Therefore, both processes potentially introduced some biases into the analysis because of the structure of their tessellation (size, shape, orientation and offset).

The kernel estimation approach [334] partially avoids these issues estimating non-parametrically the intensity function of a point process, and is suitable if it this process is suspected to be inhomogeneous. Although the end result is a raster composed of small-sized pixels, the underlying estimated intensity function could theoretically be discretized in an arbitrarily fine grid of pixels.

The computed values of the estimated intensity function are spread across multiple pixels, in a process succinctly described by Baddeley, Rubak and Turner [326, page 168]:

> "Our favorite analogy is to imagine placing one square of chocolate on each data point. Using a hair dryer we apply heat to the chocolate so that it melts slightly. The result is an undulating surface of chocolate; the height of the surface represents the estimated intensity function of the point process. The total mass of chocolate is unchanged."

Although it is sometimes confused as a probability density —and used interchangeably as such—, the result is an estimate of the intensity of the process that generated the point pattern, expressed as the expected number of random points per unit area at every location. As the unit area is generally small (meters), the expected number of points per area unit (square meters) is consequently very small.

Kernel Density Estimation (KDE) is available in leading GIS packages (GRASS, QGIS, ArcGIS) but all the computation was performed using the spatstat 1.54-0 R package, to avoid leaving the R development environment. Other R packages considered were spatialkernel 0.4-23 [335], sparr 2.1-14 [336], which offers an spatially adaptive smoothing algorithm and splancs 2.01-40 [337], which offers a space-time kernel.

After some informal testing, the output of the spatstat's Fast Fourier Transform (FFT) convolution algorithm was found to be of much better quality than the

results provided by its ArcGIS counterpart[21], and the computation speed was much faster[22] without the memory limitations[23] imposed by the large size of the data sets, although innovative approaches are still required to process very large data sets [338].

There are thee main considerations that affect the result of the computation significantly, which will be discussed in the following subsections, while the special case of kernel density estimation on a network topology will be discussed in section 8.5:

- Kernel function selection (discussed in section 6.4.2).
- Kernel bandwidth selection (discussed in section 6.4.3).
- Transformation and classification of the results (discussed in section 6.4.4).

Beyond the fixed-bandwidth smoothers discussed, which use the same kernel and bandwidth for all locations in the observation window, other suitable approaches are spatially adaptive smoothers, which use a data-driven variable bandwidth, but were found impractical because of the large number of collected points.

6.4.2 Kernel Function Selection

There are several fixed-width kernel functions available in leading spatial analysis software (Table 6.8). The most frequent is the Gaussian kernel —although ArcGIS uses a quartic kernel approximation [334, p. 76, equation 4.5] instead—, followed by the Epanechnikov kernel. The kernel functions depend on parameter u, corresponding to the distance to the center.

The kernels are defined by a symmetrical function (Fig. 6.17). From the value at the axis of symmetry the function always decreases (or in the case of the uniform kernel, remains constant) as the distance to the axis increases. The revolution around its axis of this cross-section, placed vertically on the plane, defines the radially symmetric kernel (following the analogy of Baddeley et al. on the preceding page, the cross-section of the shape of one molten piece of chocolate).

The integral of resulting surface is the expected number of points falling into the observation window. Therefore, if the area under the curve of the kernel function is defined as one, the integral of the resulting intensity function should the original number of points. However, this is not true because of edge effects, unless edge correction is applied in the kernel estimator, using for example the

[21] Details of the ArcGIS implementation of the KDE are available at *http://desktop.arcgis.com/en/arcmap/latest/tools/spatial-analyst-toolbox/kernel-density.htm* and *http://desktop.arcgis.com/en/arcmap/latest/tools/spatial-analyst-toolbox/how-kernel-density-works.htm* at the time of writing.

[22] About 4 times faster than ArcGIS, with much better quality.

[23] Many ArcGIS libraries are still 32 bit only.

Table 6.8: Availability of kernel functions in leading spatial analysis software. Note that spatstat supports any custom kernel defined as a function f(x, y) or a convolution matrix.

Kernel	Formula	ArcGIS	GRASS	spatstat
Uniform (disc)	$K(u) = \frac{1}{2}$	-	Y	Y
Triangular	$K(u) = (1 - \lvert u \rvert)$	-	Y	*
Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2)$	-	Y	Y
Quartic	$K(u) = \frac{15}{16}(1 - u^2)^2$	Y	Y	Y
Triweight	$K(u) = \frac{35}{32}(1 - u^2)^3$	-	Y	*
Tricube	$K(u) = \frac{70}{81}(1 - \lvert u \rvert^3)^3$	-	-	*
Gaussian	$K(u \mid \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{u}{\sigma}\right)^2}$	-	Y	Y
Cosine	$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right)$	-	Y	*
Silverman	$K(u) = \frac{1}{2} e^{-\frac{\lvert u \rvert}{\sqrt{2}}} \cdot \sin\left(\frac{\lvert u \rvert}{\sqrt{2}} + \frac{\pi}{4}\right)$	-	-	*
Matrix-based	Pixel image	-	-	Y
Function-based	$f(x, y)$	-	-	Y

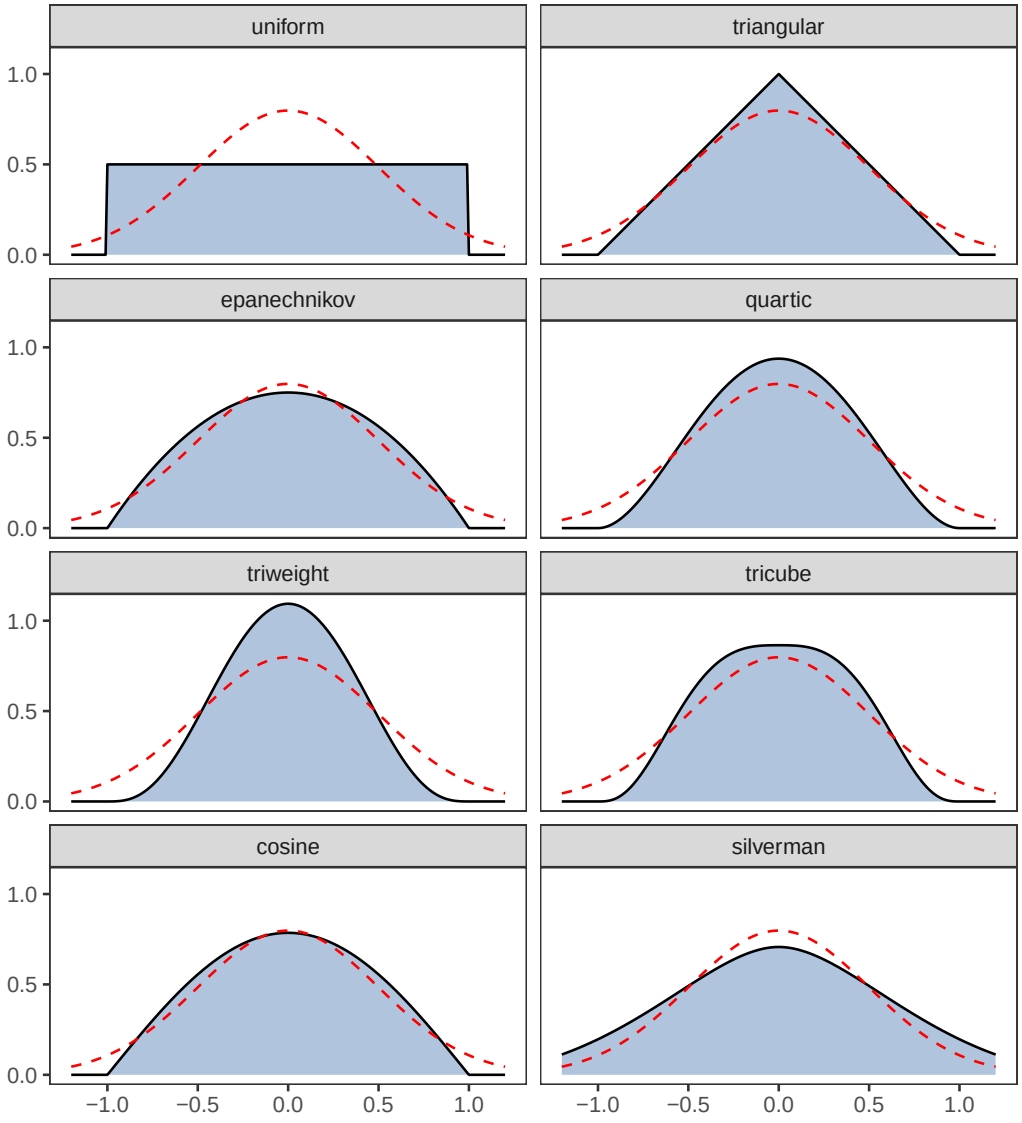


Figure 6.17: Most frequently used kernels plotted at the same scale. All kernels have the same one-unit area under the curve, and the equivalent Gaussian kernel with $\sigma = 0.5$ is overlaid for reference.

Diggle's correction [339]. The standard error can also be computed but the results are notoriously unreliable [340].

Even though relatively more expensive to compute, the kernel used in the maps (figures 6.19, 6.20, 6.21 and 6.22) was the Gaussian kernel, discretized into pixels of 25x25 meters, the same size as the pixels used in the rasterization process described in section 6.2.

6.4.3 Kernel Bandwidth Selection

The Gaussian family of functions can be parametrized with two parameters: σ (standard deviation) and μ (mean). In the case of the Gaussian kernel function, since its maximum is at K(0), the μ parameter must be zero.

In the case of other kernel functions (Fig. 6.17), they are commonly characterized in GIS software by their radius, while in the field of spatial statistics the amount of spreading is described as "half-width" or a "full width at half maximum". In the figure all functions have a radius of 1, while the corresponding overlaid Gaussian function has a σ of 0.5.

The magnitude of σ (measured in map units) determines the amount of smoothing of the result, with low values producing a rougher surface and high values resulting in a smoother surface. Therefore, too small values are not capable of smoothing noise (blurring sharp edges of high frequency detail) while too large values oversmooth the intensity and keep only the overall trend (low frequency signal).

In some cases, this σ can be conceptualized as a measure of uncertainty, because of inaccuracies in the determination of the device position using GPS technology [306] or any other factors capable of introducing some kind of positional error into the location of the collected events.

Several algorithms and rules of thumb have been developed for automatically selecting the bandwidth σ by minimizing a measure of error; among the available bandwidth selectors available in spatstat, the following were explored (Table 6.9):

Diggle uses cross-validation to choose a bandwidth to minimize the mean-square error according to the criterion defined by Diggle [339].

Stoyan rule is a rough estimate of the appropriate kernel smoothing bandwidth of the pair correlation function, using a rule of the thumb defined by Stoyan and Stoyan [341].

Loader chooses a bandwidth that maximizes the point process likelihood cross-validation criterion defined by Loader [342].

Scott is a very quick to compute rule of thumb defined by Scott [343, p. 152], useful for estimating gradual trend.

Table 6.9: Common bandwidth selection criteria and corresponding bandwidths for the retrieved sources, sorted in (mostly) ascending order.

Criterion	Panoramio	Flickr	Twitter	Instagram
Diggle	1.93 m	0.39 m	0.53 m	103.88 m
Stoyan's Rule	2.47 m	0.74 m	1.03 m	8.62 m
Loader	20.86 m	30.95 m	42.64 m	91.41 m
Scott's Rule (E-W axis)	329.16 m	165.69 m	200.63 m	454.08 m
Scott's Rule (N-S axis)	334.85 m	172.44 m	190.84 m	501.89 m
Bounding box	1,844.41 m	1,844.41 m	1,844.41 m	1,844.41 m
Fraction	3,525.99 m	3,525.99 m	3,525.99 m	3,525.99 m

Bounding box criterion selects a bandwidth based on window geometry, as a fraction of the shortest enclosing rectangle (usually 1/8).

Fraction criterion selects a bandwidth based on window geometry, as a specified quantile of the distance between two independent random points in the window (1/4 by default), based on the cumulative distribution function of the distance between two independent random points.

The Scott's rule of thumb was finally selected as the most suitable because the objective of the analysis was exploring the general trend, which roughly translated to circles of 34 ha for Panoramio, 8.6 ha for Flickr, 12.6 ha for Twitter and 64.8 ha for Instagram. Incidentally, the Loader point process likelihood cross-validation criterion produced similar to the ones chosen for the rasterization method, discussed in section 6.2.

6.4.4 Empirical Cumulative Distribution Function Transformation

The result of the kernel estimation of the intensity was a raster composed of a grid of 25x25 meter pixels, with a range of values much less extreme than the ones obtained in the quadrat count methods (discussed in sections 6.2 and 6.3), because very high local concentrations of events were spread over multiple pixels within the kernel bandwidth.

While the distributions (Table 6.10) were significantly less skewed and had much narrower ranges than the corresponding distribution in the rasterization methodology (Table 6.4), it was also necessary a transformation of the values, but in this case the log1p transformation was considered less suitable.

The proposed transformation was driven by the data, converting the values to

Table 6.10: Summary statistics of the results of the kernel density estimation in a 25x25 meter grid, converted to probabilities (per million) dividing by the corresponding total number of points. Note that the mean is always the inverse of the number of effective pixels (in this case 1/163,782) per million.

Source	Min	Max	Mean	Median	Std. Dev.	Skewness	Kurtosis
Panoramio	≈0	57.53	6.11	2.25	8.77	2.25 > 0	8.34 > 3
Flickr	≈0	227.20	6.11	0.85	16.65	5.39 > 0	39.44 > 3
Twitter	≈0	141.48	6.11	1.67	12.59	4.52 > 0	30.07 > 3
Instagram	≈0	29.61	6.11	3.33	6.76	1.39 > 0	4.36 > 3

quantiles using the corresponding empirical cumulative distribution function (ECDF) for each of the sources. An ECDF is a step function whose result for any input value is the fraction of observations that are less than or equal to this value, and therefore the range of the result is between zero and one.

To visualize ECDFs of the four retrieved sources (Fig. 6.18), they were potted as a function of the density per pixel resulting from the kernel estimation 6.18a. However, as the ranges were different across the sources, another set of ECFGs where the values were normalized across their ranges (dividing by their corresponding maximum value) was also produced (Fig. 6.18b).

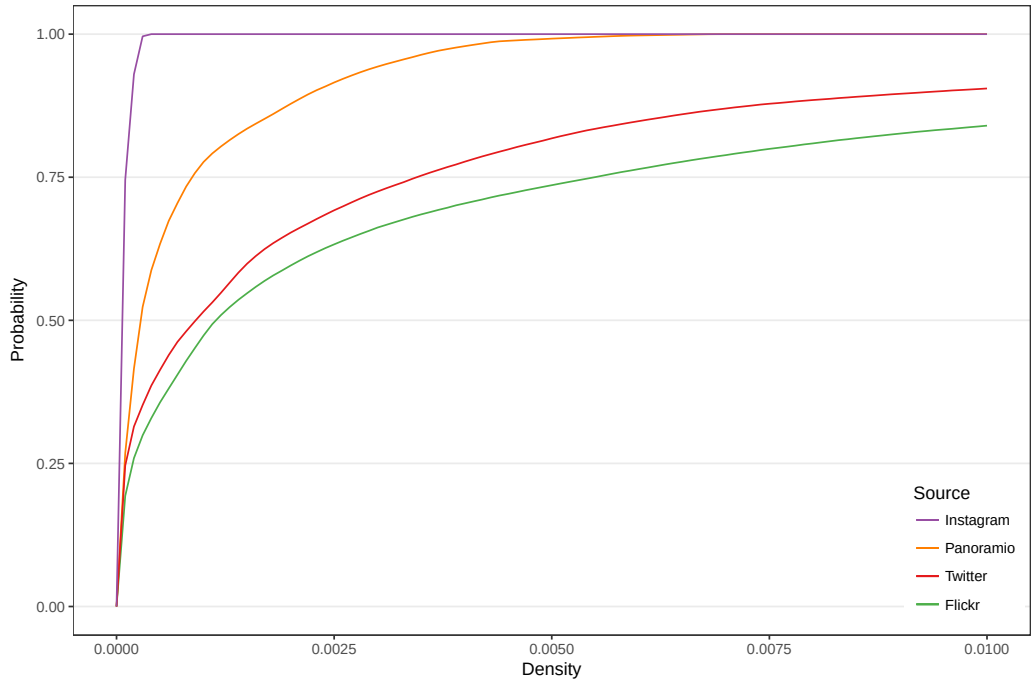
According to the ECDFs 6.18a, Instagram reaches the maximum probability earlier because of its shorter range, while Flickr exhibits the most slow-growing function as it has the broadest range (about an order of magnitude higher than Instagram).

Conversely, when using a normalized range between each source's minimum and maximum (Fig. 6.18b), the clustering or dispersion of the values (without spatial considerations) can be observed, with Instagram exhibiting proportionally more high values (well-exposed image), while Twitter values being eminently in the low end of the range (under-exposed image).

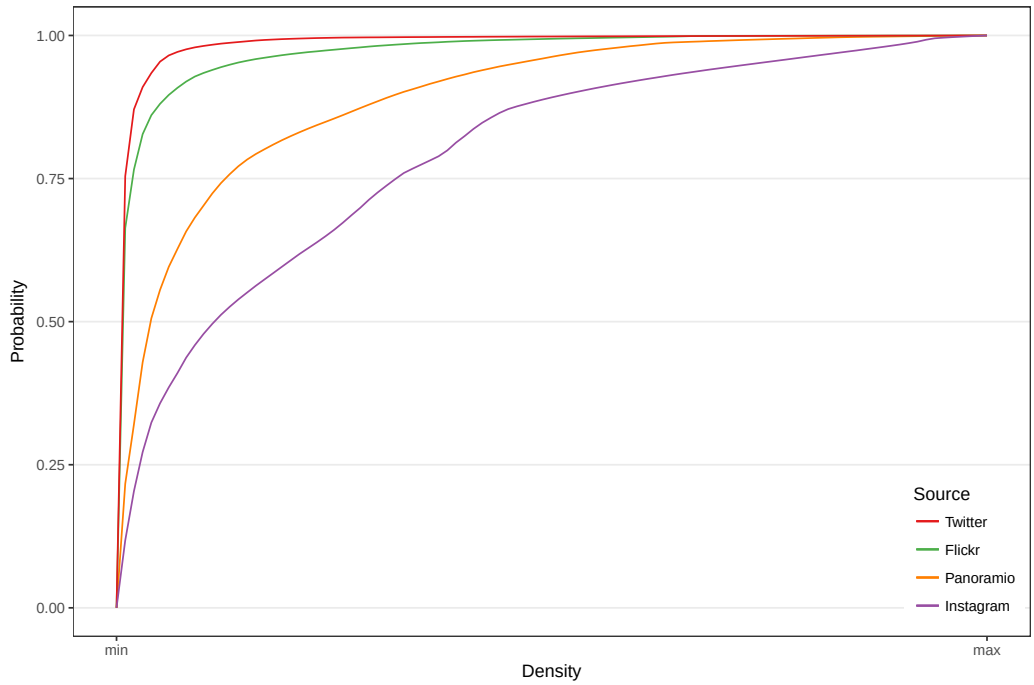
6.4.5 Classification and Representation

Using the original values as the inputs of the ECDF, the values were converted to quantiles in the range zero to one. This technique is similar to the image processing method named "histogram equalization" which increases the global contrast of raster images, especially when the detail concentrates in a narrow portion of the range.

The transformed values were the result of plugging the original values into the argument of the ECDF. The cummulative distribution of these transformed values is a straight line, whose intercept is at the origin and its slope is the inverse of the



(a) ECDFs of the probabilities plotted in the same 0–0.01 range



(b) ECDFs of the probabilities with stretched ranges between their minimum and maximum values

Figure 6.18: Empirical cumulative distributions of pixel values of the kernel density estimation for the collected sources in a common (top) and stretched (bottom) range.

maximum value in the distribution, and would appear as a diagonal line between the bottom-left and top-right corners of the bottom figure (Fig. 6.18b).

With this methodology the values were classified according to their relative position within their distribution, but it was still necessary to visualize how they were distributed spatially. This was only possible because the pixels were all the same size, and therefore the classification of the values in data space directly translated to the same proportions of area in geometric space. To emphasize the most extreme values, they were classified and displayed according to the following criteria[24]:

- The values within the interquartile range (IQR), corresponding to the middle 50% between the 25th and 75th percentiles, were colored gray.
- The values corresponding to the bottom quartile (below the 25th percentile) were colored in blue hues, while the values corresponding to the top quartile (above the 75th percentile) were colored in red hues.
- The colors in these quartiles were represented with darker and more saturated hues for increasingly extreme values, corresponding to the top 75th, 95th and 99th (reds) and bottom 25th, 5th and 1st (blues) percentiles.

The results of this representation approach resembled simplified contour maps, avoiding the visual clutter introduced by the isolines[25]. In the resulting maps of the kernel estimation of the intensity of the point data retrieved from Panoramio (Fig. 6.19), Flickr (Fig. 6.20), Twitter (Fig. 6.21) and Instagram (Fig. 6.22), the areas with outstanding high and low concentration of points appeared clearly visible, as well as the areas where the intensity was neither remarkably high or low (the values around the median within the IQR, in gray).

The kernel smoothing successfully showed the general trend in the intensity, avoiding the randomness that appeared as high-frequency detail in the rasterization approach discussed in section 6.2, because the error was spread across multiple pixels according to the automatically determined kernel bandwidth parameter, making easier to identify areas with unusual absence (deserts) or outstanding presence (hot spots) of events.

[24]Which translated to the intervals defined by the breaks [0, 1, 5, 25, 50, 75, 95, 99, 100].
[25]Contour lines are also referred as isolines, isopleths, or isarithms.

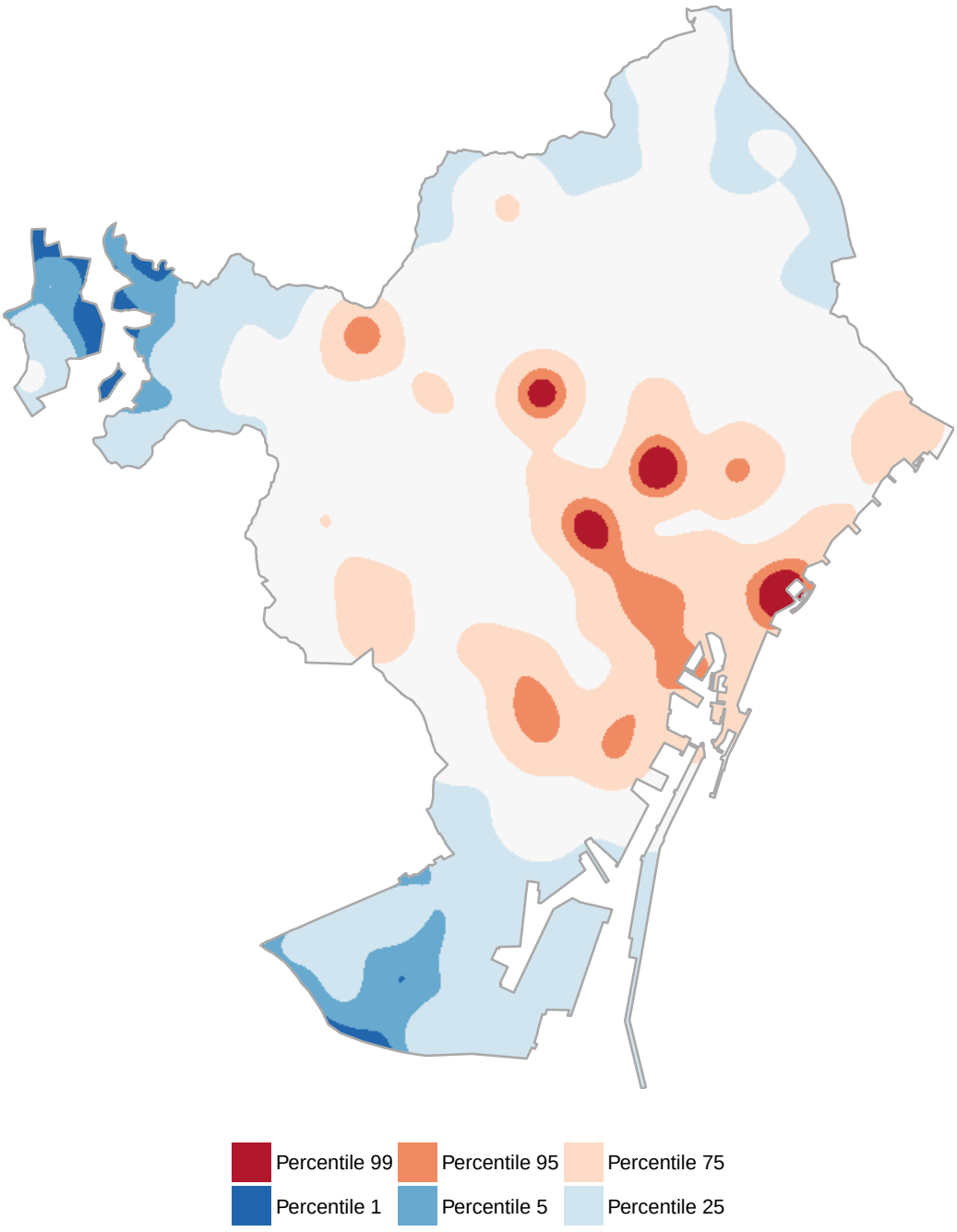


Figure 6.19: Kernel estimation of the point process intensity of the geotagged pictures of Barcelona collected from Panoramio, in a grid of 25x25 meter cells. Interquartile range is represented as neutral gray, top quartile in red hues and bottom quartile in blue hues. More extreme values use darker and more saturated colors. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

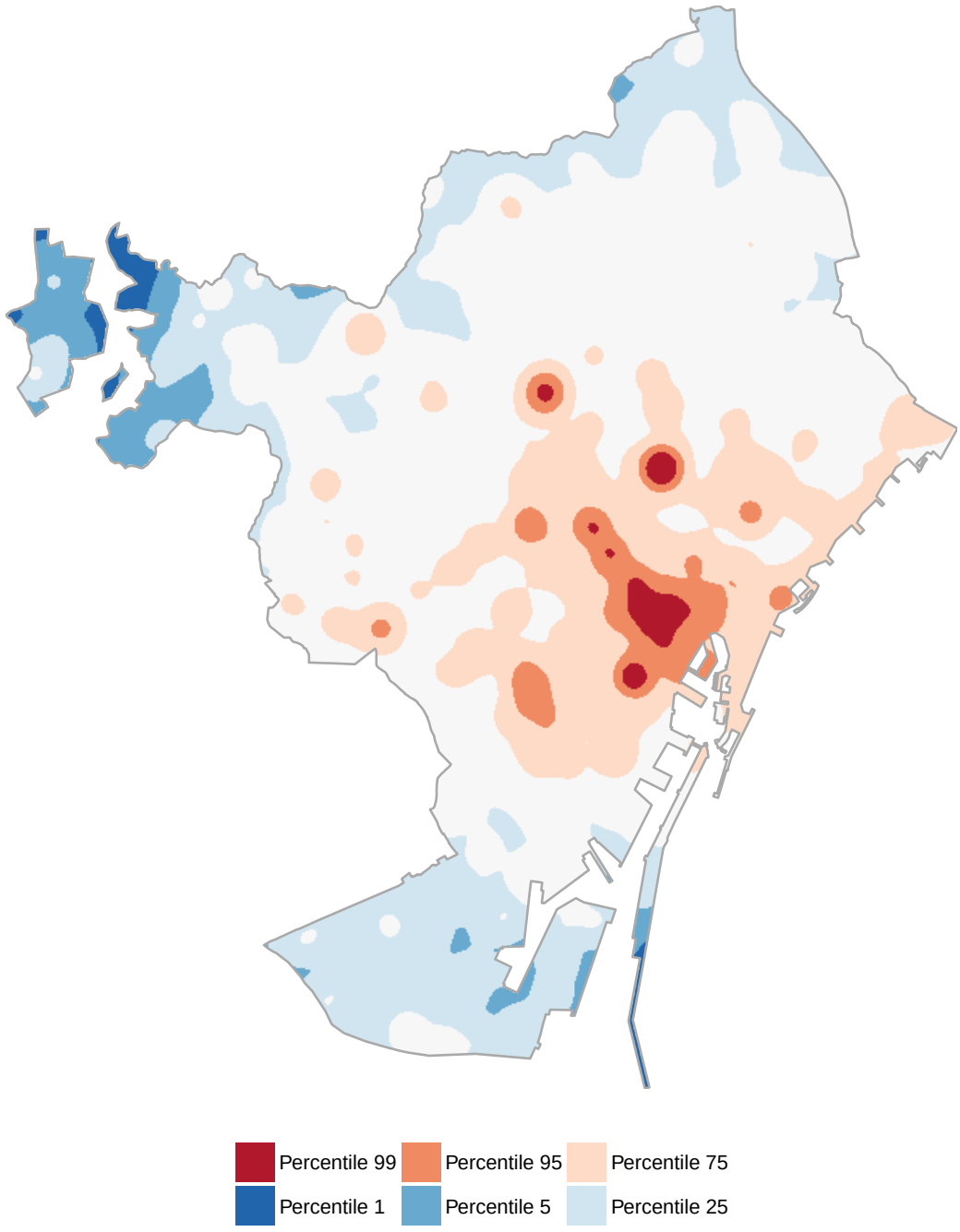


Figure 6.20: Kernel estimation of the point process intensity of the geotagged pictures of Barcelona collected from Flickr, in a grid of 25x25 meter cells. Interquartile range is represented as neutral gray, top quartile in red hues and bottom quartile in blue hues. More extreme values use darker and more saturated colors. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

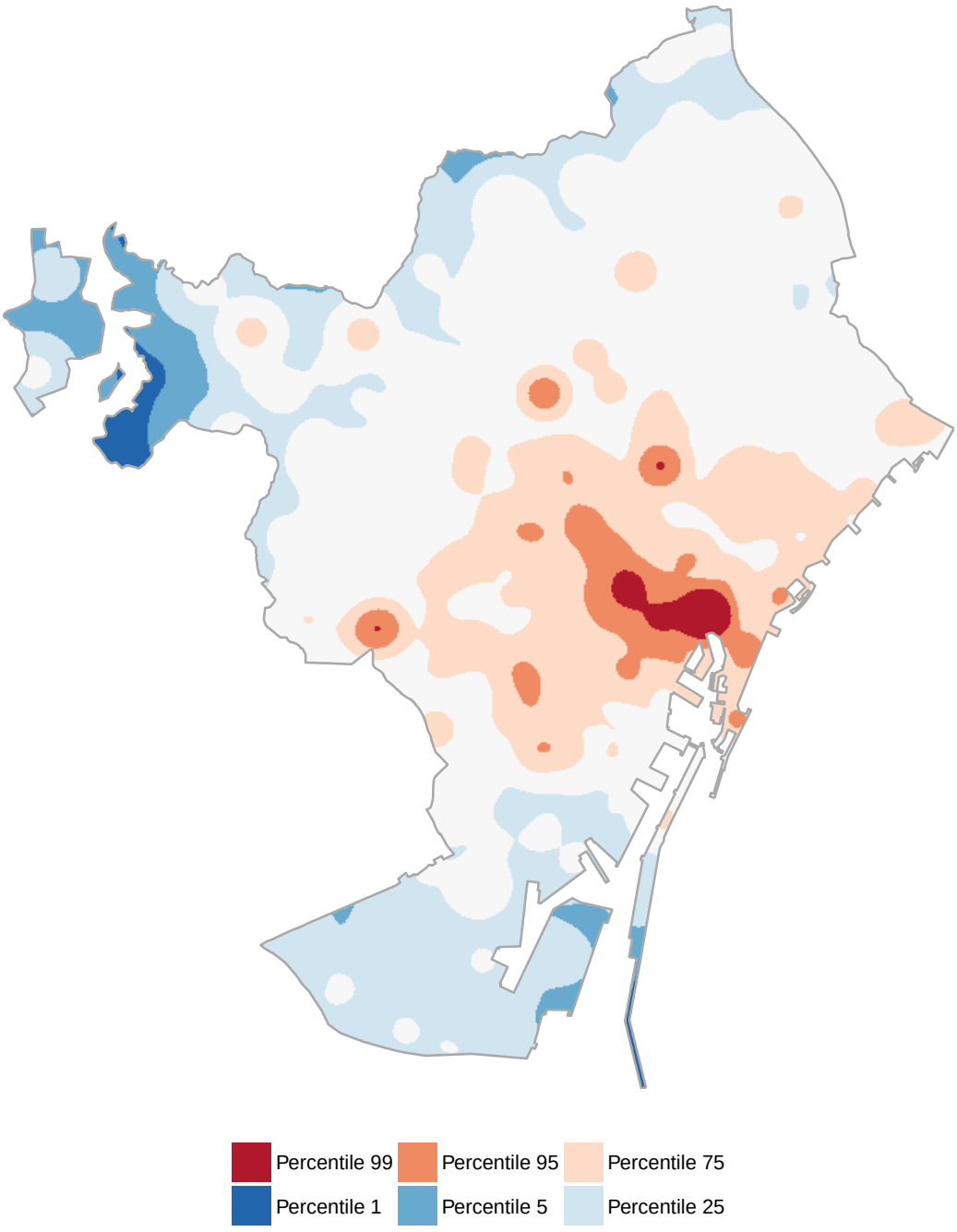


Figure 6.21: Kernel estimation of the point process intensity of the geotagged status messages collected from Twitter, in a grid of 25x25 meter cells. Interquartile range is represented as neutral gray, top quartile in red hues and bottom quartile in blue hues. More extreme values use darker and more saturated colors. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

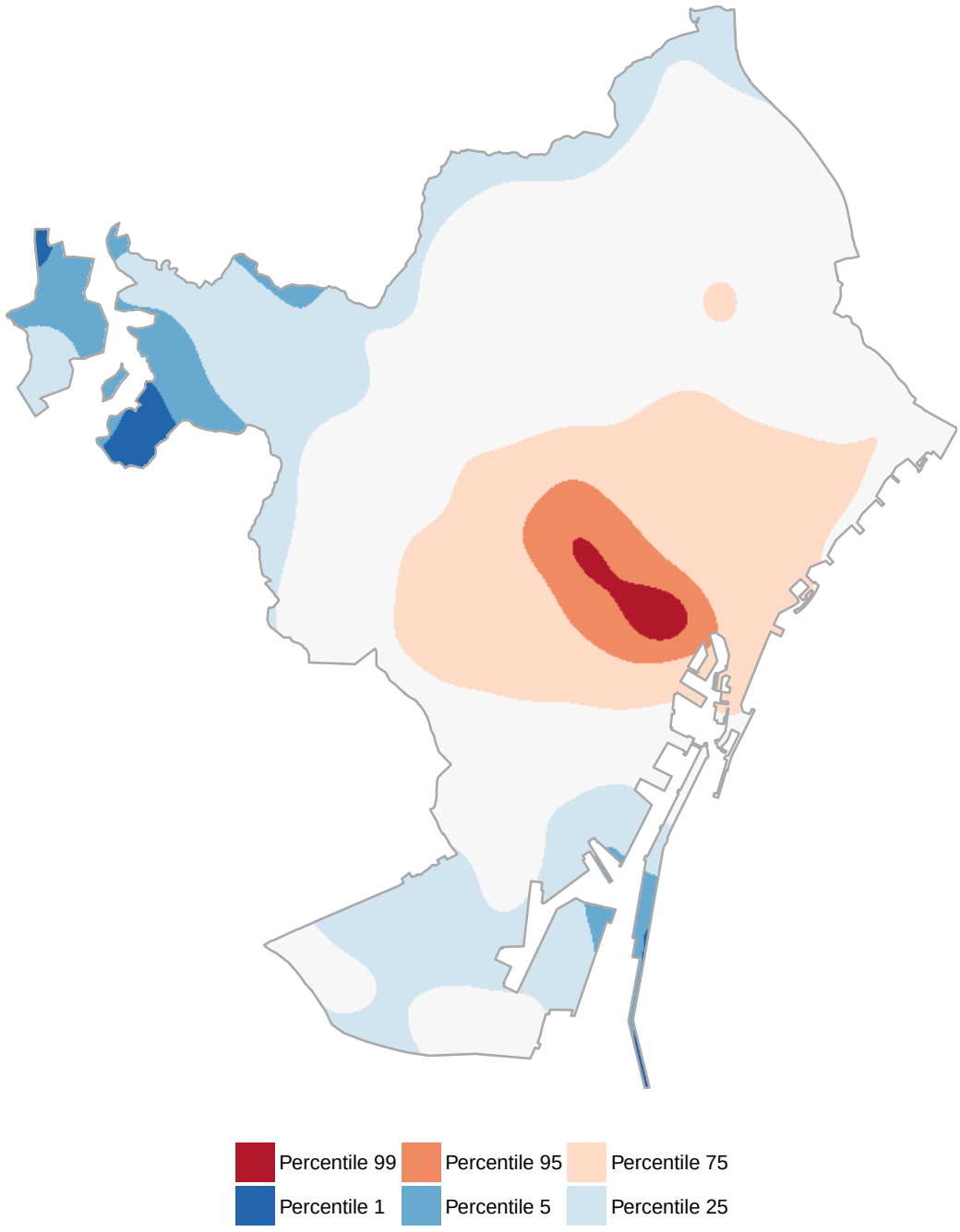


Figure 6.22: Kernel estimation of the point process intensity of the unique locations of Barcelona collected from Instagram, in a grid of 25x25 meter cells. Interquartile range is represented as neutral gray, top quartile in red hues and bottom quartile in blue hues. More extreme values use darker and more saturated colors. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. City limits cartography from CartoBCN, under CC BY 3.0.

Chapter 7

The Neighborhood Scale

> "The point of cities is multiplicity of choice"
>
> Jane Jacobs

7.1 Visualizing Neighborhood-Scale Data

7.1.1 Administrative Boundaries

The city of Barcelona is divided into 10 districts (Table 7.1), each of them administered by a "regidor" (councilor), with powers in local issues such as urbanism, making these divisions both administrative and political.

The current delimitation was approved in 1984 and obeys in part to historic reasons, being some of the districts former towns[1] around the city of Barcelona, incorporated in the 19th and 20th centuries. Since 2009, the districts were further divided into a total of 73 neighborhoods, whose delimitation respond to different characters granted by their historic background, complementing —without substituting— the existing district division.

Since the neighborhood and district boundaries obey to real differences in the character of the different units, they present the opportunity to study the distribution of the observed phenomena with a high degree of confidence that are not biased by the modifiable area problem [305, 286], which occurs when point observations are aggregated into arbitrarily defined units.

[1] Sarrià, Les Corts, Sant Andreu de Palomar, Gràcia or Sant Martí de Provençals.

Table 7.1: Number of neighborhoods per district and corresponding numeric codification.

Code	Name	Neighborhoods
1	Ciutat Vella	4
2	Eixample	6
3	Sants-Montjuïc	8
4	Les Corts	3
5	Sarrià-Sant Gervasi	6
6	Gràcia	5
7	Horta-Guinardó	11
8	Nou Barris	13
9	Sant Andreu	7
10	Sant Martí	10

7.1.2 Limitations of Choropleth Maps

The number of neighborhoods —each with its own character— makes displaying data while at the same time providing a spatial context challenging. The usual approach to provide this context is to display data in a map.

Choropleth maps are the most used[2] visualization to represent spatial data, where the polygons representing aggregation units are filled with a color scale according to a measured magnitude. This technique has some limitations:

- When using maps a lot of the space in the visualization is wasted showing the geometry.
- Only a single variable can be displayed (which can be partially solved using small multiples, a collection of maps with the same size displayed side by side).
- As values are classified into intervals before being encoded as colors, much of the variance is lost.
- Colors can be difficult to perceptually interpret correctly, especially in individuals with color blindness[3].
- Only magnitudes can be shown, excluding line graphs to visualize time series or scatter plots to visualize related variables.
- It is not possible to sort and accurately compare values as in bar graphs.
- If not showing densities, the different areas of the polygons distort the data, and in large-scale maps the projection choice distorts the geometry itself.

[2] And some times overused.

[3] Achromatopsia (monochrome vision), tritanopia (blue), deuteranopia (green) or protanopia (red).

Other types of visualization —scatter plots, bar graphs, line graphs— are adequate to represent certain types of data but lose the spatial context, particularly if the audience is not already familiar with the geography.

7.1.3 Linked Micromaps

Linked micromaps [344, 345] offer the possibility of using multiple visualization types while providing a limited geographic context in the form of a micromap (a map thumbnail) highlighting the displayed (linked) subset of data.

The methodology generally begins sorting the features according to a given variable, not necessarily represented. Features are then sorted from top to bottom and partitioned according to this order into groups of roughly the same number of members.

All the information corresponding to a group is laid out in horizontal stripes. Each of these groups has a map, with its members highlighted in different colors. These colors are recycled in each of the groups and match the corresponding linked panels.

The stripes can include any number of specialized statistical visualizations, containing only the members of the group, but sharing the same vertical and horizontal scales between related panels, which are laid out vertically, creating columns of related charts.

7.1.4 Representation Challenges

While a high level of cartographic detail is generally desirable —and often highly sought after, despite the cost of producing accurate cartography—, one of the issues with the map thumbnails used in micromaps arise when excessive detail —relative to the scale— is present in the polygon boundaries.

In this case, especially when polygon boundaries have a large fractal dimension, the meandering nature of the outlines —strokes in illustration programs— make them appear as thick lines because the accumulation of bends in a short space, leaving less available surface area to color-code the polygon fill.

This effect is produced because, according to the Nyquist-Shannon theorem, the borders need to be at least a pixel wide to be visible (being the pixel resolution she spatial version of the sampling rate), and as the map gets smaller the border to area ratio increases. This issue is aggravated in the case of polygons that are smaller at the displayed scale.

To overcome these issues, the original high resolution cartography was transformed —while keeping the features recognizable— with two objectives, the details

of which are discussed in section 7.2:

- Reducing detail through geometry simplification.
- Increasing the area of small polygons.

7.2 Micromap Generation

7.2.1 Software

The original Linked Micromaps[4] software is available from the Geographic Information Systems and Science for Cancer Control[5] within the National Cancer Institute of the U.S. Department of Health & Human Services, to simplify comparing multiple variables across regions (generally states and counties) as well as across time.

Unfortunately, its functionality is limited to linking flat files —formatted as comma-separated values (CSV)— with the provided United States cartography only, without the option of using another cartographic source. The software is written in Java [346] and runs on the Microsoft Windows operating system, but has not been updated since 2008. As alternatives, at the time of writing there were two R packages[6] capable of producing this kind of visualizations:

micromap in its version 1.9.2 [347], released[7] on February 6, 2015 and developed by personnel of the US Environmental Protection Agency (EPA) and the Oregon State University, and dependent on the ggplot2 [172] R package.

micromapST in its version 1.1.1 [348], released[8] on December 7, 2016, which can be considered the spiritual successor to the original Linked Micromaps software and includes Daniel B. Carr of George Mason University among its developers.

Both packages offered multiple glyphs but micromapST had more options available out-of-the-box (Table 7.2). For this reason, the chosen package was micromapST[9],

[4]The Linked Micromaps software is available at *http://gis.cancer.gov/tools/micromaps/* at the time of writing.

[5]The the Geographic Information Systems and Science for Cancer Control website is available at *http://gis.cancer.gov/* at the time of writing.

[6]With the added benefit of being multi-platform (Windows, macOS and Linux).

[7]The micromap package is available from CRAN at *http://cran.r-project.org/package=micromap* at the time of writing.

[8]The micromapST package is available from CRAN at *http://cran.r-project.org/package=micromapST* at the time of writing.

[9]A fork with a patch by the author is available on GitHub at *https://github.com/Suppaman/micromapST* at the time of writing.

Table 7.2: Comparison of the glyphs available in the two R packages capable of producing linked micromaps at the time of writing.

Glyph Type	micromapST 1.1.1	micromap 1.9.2
Label	Yes	Yes
Map	4 groupings	4 groupings
Arrow	Yes	-
Bar plot	4 types	2 types
Box plot	Yes	Yes
Dot plot	4 types	2 types
Scatter plot	Yes	-
Time series	Yes	-

despite being arguably less developer-friendly when using non-US cartography and having less flexibility in its grouping functionality.

7.2.2 Source Data

The cartography of the Barcelona administrative divisions was retrieved on Jun 23, 2017 from the official CartoBCN[10] portal in the ESRI Shapefile[11] format. The downloaded data[12] used the ETRS89 / UTM ZONE 31N (EPSG:25831) coordinate reference system, and was distributed under the Creative Commons CC BY 3.0 license[13]. According to its metadata[14], this data was produced on March 15, 2011 by the *Departament del Pla de la Ciutat* (Department of the City Map) and contained the following divisions (Fig. 7.1):

Districts Dataset produced digitizing the division of the city limits into 10 districts, approved on March 7, 1984 and published in the *Gaseta Municipal* (Official Gazette) number 9 of March 30, 1984.

Neighborhoods Dataset produced digitizing the neighborhood division, defined by the *Gabinet Tècnic de Programació de l'Ajuntament de Barcelona* (Barcelona City Council Programming Technical Bureau), approved by the City Council on December 22, 2006 and published in the *Gaseta Municipal* (Official Gazette) number 7 of February 28, 2007.

[10]CartoBCN is vailable at *http://w20.bcn.cat/cartobcn/* at the time of writing.
[11]CartoBCN also offered the cartography in DGN (v8), DWG and KMZ formats.
[12]The cartography used was named "BCN_Barri_ETRS89_SHP".
[13]Details available at *http://creativecommons.org/licenses/by/3.0/* at the time of writing.
[14]In the accompanying file named "ETS89_Info_BCN_DIVADM_CA.pdf"

In the feature class, each of the 73 neighborhoods was represented as a polygon[15], with fields storing the corresponding name and unique numeric code, as well as the name and unique numeric code of the district it belonged to. Each polygon contained population data for male and female inhabitants, with the total population being the sum of both quantities.

Although the metadata did not provide a nominal scale for the source data, its geometry was very detailed. To be usable to produce linked micromaps, the polygon boundaries had to be simplified and the smallest polygons enlarged.

7.2.3 Geometry Simplification

Cartographic generalization is a process to produce a coarse map from detailed geometry that would otherwise be only suitable for small scale —small denominator, large fraction— maps. One of the benefits of this technique is that it is only necessary and maintain a single set of cartographic data, with the maximum possible accuracy, and derive the rest of the cartographic products from its geometry.

One of the most common generalization methods is geometry simplification, where the objective is to eliminate unnecessary detail while preserving the overall shapes. The resulting geometries are less complex and suitable for display at larger scales —large denominator, small fraction— while still being recognizable.

The simplification algorithm is given an objective, which is generally the proportion of nodes to be removed[16], and in some cases one or more threshold parameters to determine the conditions under which a node should be removed or kept (to control how much detail is preserved).

The parameter choices are highly dependent on the shape of the original geometries, and insufficient simplification can result in unsuitable geometries for the intended display scale, while excessive simplification can produce unrecognizable shapes and/or undesirable removal of the smallest features (Fig. 7.2).

7.2.4 Algorithm Choice

One of the earliest examples of geometry simplification algorithms is the Ramer-Douglas-Peucker algorithm [349, 350] —generally referred to as the Douglas-Peucker algorithm or iterative end-point fit algorithm—, which is still one of

[15] Some of the entities were actually multipolygons (in the simple feature geometry specification), as they contained multiple parts.

[16] An interactive demonstration of a simplification algorithm is available at *http://bost.ocks.org/mike/simplify/* at the time of writing.

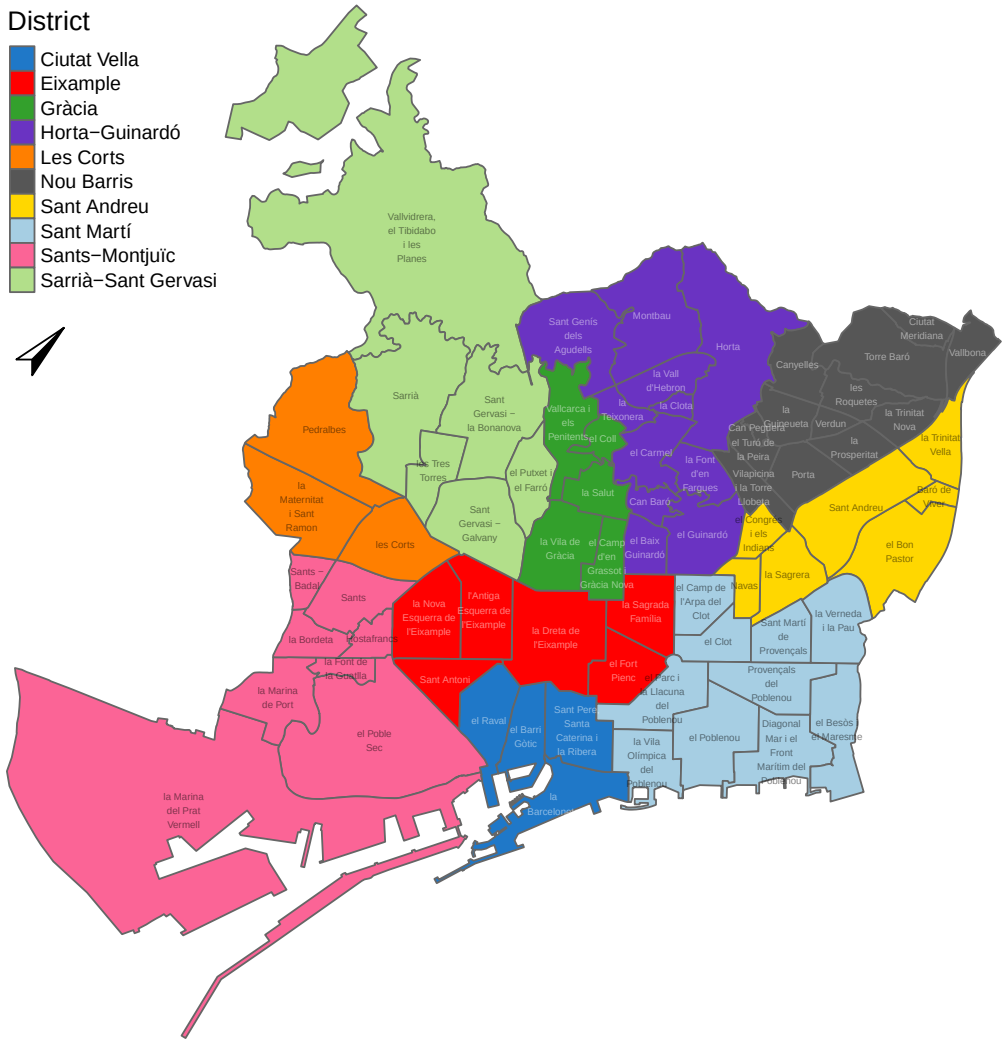


Figure 7.1: Map of the 73 neighborhoods of Barcelona, colored according to the district where they belong. Map is rotated 45 degrees clockwise. Administrative divisions cartography from CartoBCN, under CC BY 3.0.

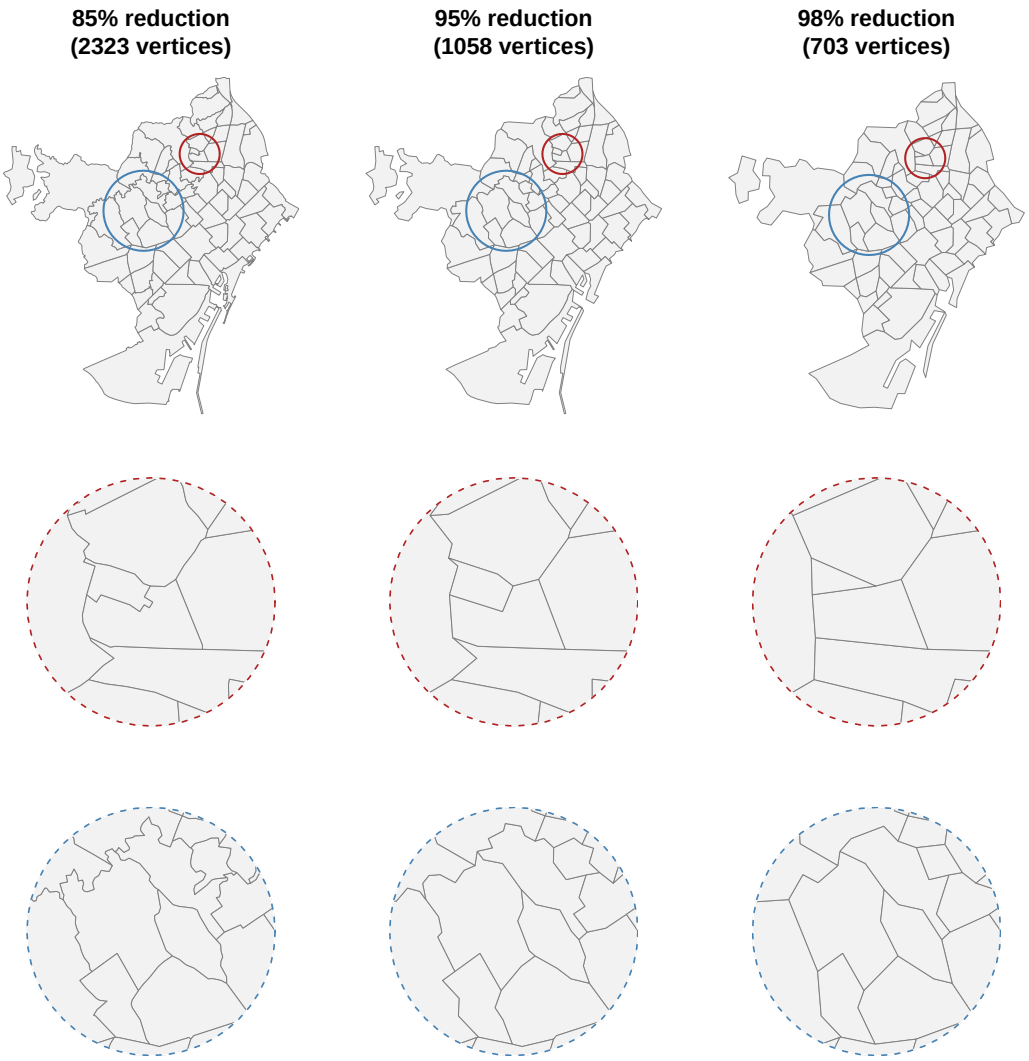


Figure 7.2: Effect of different simplification ratios applied to the geometry of the Barcelona neighborhoods, subjectively sorted from insufficient simplification (left) to excessive simplification (right), according to the intended display resolution. The blue circle diameter is double of the red circle (and four times the area). Maps produced using the using the Visvalingam-Whyatt algorithm as implemented in the mapsharper R package. Administrative divisions cartography from CartoBCN, under CC BY 3.0.

the most widely implemented methods. The algorithm produces a subset of points from the original curve, keeping always the starting and end points, and recursively removing the points that are closer than a defined threshold to the segment defined by these points.

In contrast, the Visvalingam-Whyatt [351] methodology is more intuitive, progressively removing points that result in a less dissimilar curve, defined by the sum of the all the areas of the triangles defined by three consecutive points in the curve.

Both the Ramer-Douglas-Peucker and the Visvalingam-Whyatt algorithms were implemented by the R package rmapshaper 0.3.0 [352], a wrapper around the mapshaper[17] JavaScript library by Matthew Bloch that powers the Mapshaper[18] online editor for map data.

Two additional algorithms implemented in the ArcGIS generalization toolset[19] were also considered: the shape recognition technique defined by Wang and Müller [353] that replaces bends in features with straight segments within a tolerance, and the weighted area algorithm defined by Zhou and Jones [354] which weights triangles of effective area for each vertex, in an approach similar to the Visvalingam-Whyatt algorithm.

Finally, it is worth mentioning that the open source GRASS GIS [355] offers additional simplification algorithms (Lang and Reumann-Witkam) and also six smoothing algorithms in its v.generalize[20] command.

After as series of tests, the Visvalingam-Whyatt was the chosen simplification algorithm because it produced the most pleasing results, with less artifacts while efficiently reducing the number of vertices in the polygons (Fig. 7.3).

7.2.5 Preserving Topological Integrity

The administrative divisions polygons tessellate the space within the city limits[21], without neither overlaps (a region of space that belong to more than one polygon) nor gaps (a region of space not covered by any polygon).

A naive algorithm tries to simplify each feature independently, allocating the

[17] The mapshaper source code is available on GitHub at *http://github.com/mbloch/mapshaper/* at the time of writing.

[18] The Mapshaper website is available at *http://mapshaper.org/* at the time of writing.

[19] An overview of the ArcGIS generalization toolset is available at *http://desktop.arcgis.com/en/arcmap/latest/tools/cartography-toolbox/an-overview-of-the-generalization-toolset.htm* at the time of writing.

[20] GRASS vector based generalization documentation is available at *http://grass.osgeo.org/grass74/manuals/v.generalize.html* at the time of writing.

[21] Excluding the harbor waters.

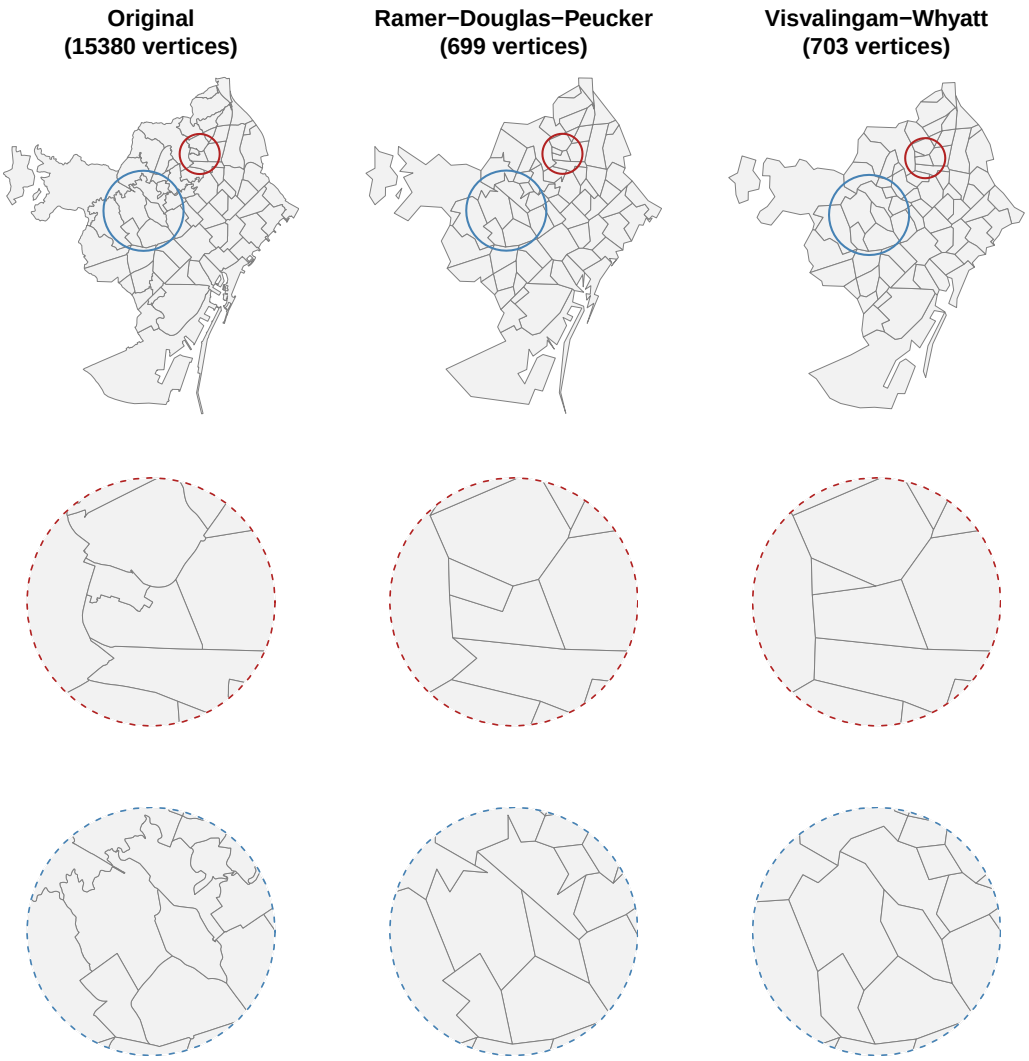


Figure 7.3: Comparison of the original Barcelona neighborhoods cartography and its generalization with the two simplification methods available in the rmapshaper R package, both retaining 2% of the original points. Administrative divisions cartography from CartoBCN, under CC BY 3.0.

vertex budged according to the shape of each polygon. While this approach can produce slightly better results if each feature is inspected in isolation, it is not adequate in the case of administrative divisions.

With this approach, as the shared vertices and edges originally present in neighboring polygons are kept or discarded independently for each feature, the process does not preserve the original topology, and can potentially introduce overlaps and gaps not present in the original geometry.

In contrast, topologically-aware geometry simplification (Fig. 7.4) preserves these properties, ensuring that shared borders between features are altered geometrically but not topologically during the simplification process.

In addition, since districts are divided into neighborhoods, and any given neighborhood can belong to a single district, the process guarantees by design that the resulting simplified neighborhoods can be aggregated into new simplified districts using a unary union (dissolve) operation, and the result of this process will yield a topologically correct tessellation of districts, with boundaries matching the outer perimeter of their neighborhoods.

7.2.6 Small Polygon Exaggeration

The simplification process addressed most of the issues to produce a suitable geometry for the map thumbnails. However, due to the small relative size of some of the neighborhoods, the corresponding polygons in the thumbnail map were difficult to identify, as the area occupied by their borders almost covered their interior area. Furthermore, some small polygons were reduced by the simplification process to small sliver-like triangles (Fig. 7.2).

In was therefore necessary to produce a "map caricature", in the sense that the size of some features had to be enlarged, while preserving the overall readability of the administrative divisions, recognizable through their shape, position and magnitude. Generally, this process is conducted by hand —as in the case of the US map in the original Linked Micromap software—, but in this occasion the process was automated by creatively using the cartogram representation technique.

A cartogram is a map where a variable of interest is represented by distorting the geometry, generally the area in the case of magnitudes or the distance in the case of travel time. There are several variants of cartogram:

Continuous area cartograms which preserve the topology and the overall shape of the original non-distorted polygons[22].

[22]Examples of continuous area cartograms can be found at *http://www.worldmapper.org/* at the time of writing.

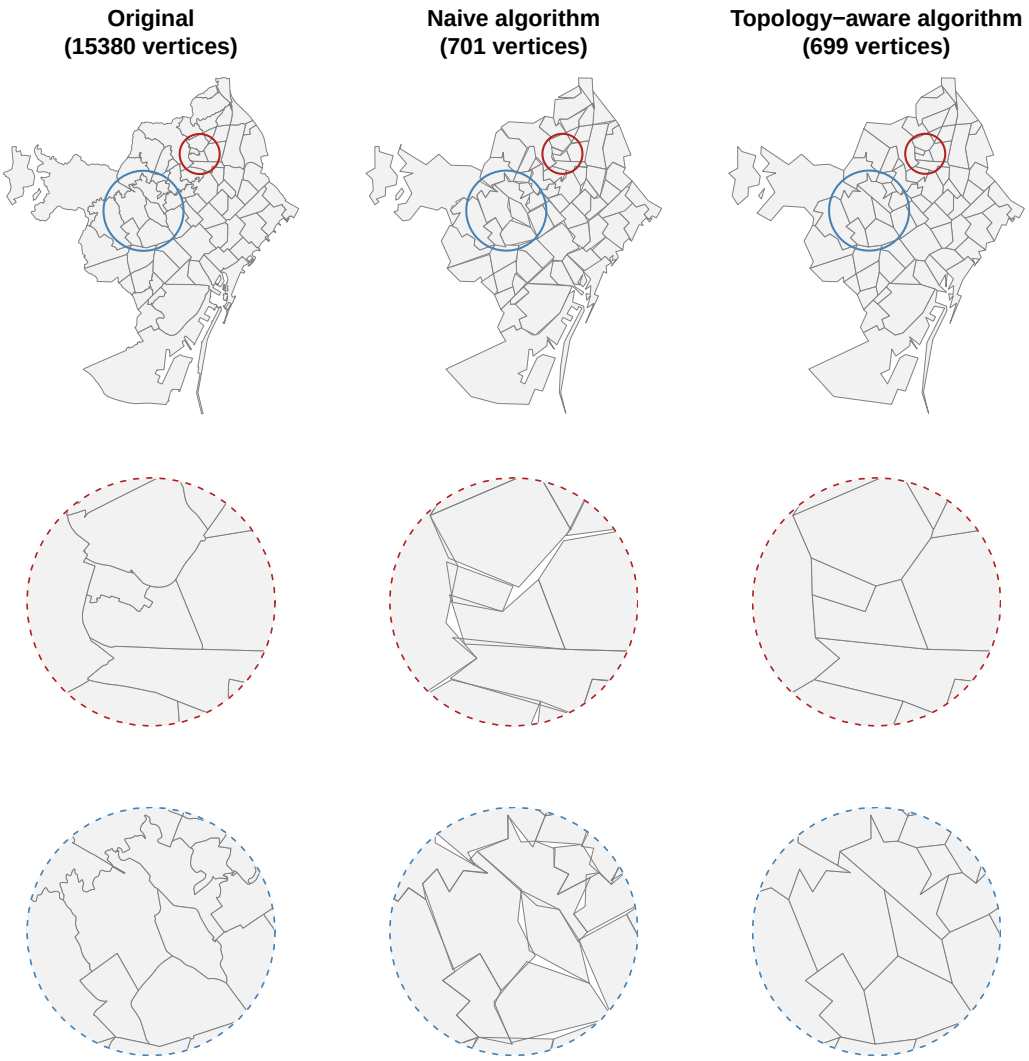


Figure 7.4: Geometry simplification, using the Ramer-Douglas-Peucker algorithm, comparing a naive algorithm and a topology-aware algorithm results to the original geometry. The naive algorithm introduces gaps and overlaps not present in the original data. Administrative divisions cartography from CartoBCN, under CC BY 3.0.

Scaled-down cartograms which only preserve the shapes but reduce their sizes, keeping the original centroids, according to the variable of interest.

Dorling cartograms, which replaces the actual shapes with proportionally-sized circles in the same overall position as the original shape (treemaps, discussed in section 5.2.4 can be considered related to this technique).

To increase the area of small polygons, subtly distorting the map while keeping the features recognizable, an algorithm to construct continuous area cartograms [356] was used, from the R package cartogram 0.0.2 [357], among the different algorithms available [358].

The variable represented was the area of the neighborhood itself, tweaked to increase the magnitude of low values —which corresponded to small polygons— instead of a 1:1 relationship. Two adjustment curves were tested (Fig. 7.5), and the selected adjustment method was the second (lifting without clipping), which seemed to produce better results visually:

Clipping the lowest values below a defined *a'* threshold, where the tweaked area was defined for each area *a* as:

$$f(a) = \begin{cases} a' & a \leq a' \\ a & a > t \end{cases}$$

Lifting the lowest values, scaling the range using linear interpolation between the true area range (from *u* to *v*) and a new range between a new minimum value *u'* and the original maximum value (from *u'* to *v*):

$$f(a) = \frac{(a-u)\cdot(v-u')}{v-u} + u'$$

In both cases, the resulting values were re-scaled so their sum was the same as the sum of the true areas as follows, and therefore the aggregate size of the distorted polygons was the same as the original ones:

$$a_i = f(a_i) \cdot \frac{\sum a}{\sum f(a)}$$

7.2.7 Map Orientation

The traditional map orientation or Barcelona rotates the north orientation 45 degrees clockwise (Fig. 7.1) to place the harbor approximately parallel to the bottom edge of the map, the Besòs (east) and Llobregat (west) rivers at the sides of the map, and the Collserola mountain range at the top of the map.

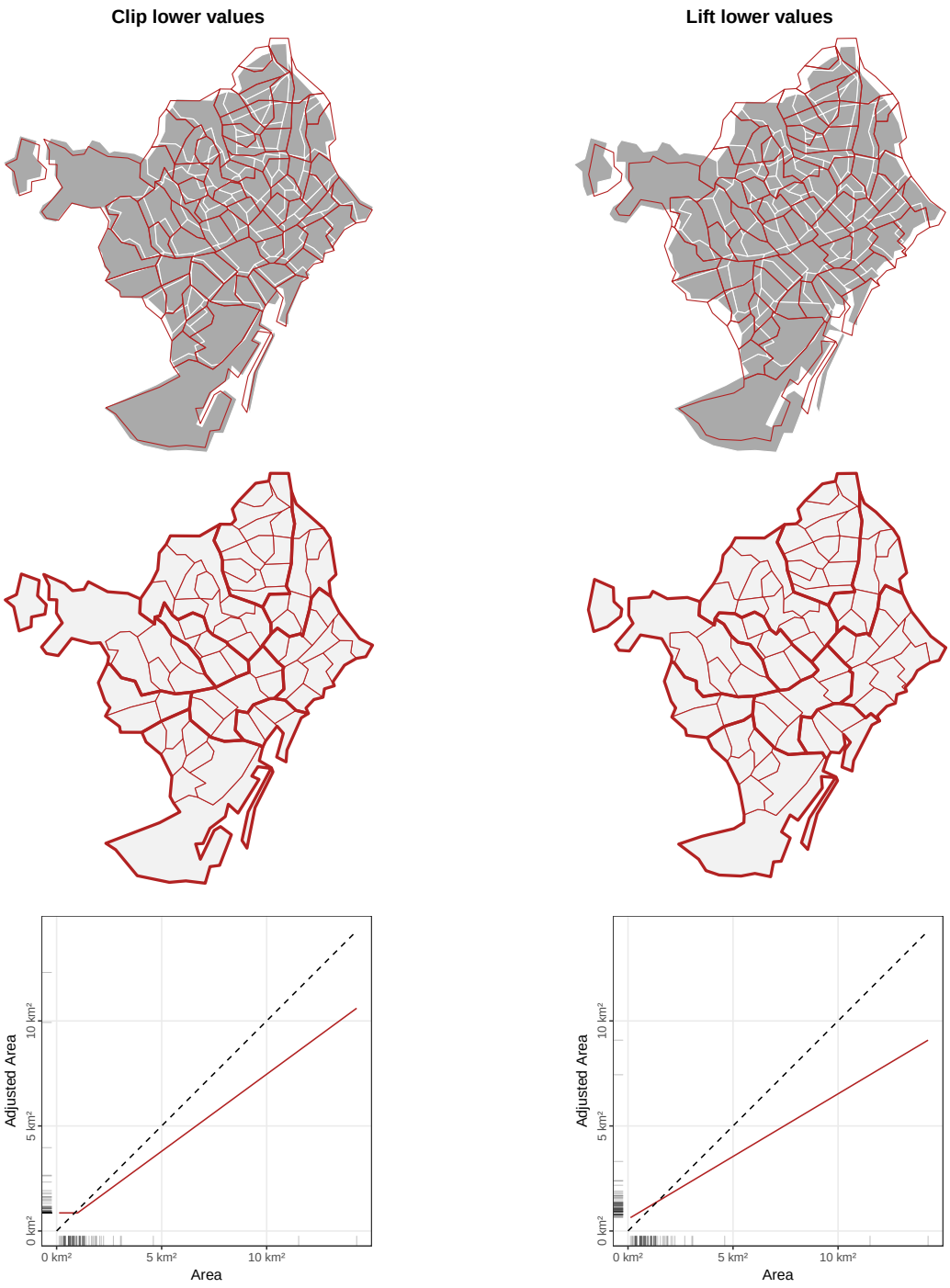


Figure 7.5: Adjustment methodologies tested to exaggerate the dimensions of the smallest polygons. The top row shows the results (red) compared to the geometry without adjustment, both using the Visvalingam-Whyatt algorithm. The second row shows only the tweaked geometry with the districts boundaries as thicker lines. The third row shows the functions used to transform the areas. Administrative divisions cartography from CartoBCN, under CC BY 3.0.

Figure 7.6: Map of the Barcelona subway network, with the north rotated 45 degrees clockwise. Source: Transports Metropolitans de Barcelona website.

This orientation roughly follows the historical axes of the roman *decumanus maximus* (today Carrer del Bisbe and Carrer Ciutat) and *cardo maximus* (today Carrer Ferran and Carrer Jaume I[23]) intersecting at Plaça de Sant Jaume, and more recently, the street network in the layout of the *Eixample* by Ildefons Cerdà.

Even in the age of mobile devices equipped with mapping applications that favor either the usual "north up" orientation or the "first person perspective" of GPS navigation, this orientation is still favored in maps of Barcelona used daily, such as the map of the subway public transportation network (Fig. 7.6).

Therefore, the map thumbnails were rotated accordingly, as their objective was to easily link the neighborhood location with the presented data, and this orientation was more familiar to the audience of the maps. The rotation operation did not cause additional distortion, as the simplification algorithms were invariant to the rotation of the geometry and did not favor any particular orientation.

[23] Or, according to some sources, carrer del Call and carrer Llibreteria.

7.3 Picture Intensity across Neighborhoods

7.3.1 Measures of Intensity

In chapter 6 it was established that the distribution of the retrieved locations was not homogeneous, concentrating in some areas and leaving large areas devoid of any event. The adopted approach, while robust and reproducible, ignored the differences in the character of different parts of the city.

This section introduces a new complementary approach —more grounded on the actual integrated structure of the city— through its division into neighborhoods. This approach focuses on three intensity measures computed to identify their variation across the neighborhoods, using the linked micromaps discussed in section 7.2:

- Intensity per gross area unit (discussed in section 7.3.2).
- Intensity per street length unit (discussed in section 7.3.3 and further developed in chapter 8).
- Intensity per capita (discussed in section 7.3.4).

All location counts were computed as a yearly average to obtain a more meaningful metric, not dependent on the temporal range of the data. Therefore, the corresponding counts were divided by the number of yearly periods covered by the data set (which in the case of Flickr locations was 12 years).

Since all three indicators were a quotient, it was not sufficient to provide a single figure —as higher values could be the result of large numerators and/or small denominators, and vice versa with lower values—, and it was necessary to visualize the results using a bi-variate charts (scatter plots).

In the linked micromaps, the neighborhoods were sorted according to the corresponding intensity metric. In the linked scatter plots, which showed the numerator in the vertical axis and the denominator in the horizontal axis, this sorting order could be imagined as determined by an imaginary line sweeping the points clockwise, starting from the vertical axis.

A more traditional approach using a scatter plot —losing the spatial reference provided by the micromaps, but benefiting from more space available for the figure itself— was also included for comparison for each of the computed intensities. Considering the intended audience was familiar with the neighborhood and district locations, this shortcoming was reduced labeling the features —in this research only the five neighborhoods with outstanding high or low values were labeled—, and coloring the points according to the neighborhood they belonged to (with the same color scheme as in figure 7.1 on page 227).

Table 7.3: Area, average yearly pictures and corresponding densities for the ten districts of Barcelona.

District	Gross area (ha)	Yearly pictures (count)	Yearly density (pictures / ha)
Ciutat Vella	436.85	29,051.83	66.50
Eixample	747.64	22,233.83	29.74
Gràcia	418.55	6,627.42	15.83
Horta-Guinardó	1,194.71	2,148.17	1.80
Les Corts	601.76	3,284.42	5.46
Nou Barris	804.15	422.58	0.53
Sant Andreu	656.53	840.17	1.28
Sant Martí	1,052.38	6,860.92	6.52
Sants-Montjuïc	2,294.05	14,248.92	6.21
Sarrià-Sant Gervasi	2,009.29	4,844.50	2.41

7.3.2 Intensity per Gross Area

The spatial distribution of picture intensity was discussed in chapter 6, where all the analysis units had the same size and shape, adopting a regular tessellation. However, when analyzing the same data per administrative unit, the aggregation units do not have the same area, as evidenced when broken down by district (Table 7.3), and more notably when comparing the neighborhoods, where the smallest (*Can Peguera*) has an extension of less than 12 hectares while the largest (*la Marina del Prat Vermell*) measures almost 1,500 hectares.

As a consequence, even under of the assumption of homogeneous intensity (hypothesis rejected in section 6.4), the expectation of the number of events retrieved would be proportional to the size of the aggregation unit. It was therefore reasonable to homogenize the event counts dividing the obtained counts by the neighborhood total gross area (two other homogenization approaches are discussed in section 7.3.3 and section 7.3.4).

The resulting linked micromaps (Fig. 7.7) show that the old quarter (*el Barri Gòtic*) is by far the most photographed neighborhood, closely followed by a group of popular tourist destinations (*la Sagrada Família*, *Sant Pere*[24], *el Raval*, *la Salut* and *la Dreta de l'Eixample*) as shown in the dot plot column.

The density values in these dot plots are used as the sorting criteria for the neighborhoods, and provide a summary of the distribution of the intensity of picture-taking activities within Barcelona, while the corresponding micromaps

[24]The proper name of the neighborhood is "Sant Pere, Santa Caterina i la Ribera".

allow locating each neighborhood and recognize the clustering of the most picture-dense areas.

Since the density is a quotient, it summarizes the relative magnitudes of both numerator and denominator into a single value. Therefore, a high density can be consequence of the presence of many pictures and/or the small size of the aggregating unit, and conversely a low density can be the result of the opposite situation, along with intermediate situations in between.

Accordingly, the scatter plots beside the dot plots allow relating these picture counts and areas. This is useful to conclude that *el Barri Gòtic* is intensely photographed despite its small size —recording as many pictures as *el Poble Sec*, whose area is much larger— or that the largest neighborhoods (*Vallvidrera*[25] and *la Marina del Prat Vermell*) are not very photographed despite their large size.

In contrast, when using a traditional scatter plot (Fig. 7.8), the reference to the location of the neighborhood is lost, and the density is encoded in the positions of the points as the angle between them and the origin from the horizontal axis (provided non-transformed scales are used). However, an extra covariate can be included as the color, size or shape of the dots (in the case of the figure, the population of the neighborhood is shown as the size of the dot, with the color of the district it belongs to).

7.3.3 Intensity per Street Length

As in many cities, and in particular those with a long history, the street network configuration is not uniform across Barcelona (discussed in section 8.2). In particular, the ratio between road length and gross area[26] is very variable and reflects the differences in the urban fabric of neighborhoods and districts (Table 7.4). This metric was used instead of the area occupied by the streets themselves, because it was considered that could reflect more accurately the relation of users walking *along* streets.

Because of the limitations of the density per area unit (discussed in section 7.3.2), a new metric was developed, computed per neighborhood as the quotient of the number of pictures taken within its administrative limits, divided by the length of its street network (computed from the geometry, as discussed in section 8.2.3). This metric captures two linked factors simultaneously:

- The pressure of the visitors (but also of locals, as discussed in chapter 5) on certain street segments as a result of their picture-taking activity.

[25]The proper name of the neighborhood is “Vallvidrera, el Tibidabo i les Planes”.
[26]Gross area included streets as well.

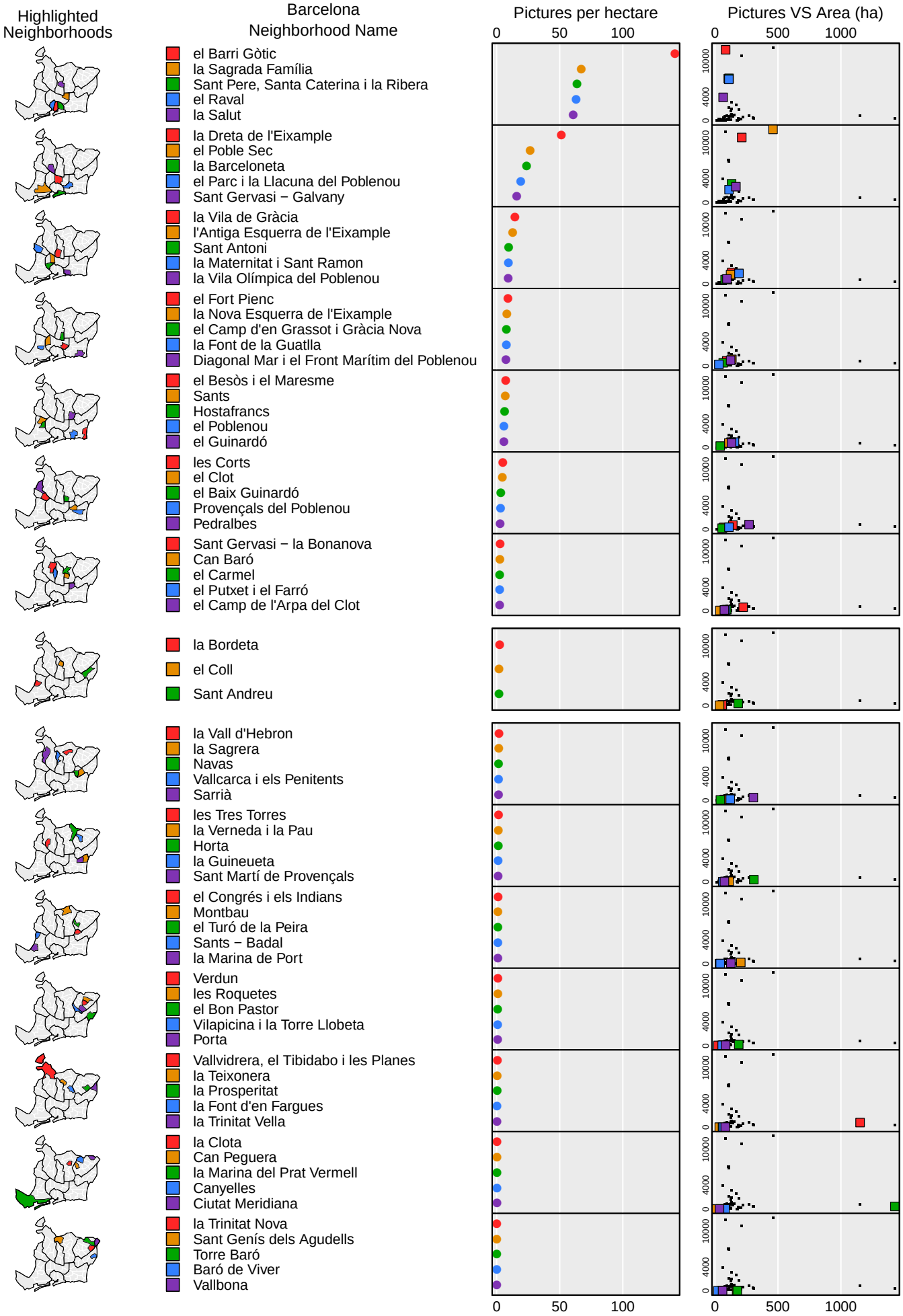


Figure 7.7: Average yearly picture intensity per gross area unit of the geotagged pictures of Barcelona retrieved from Flickr.

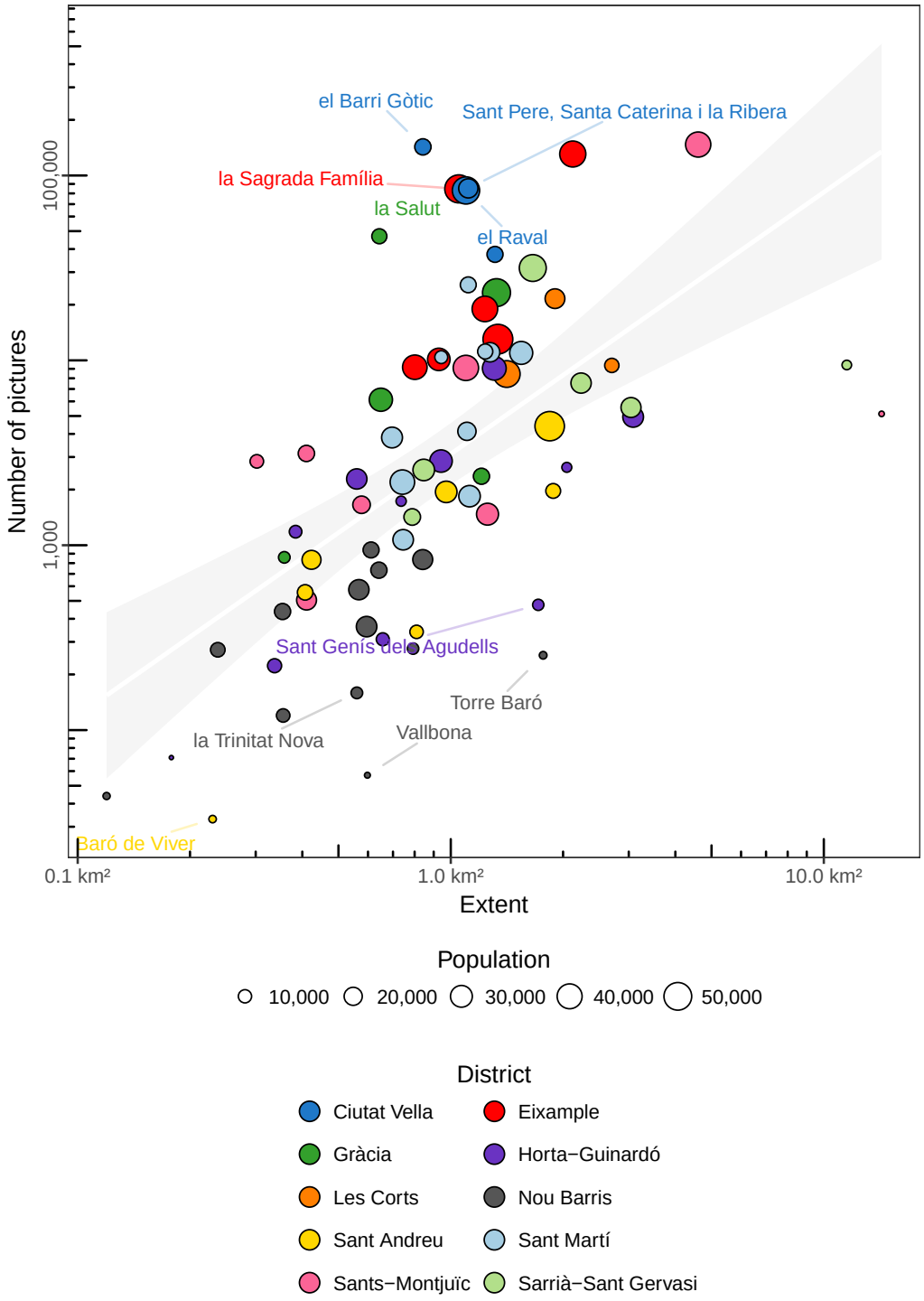


Figure 7.8: Relationship between the number of geotagged pictures of Barcelona collected from Flickr and gross area for the neighborhoods of Barcelona. Picture counts are log-10 transformed. Size is proportional to the population. The color scheme match the districts of the map in figure 7.1. The five neighborhoods with the highest density are labeled in the color of their corresponding district and the ones with the lowest densities are labeled in dark gray.

Table 7.4: Street length, street density per gross area and average yearly picture intensity per street length for the ten districts of Barcelona.

District	Street length (km)	Street density (m / ha)	Yearly picture intensity (pictures / km)
Ciutat Vella	89.49	204.85	324.64
Eixample	121.10	161.98	183.6
Gràcia	89.75	214.43	73.84
Horta-Guinardó	158.87	132.98	13.52
Les Corts	79.00	131.28	41.57
Nou Barris	125.40	155.94	3.37
Sant Andreu	109.21	166.34	7.69
Sant Martí	182.30	173.23	37.64
Sants-Montjuïc	195.14	85.06	73.02
Sarrià-Sant Gervasi	217.17	108.08	22.31

- The attractiveness of specific landmarks within the city, but also its seldom photographed areas.

The resulting linked micromap (Fig. 7.9), sorted the neighborhoods according to this average yearly intensity of pictures per network segment. According to this metric, *el Barri Gòtic* had again the most photographed streets, followed at a certain distance by two neighborhoods home to world-class landmarks in Barcelona: *la Salut* (site of Park Güell) and *la Sagrada Família* (site of the temple of the same name).

Four other neighborhoods followed (*la Dreta de l'Eixample*, *Sant Pere*, *el Raval* and *el Poble Sec*), after which the distribution began to flatten out until about the end the first upper third of the distribution, where values were indistinguishable from noise.

The linked scatter plots, with average yearly pictures in the vertical axis and kilometers of street network in the horizontal axis, shows whether a high ratio is the result of a large picture count, a short road length or a combination of both (and vice-versa in the case of a low ratio).

The same data, visualized in a scatter plot (Fig. 7.10) with the road length as the explanatory variable and the number of pictures as the dependent variable, shows a positive relationship when the picture count is plotted used a log-10 scale. In this plot, the 5 neighborhoods with the higher and lower ratios are labeled and correspond to the blocks at the top and bottom of the corresponding linked micromap (Fig. 7.9).

However, this approach does not allow visualizing these neighborhoods spatially,

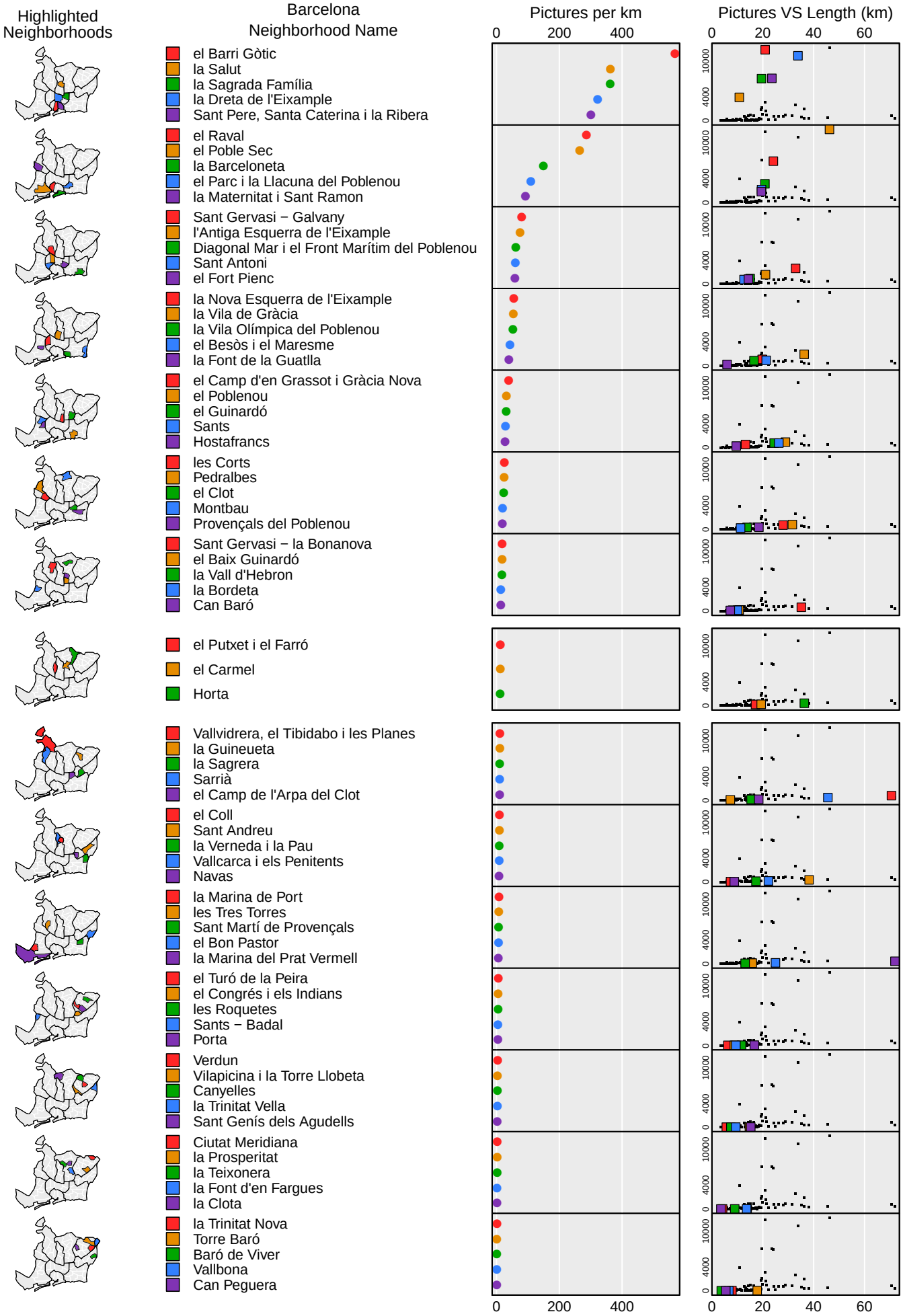


Figure 7.9: Average yearly picture intensity per street length unit of the geo-tagged pictures of Barcelona retrieved from Flickr.

for example to detect indication of clustering, the influence of external factors (e.g. proximity to the sea), the size of the spatial units as potential confound, or the connection of the quotients between both variables. In addition, it makes impossible to situate the neighborhoods for readers not familiar with the area of study.

7.3.4 Intensity per Capita

To obtain a measure of the pressure of street photography on the daily lives of the inhabitants of the city, the number of collected geotagged pictures per neighborhood was adjusted to account for the different sizes of their populations. If the majority of pictures were taken by the residents of the neighborhoods themselves, it would be expected that the ratio between the number of pictures taken and the population would be roughly constant across all neighborhoods.

Therefore, wide differences in the number of pictures taken compared to the population would indicate a inhomogeneous distribution of their attractiveness, and could be an —albeit imperfect— measure of the pressure of visitors (not necessarily tourists only) on the daily activities of the residents.

Population data was retrieved form the official bureau of statistics of the city through the Open Data BCN portal[27], which covered every year in the 2010–2016 period. The data used in the calculations correspond to the most recent data at the time of research, corresponding to year 2016.

The distribution of the population is not homogeneous across the neighborhoods, as there are for example parks without any residents. Therefore, the lack of ancillary information to support a better distribution of the population (as used in dasymetric maps to show true densities) imposes an unrealistic assumption of homogeneity, which should be somewhat mediated by the relative small size of neighborhoods.

The districts had very different number of residents, even accounting for their areas computing their population densities (Table 7.5); similarly, and the neighborhoods' population ranged from 537 inhabitants in the least populated (*la Clota*), to a population of 58,224 in the most populated (*la Nova Esquerra de l'Eixample*).

When visualizing the picture/population ratio using linked micromaps (Fig. 7.11), the old quarter (*el Barri Gòtic*) is again the most photographed neighborhood according to Flickr data, followed by *la Marina del Prat Vermell* which despite its very small population (1,138 inhabitants) manages to capture 5,137 pictures (as evidenced by the accompanying scatter plot).

[27]Population data was downloaded from *http://opendata-ajuntament.barcelona.cat/data/ca/dataset/ine-ine01* on June 2017.

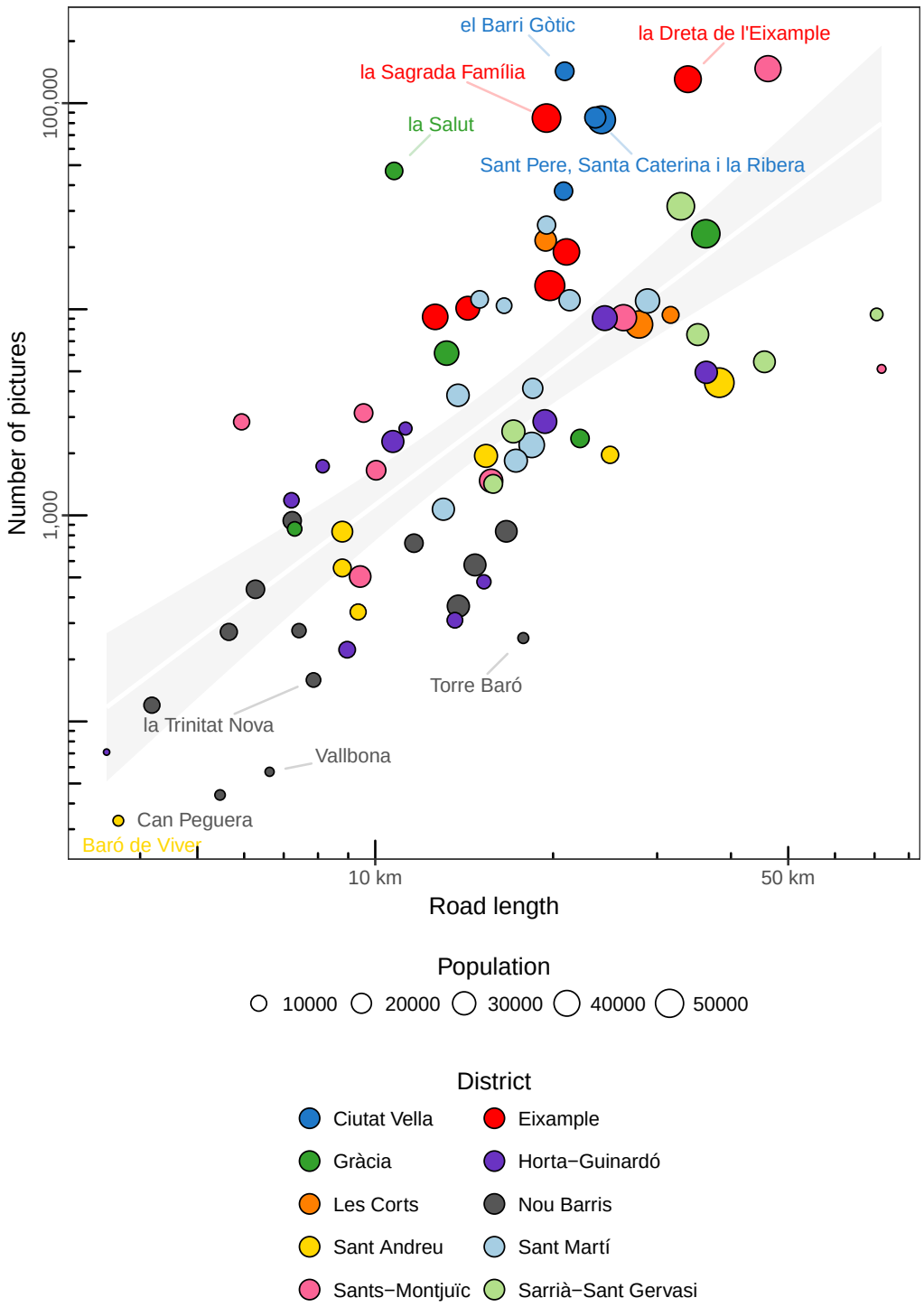


Figure 7.10: Relationship between the number of geotagged pictures of Barcelona collected from Flickr and road length for the neighborhoods of Barcelona. Picture counts are log-10 transformed. Size is proportional to the population. The color scheme match the districts of the map in figure 7.1. The five neighborhoods with the highest density are labeled in the color of their corresponding district and the ones with the lowest densities are labeled in dark gray.

Table 7.5: Population, population density and average yearly picture intensity per capita of the ten districts of Barcelona.

District	Population (inhabitants)	Population density (inhabitants / ha)	Yearly intensity (1000x inhabitants)
Ciutat Vella	102,347	234.29	283.86
Eixample	266,477	356.42	83.44
Gràcia	121,502	290.29	54.55
Horta-Guinardó	168,092	140.70	12.78
Les Corts	82,270	136.72	39.92
Nou Barris	166,310	206.82	2.54
Sant Andreu	147,732	225.02	5.69
Sant Martí	235,719	223.99	29.11
Sants-Montjuïc	183,120	79.82	77.81
Sarrià-Sant Gervasi	147912	73.61	32.75

These neighborhoods are followed —as shown by the linked dot plot— by the three neighborhoods of *Sant Pere*, *el Poble Sec* and *la Salut*, with very similar ratios to *la Marina del Prat Vermell* but with widely different populations, as evidenced by the linked

scatter plot 7.12, where they are almost aligned on a line passing through the origin.

Because the neighborhoods are sorted according to this picture per capita ratio, the linked micromaps show evidence of clustering of high values in the most popular visitor destinations near the center (as expected), and of low values in the neighborhoods in the periphery.

When visualizing the same information using a scatter plot (Fig. 7.12), the disconnection between the population and the number of pictures taken becomes more apparent, with some neighborhoods capturing a significant number of pictures, compared to what would be expected considering their population.

While approach allows visualizing simultaneously the dependent (picture count) and independent (population) variables, as well as the area of the neighborhood (dot size), the location is reduced to identifying the corresponding district as a color, instead of showing its size, shape and location on a map.

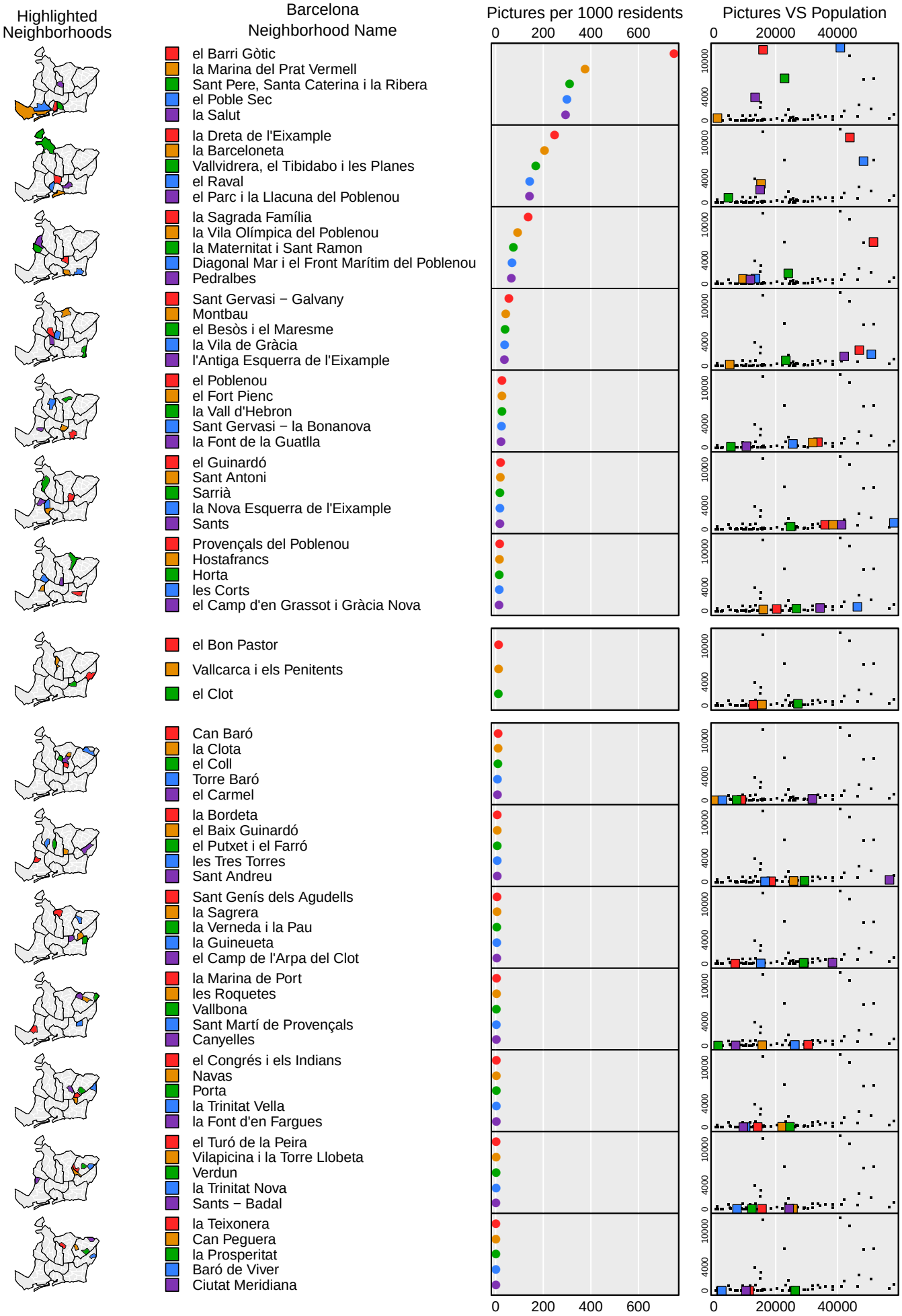


Figure 7.11: Average yearly picture intensity per capita of the geotagged pictures of Barcelona retrieved from Flickr.

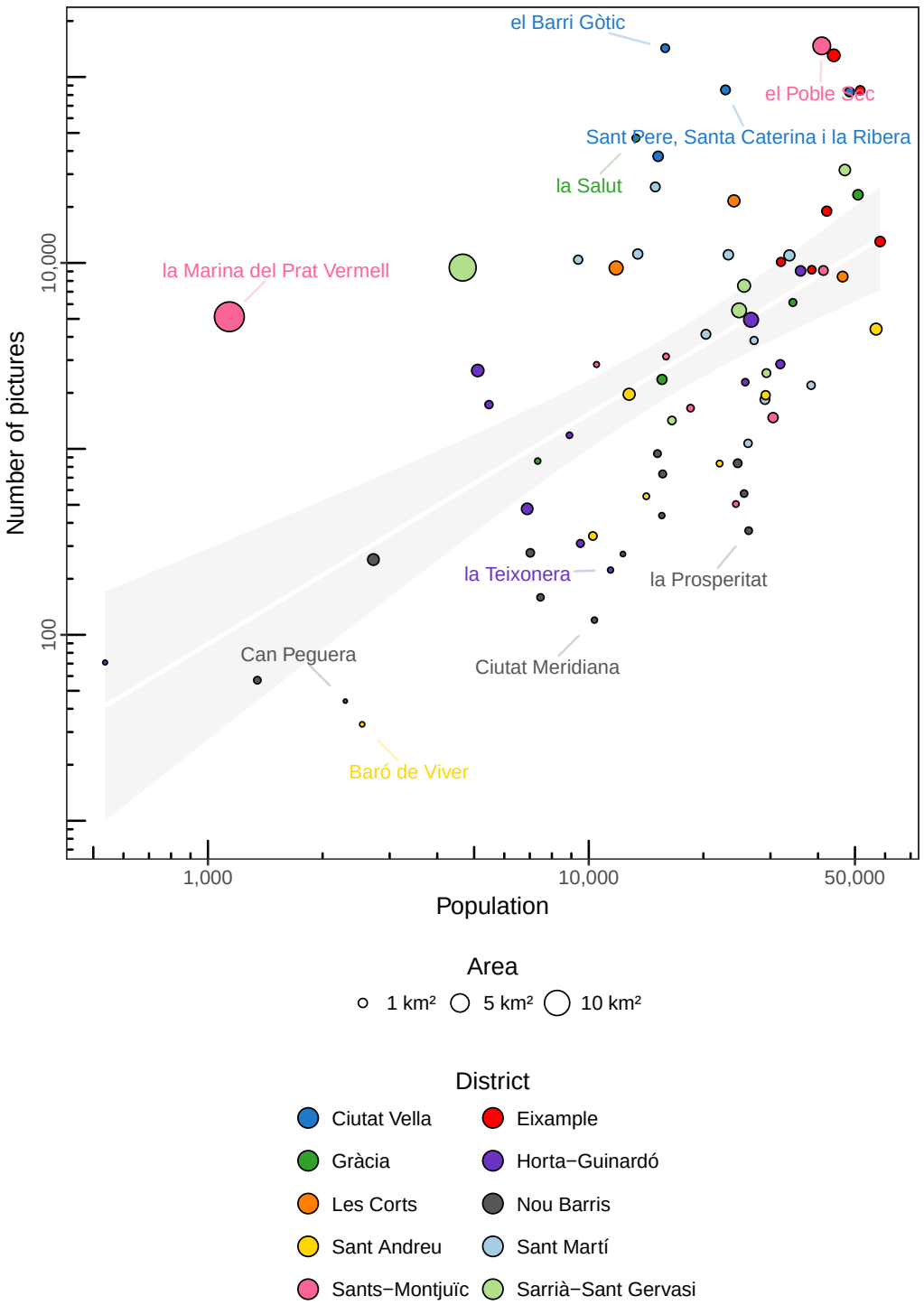


Figure 7.12: Relationship between the number of geotagged pictures of Barcelona collected from Flickr and population for the neighborhoods of Barcelona. Picture counts are log-10 transformed. Size is proportional to the gross area. The color scheme match the districts of the map in figure 7.1. The five neighborhoods with the highest density are labeled in the color of their corresponding district and the ones with the lowest densities are labeled in dark gray.

7.4 Micromaps and Temporal Data

7.4.1 Daytime and Nighttime Distribution

The daily cycle —with its alternating daytime and nighttime periods— is discussed in section 9.7, but isolated from its spatial context. As an approach to providing this context using micromaps, the variation of this cycle across the different neighborhoods was investigated. Since nighttime noise is a source of conflicts, this approach should provide and indirect measure of which neighborhoods have more nocturnal activity.

Since the hourly temporal information available was very sensitive to outliers or inaccuracies in the time stamps retrieved form the picture metadata, only the data with the highest certainty in its hourly temporal accuracy was used, as discussed in section 9.7.1.

The resulting information-dense visualization included three pieces of information in addition to the name of the neighborhood and its location in the map thumbnail (Fig. 7.13), sorted according to their seasonally adjusted daytime/nighttime ratio, using three linked chart types arranged in columns, from left to right:

- In the leftmost chart, a dot plot of the percent of the pictures that were taken during daytime, with a dotted line at the expected proportion if the process was random (50%).
- In the center chart, a diverging bar chart with the number of pictures taken per year on average during daytime (left bar) and nighttime (right bar), with the combined length of the bars corresponding to the yearly average number of pictures taken, and its horizontal position their relative proportions, where a centered bar would imply equal distribution of daytime and nighttime pictures.
- In the rightmost chart, a filled bar chart of the proportion of daytime and nighttime pictures, each seasonally adjusted according to the lengths of day and night at the time it was taken (as discussed in section 9.4).

As shown in the diverging bar chart in the center column, some neighborhoods lacked a sufficient amount of data (denoted by the combined length of both sides the bar) to reach a significant conclusion. However, this limitation did not apply to all the neighborhoods.

La Salut appears as the first neighborhood with a significant picture count that exhibits an overwhelming majority of daytime pictures, probably because its main attraction is the *Park Güell*, which remains closed during the night. *La Sagrada Famíla* exhibits a very similar pattern, arguably for the same reason in relation to the basilica that gives the neighborhood its name.

At the other end of the spectrum appears *el Parc i la Llacuna del Poblenou* as

the neighborhood with a significant picture count with the highest adjusted proportion of nighttime pictures. The presence in the neighborhood of the *Torre Glòries*, formerly known as *Torre Agbar*, is the most reasonable explanation as this building is arguably more attractive when illuminated by its more than 4500 LEDs.

7.4.2 Workdays and Weekends Distribution

Beyond the daily cycle, which for most people dictates the wake/sleep cycle, the weekly cycle is related to the convention of working and/or studying on work days (in Barcelona from Monday to Friday) and leisure activities on weekends (in Barcelona on Saturday and Sunday).

This cycle is not perfectly regular, as public (national and local) holidays often occur on workdays and on occasions —provided they fall into Tuesday or Wednesday— are extended into the closest weekend[28]. In addition, most people are on vacation in the months of July and August, and therefore the division is blurred during these months. Finally, for students the leisure calendar would include the Christmas Holidays and half June and September.

Despite these distortions, the cycle is prevalent during the duration of the year and the time stamps of the retrieved pictures allowed exploring the distribution of workdays and weekends across the different neighborhoods using a linked micromap (Fig. 7.14). The neighborhoods were sorted according to their seasonally adjusted workday/weekend ratio, with the neighborhoods with a larger proportion of weekend pictures at the top, using three linked chart types arranged in columns, from left to right:

- In the leftmost chart, a dot plot of the percent of the pictures that were taken during weekends, with a dotted line at the expected proportion if the process was random (2/7, about 29%).
- In the center chart, a diverging bar chart with the number of pictures taken per year on average during workdays (left bar) and weekends (right bar), with the combined length of the bars corresponding to the yearly average number of pictures taken, and its horizontal position their relative proportions, where a centered bar would imply equal distribution of pictures taken on workdays and weekends.
- In the rightmost chart, a filled bar chart of the proportion of pictures taken in workdays and weekends, adjusted according to the relative proportions of both groups in a typical week (as discussed in section 9.4), with a weight

[28] A long weekend is referred to as a *puente* (Castilian) or *pont* (Catalan).

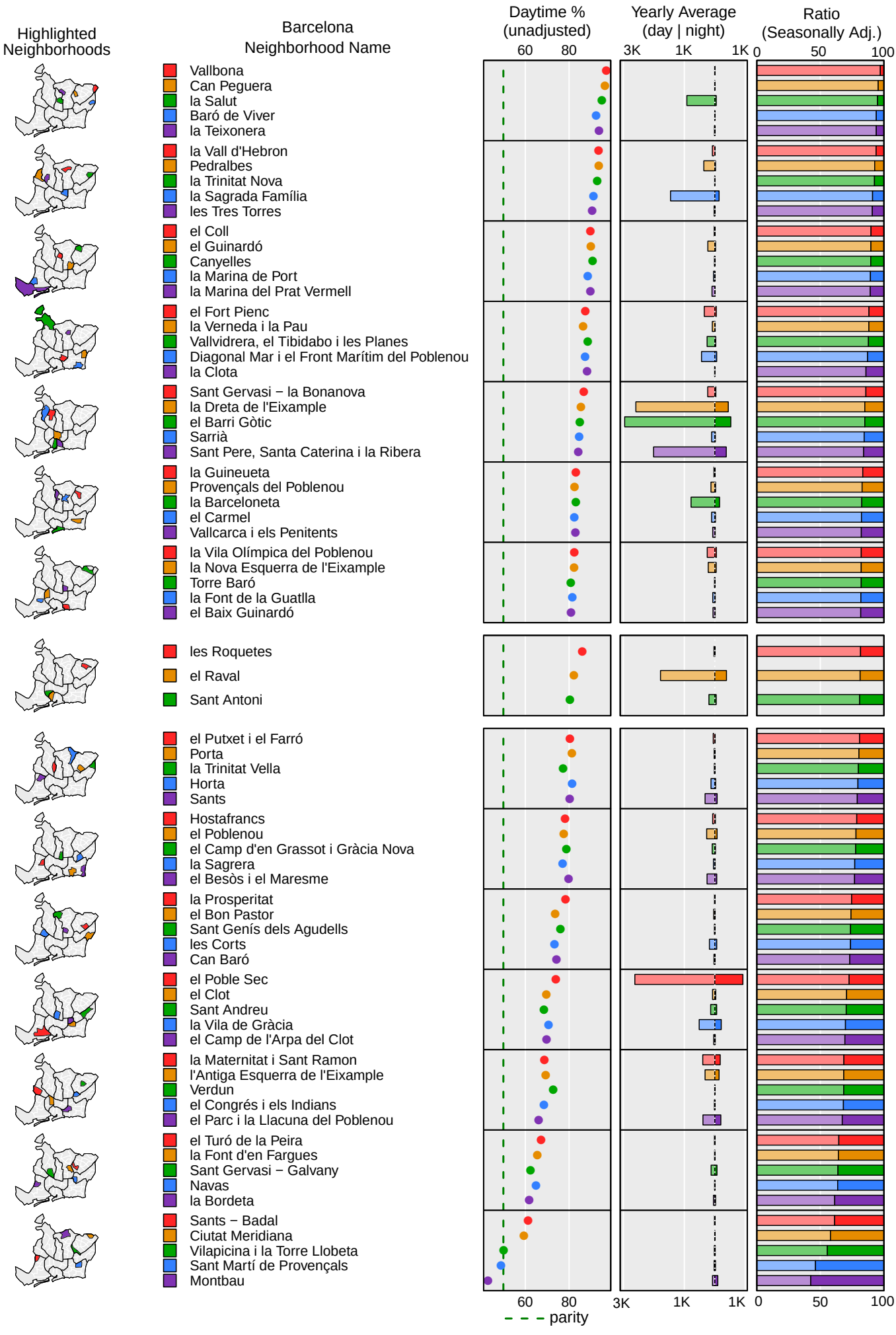


Figure 7.13: Daytime and nighttime distribution per neighborhood of the time stamps of the geotagged pictures of Barcelona retrieved from Flickr.

of 2/7 for workdays[29] and 5/7 for weekends[30] (each weighted with the probability of a random event to fall into the *other* category).

Unsurprisingly —considering all analyses of the temporal behavior within neighborhoods use the same source data—, only some of the neighborhoods had a sufficient number of geotagged pictures within their boundaries to be able to extract significant conclusions from the weekly distribution of their time stamps.

The most photographed neighborhoods appear at the bottom half of the figure, because all of them have a smaller proportion of weekend pictures than the median, arguably because their attractiveness does not change during the weekly cycle; this fact seems to be confirmed because their adjusted ratio is close to 50%.

Of the neighborhoods with a significant amount of pictures, *Sant Pere* has a large proportion of weekend pictures, probably because it contains *el Parc de la Ciutadella* and the Barcelona Zoo, while *la Sagrada Família* is at the other end of the spectrum, barely changing its pattern throughout the weekly cycle.

7.4.3 Seasonal Distribution

The longest cycle analyzed was the yearly cycle (discussed in detail in section 9.9). As a natural cycle it influences the climate along the seasons as well as the duration of daytime and nighttime, and indirectly many events that occur with a yearly periodicity, the most relevant for this research being the summer holidays.

For the analysis of the seasonal variation across neighborhoods (Fig. 7.15), the timestamps of the geotagged pictures of Barcelona retrieved from Flickr were classified into months, which in turn were classified into four seasons. The neighborhoods were ordered according to the proportion of pictures taken during the warmer seasons and included three columns of data panels, from left to right:

- In the leftmost chart, a dot plot of the percent of the pictures that were taken during the warmer seasons. This classification was informally defined as warm (summer and spring) and cold (winter and autumn) seasons. In the panels, a dotted line marked the expected proportion if the process was random (50%).
- In the center chart, a diverging bar chart with the number of pictures taken per year on average during cold seasons (left bar) and warm seasons (right bar), with the combined length of the bars corresponding to the yearly average number of pictures taken, and its horizontal position their relative proportions, where a centered bar would imply equal distribution of pictures taken on cold and warm seasons.

[29] Corresponding to 1 - 5 / 7.

[30] The sum of both weights is one.

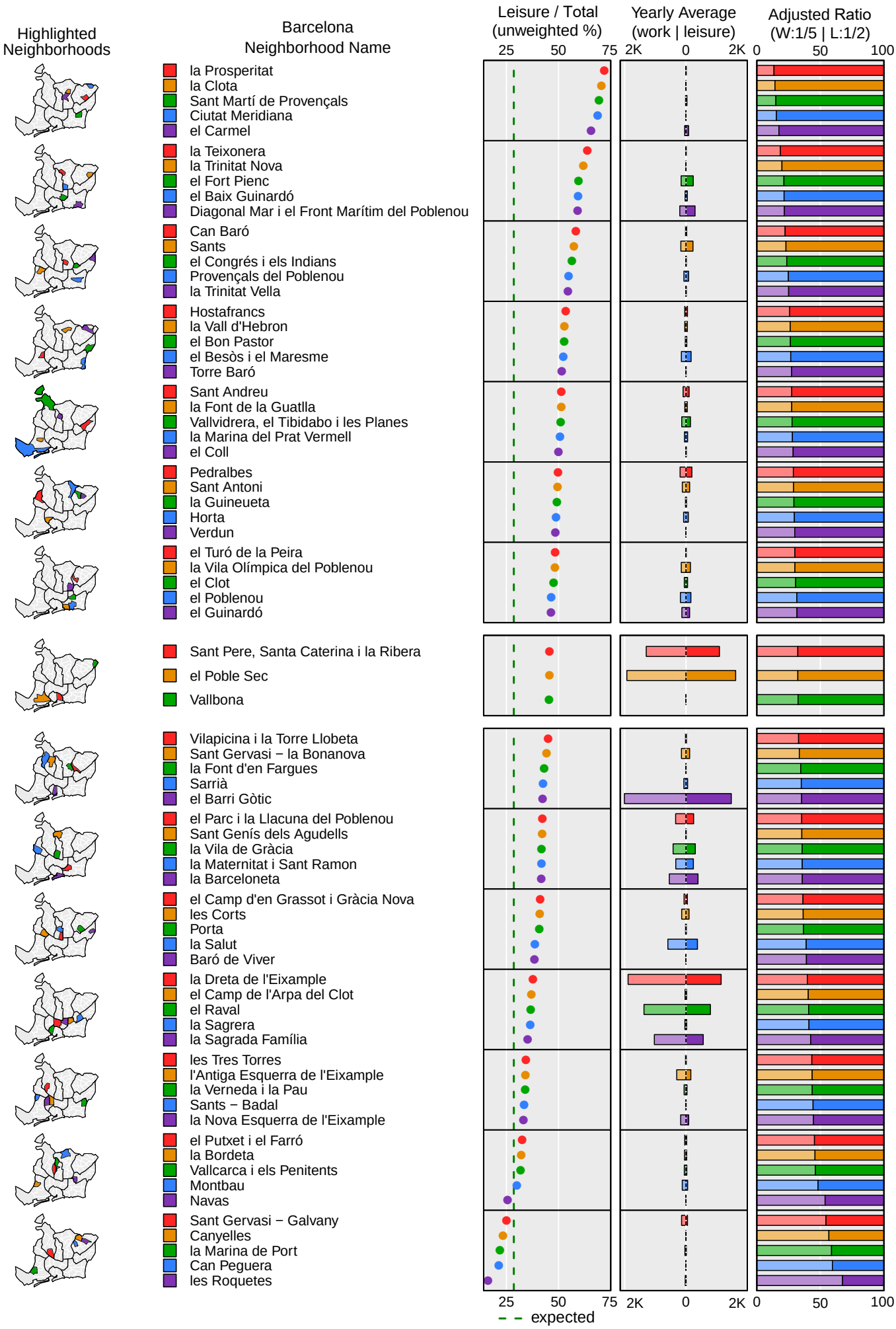


Figure 7.14: Workday and weekend distribution per neighborhood of the time stamps of the geotagged pictures of Barcelona retrieved from Flickr.

- In the rightmost chart, a filled bar chart of the proportion of pictures taken in the four meteorological seasons, where seasons correspond to groupings of three complete months: autumn (September to November), winter (December to February), spring (March to May) and summer (June to August). As the duration of all seasons was almost the same, no adjustment was necessary.

The results show that the neighborhoods with a significant number of retrieved locations are clustered around the median, with small deviations from an even distribution of pictures between cold and warm seasons.

7.4.4 Popularity Evolution

While time series are discussed in detail in section 9.5, to further investigate the suitability of micromaps to visualize sequence data, they were used to plot the evolution of the popularity of the different neighborhoods in Barcelona. In the resulting linked micromap (Fig. 7.16) the geotagged pictures collected from Flickr were visualized using two charts, along with their corresponding linked micromaps to facilitate their identification:

- The total number of retrieved pictures per neighborhood, which was also used as the ordering criteria in descending order from top to bottom, visualized as a bar chart (left column).
- The yearly proportion of pictures taken in each neighborhood as a time series, visualized as a line chart (right column).

This approach required counting the geotagged and time-stamped pictures under two grouping criteria —one spatial and one temporal—, which was stored in the two dimensions of a matrix, with the rows corresponding to the 73 neighborhoods and the columns corresponding to the 12 years covered by the dataset. In this matrix, the count for any given neighborhood in a specific year was stored in the cell at the intersection of the corresponding row (neighborhood) and column (year).

However, as the number of retrieved pictures was variable across the years (discussed in section 9.5), all the counts in the matrix were divided by the corresponding total per year across all neighborhoods (the sum all the values in its column, or column standardization) as follows:

$$p_{ny} = \frac{c_{ny}}{\sum_{n=1}^{N} c_{ny}}$$

Where the elements present in the formula are the following:

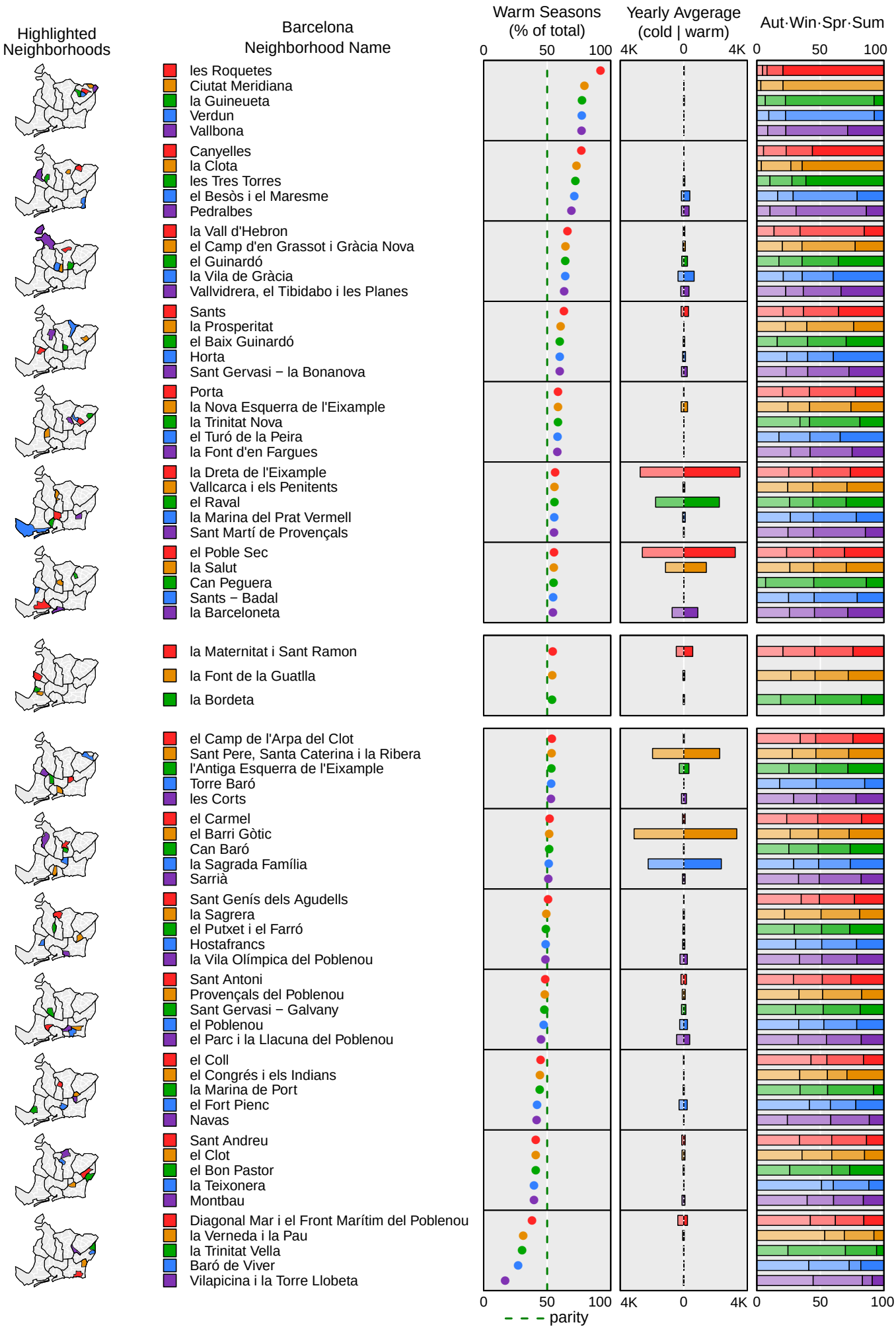


Figure 7.15: Seasonal distribution per neighborhood of the time stamps of the geotagged pictures of Barcelona retrieved from Flickr.

p is the standardized proportion of pictures, and p_{ny} is the percent of all pictures in a specific neighborhood n on year y.
c is the count of pictures collected from Flickr, and c_{ny} is the count of pictures in neighborhood n on year y.
n is the neighborhood, determined from the picture coordinates.
y is the year the picture was taken, according to its metadata.
N is the set of all n neighborhoods.

The results show that the six of the neighborhoods are overwhelmingly more photographed across the studied time period (*el Poble Sec*, *el Barri Gòtic*, *la Dreta de l'Eixample*, *Sant Pere*, *la Sagrada Família* and *el Raval*), and are clustered together, as shown in the corresponding micromap.

However, their popularity changes along the years, with *el Barri Gòtic* and *la Dreta de l'Eixample* peaking around 2006, *el Poble Sec* around 2010-2011, and *Sant Gervasi* around 2012-2013, while the temporal patterns of the rest of the neighborhoods remain relatively constant.

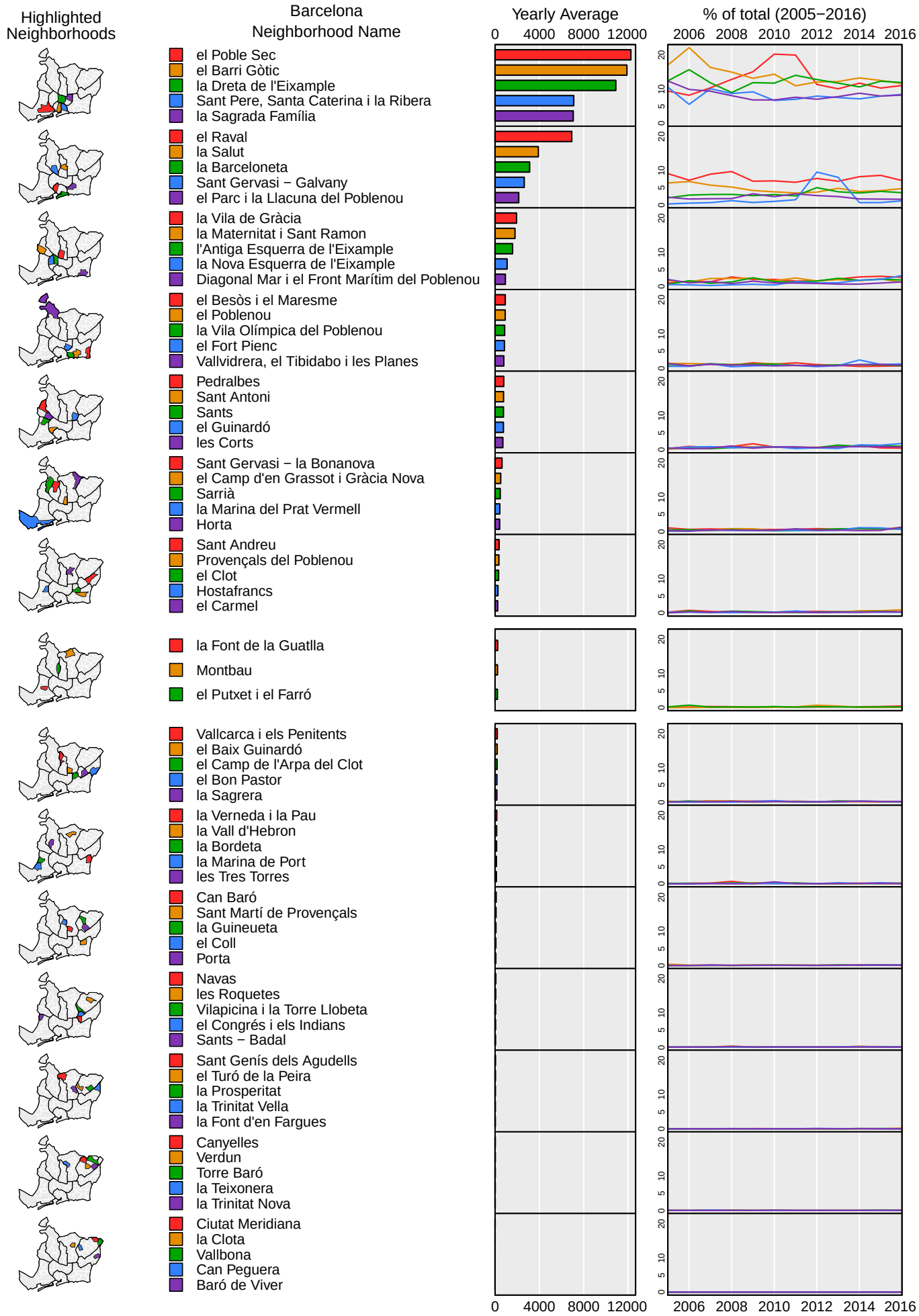


Figure 7.16: Popularity evolution per neighborhood across the analyzed 12-year period, according to the geotagged pictures of Barcelona retrieved from Flickr.

Chapter 8

A Network Approach

> "Caminante, no hay camino, se hace camino al andar"
>
> Antonio Machado

8.1 Introduction

8.1.1 Urban Networks

Cities are the most complex artifacts produced by humanity, and urban networks —transportation, communications, social, road, utilities, distribution, recycling— provide the essential support to all the interrelated activities developed within them.

Most urban indicators follow a scaling law [2], and doubling the population of a city typically requires only about a 85% increase in infrastructure [131], as efficiency increases with higher densities [359]. Human interactions increase superlinearly as well [34], and there is evidence that the scale of the cities [360] is determined by limitations imposed by the growth of networks [132].

The study of networks is quickly becoming an interdisciplinary field [361, 362], discovering mechanisms of network formation that exhibit scaling themselves [363, 364]. However, in the case of *spatial* networks, their particular properties require a distinct approach [365].

The scientific study of pedestrian flows pioneered in the decade of 1970 by Pushkarev [203], was followed by attempts to conceptualize the street network from its geometric configuration [366], perhaps the most successful being the space syntax perspective [367, 368], an approach not exempt from criticism [369].

Other avenues of research focused on modeling the navigation of pedestrians as a choice [370] between shortest or simplest paths [371], while different approaches favored the study of the evolution of street networks themselves [372] and model them [373] or quantify their scaling laws [374].

8.1.2 Chapter Outline

This chapter focuses on street networks and the influence of the activities of persons that share these public spaces, which in Barcelona face the challenge of making compatible activities such as bar terraces [375], motorbikes parked on the sidewalk, bicycles circulating and pedestrians performing their daily activities, along with the increasing presence of tourists enjoying the city.

The study of these patterns [376] and their representation[1] [377] should should provide knowledge to plan better pedestrian networks [378] and designing streets more adequate to the needs of their users [379]. This chapter proposes a multi-faceted approach to the analysis and visualization of geotagged data on an urban street network:

- Section 8.2 establishes the necessity of accounting for the heterogeneity of the street network in urban spatial analysis, and demonstrates its variation across Barcelona.
- Section 8.3 discusses the methodologies developed to count the location density per segment, and the method to determine the streets with most locations.
- Section 8.4 explains the results of the computation of the linear density of the segments in the street graph of Barcelona, applied to the collected data from Panoramio, Flickr and Twitter as cases of study.
- Section 8.5 introduces the network-constrained kernel density estimation as an alternative to the computation of the density per street segment in detailed analyses, and presents the results for the geotagged pictures collected from Panoramio and Flickr.

8.2 Heterogeneity of the Urban Fabric

8.2.1 Network and Address Density

The road network structure of Barcelona has been determined by its history, with a densely packed old quarter (Ciutat Vella), surrounded by the regular pattern

[1]The *cf. city flows* website is available at *http://uclab.fh-potsdam.de/cf/* at the time of writing.

of the Eixample, which extends until it reaches the former towns that existed around Barcelona, such as Gràcia or Sant Andreu (Fig. 7.1).

When analyzing events that are primarily occur within this network —as is the case of the studied geotagged picture locations or messages—, this rich and varied urban fabric can bias the results, since it is more likely to take a picture of a street if the road network is more extensive, all other things being equal.

To determine the necessity of correcting this bias, two elements of the road network were analyzed to determine their homogeneity —or lack of thereof— and the establish the necessary corrections in the developed methodology, which were also taken into consideration in the results discussed in section 7.3.3 in the context of the different neighborhoods:

- The density of the road network, expressed as the length of the road network per area unit (section 8.2.3).
- The distance from the entrances of the buildings to the axis of their corresponding street segment (section 8.2.4).

8.2.2 Source Data

The cartographic data was retrieved from the official sources at CartoBCN[2], the cartography portal of the municipality of Barcelona, on November 17, 2016. Of the available cartography, two data sets were used in this chapter:

Street network consisting in a zip file[3] which contained two shapefiles using the EPSG:25831 coordinate reference system: the segments[4] corresponding to the road axes (14,821 records) and the nodes[5] corresponding of the road intersections (9,484 records). According to the metadata, this dataset was produced by the in-house cartographic services (*Departament del Pla de la Ciutat*) on March 15, 2011. The spatial heterogeneity of the street network is discussed in section 8.2.3.

Postal addresses consisting in a zip file[6] which contained four shapefiles using the EPSG:25831 coordinate reference system. One of these shapefiles contained the postal addresses[7] of all the parcels (144,957 records) corresponding to the entrances to the buildings. According to the metadata, this dataset was produced by the in-house cartographic services (*Departament*

[2]CartoBCN data is available at no cost (upon registration) at *http://w20.bcn.cat/cartobcn/* at the time of writing.
[3]The compressed street network data was named "BCN_GrafVial_ETRS89_SHP.ZIP".
[4]The road segments shapefile was named "BCN_GrafVial_Trams_ETRS89_SHP".
[5]The node shapefile was named "BCN_GrafVial_Nodes_ETRS89_SHP".
[6]The compressed address data was named "BCN_Adreces_ETRS89_SHP.ZIP".
[7]The postal addresses shapefile was named "BCN_Adreces_CarrerNum_ETRS89_SHP".

del Pla de la Ciutat) on June 9, 2011. The spatial heterogeneity of the postal addresses is discussed in section 8.2.4.

8.2.3 Street Network Density

The density of the street network per area unit can only be approximated, because it varies continuously across the area within the city limits of Barcelona. It was therefore necessary to perform the computation within small aggregating polygons but large enough to provide meaningful results.

The approach, inspired in focal statistics applied to spatial analysis, discretized the area of study into a lattice of hexagons whose centers were 266.6 meters apart, corresponding to the canonical axis separation between the streets in the dominant urban fabric in Barcelona, the Eixample central district.

Within each tile of the hexagon lattice, the density of the street network per area unit was computed as the sum of the lengths of all portions of line segment (street axes, measured in meters) that lied inside, divided by the corresponding area covered by each tile[8] (measured in hectares).

The actual area occupied by the street network was not taken into account because it was not possible to separate areas destined to vehicles from areas destined to pedestrians with the cartography available, and street length was considered a better indicator of the overall perception from the point of view of the pedestrians [6].

The results (Fig. 8.1) showed the variation of the network density within Barcelona, which was higher (as expected) in the Barceloneta neighborhood and within the former city walls (old quarter), but also in the villages incorporated into Barcelona at the end of the 19th century, while the Eixample displayed more uniformity as a consequence of its regular pattern.

In the map, the density is color-coded according to a perceptually uniform color ramp, showing the differences between the most dense areas (around 500 m/ha) compared to the Eixample (with a theoretical density of 150 m/ha). However, while the density map provided a starting point to understand the variability of the street network across the city, a more rigorous approach was necessary to understand the observed spatial pattern, using the same approach discussed in section 6.3.

The local Getis-Ord G* cluster map (Fig. 8.2) provided a clearer representation of the variation of the road density across the city. The methodology provided a pseudo p-value —computed using a Monte Carlo method—, which was classified into three different statistical significance thresholds (0.05, 0.01 and 0.001),

[8]Except the tiles that lied in the perimeter and were clipped, all hexagons were the same size.

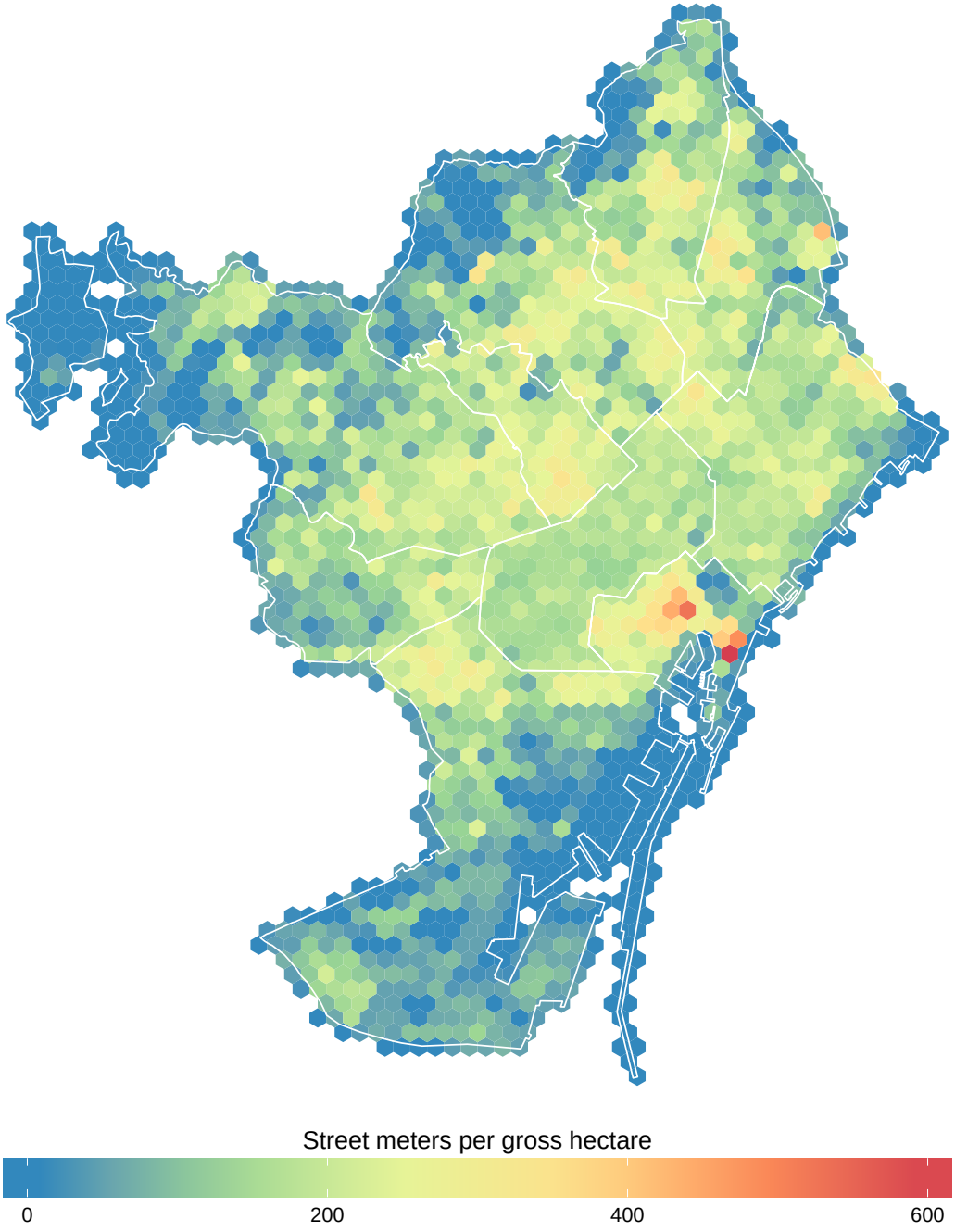


Figure 8.1: Road density variation within the limits of Barcelona, computed within a lattice of regular hexagons. The 10 districts of Barcelona are overlaid in white for reference. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. Road network and district cartography from CartoBCN, under CC BY 3.0.

represented as saturation changes in the hues corresponding to high (red) or low (blue) density clustering.

The resulting map shows the clustering of high street density areas in the old quarter and the former villages around Barcelona, while low density clusters appear near the Collserola mountain range, the harbor and the Zona Franca industrial area. Almost all the Eixample has a non-significant road density.

8.2.4 Postal Address Density

As an additional measure of the heterogeneity of the urban fabric of Barcelona, the shortest distance of the 144,957 postal addresses to their corresponding road axis within the 14,821 road segments of the official cartography was computed (further discussed in section 8.3.3).

The map of the computed distances (Fig. 8.3), classified into five categories corresponding to the partition of the distance values into five quintiles —each category with 20% of all addresses—, showed again the overall higher densities of the older urban fabrics, but also the narrow streets of the sparsely populated neighborhoods Vallvidrera and Torre Baró. According to the classification, the addresses on the streets of the Eixample appear in light blue, while on the wider streets[9] appear in dark blue.

8.3 Segment Aggregation Methodologies

8.3.1 Limitations of a Distance-Based Methodology

Using a "brute force" approach, the points were aggregated according to their distance to the set of road segments in the street network (or in other words, the number of points that were closer to a specific road segment in the street network than to any other segment were counted for each segment). The methodology consisted in the following steps:

1. The distance to the closest feature was computed and the feature identified, computing the euclidean distances[10] instead of the arguably more accurate network distances, but was considered appropriate in the case of an urban setting.

[9] Avinguda Diagonal, Gran Via de les Corts Catalanes, Passeig de Gràcia, Passeig de Sant Joan, Carrer de la Marina, Avinguda Meridiana, Avinguda del Paral·lel, Via Laietana, Carrer d'Aragó and Avinguda de Roma.

[10] Sometimes referred to with the idoms "as the crow flies" or "in a beeline".

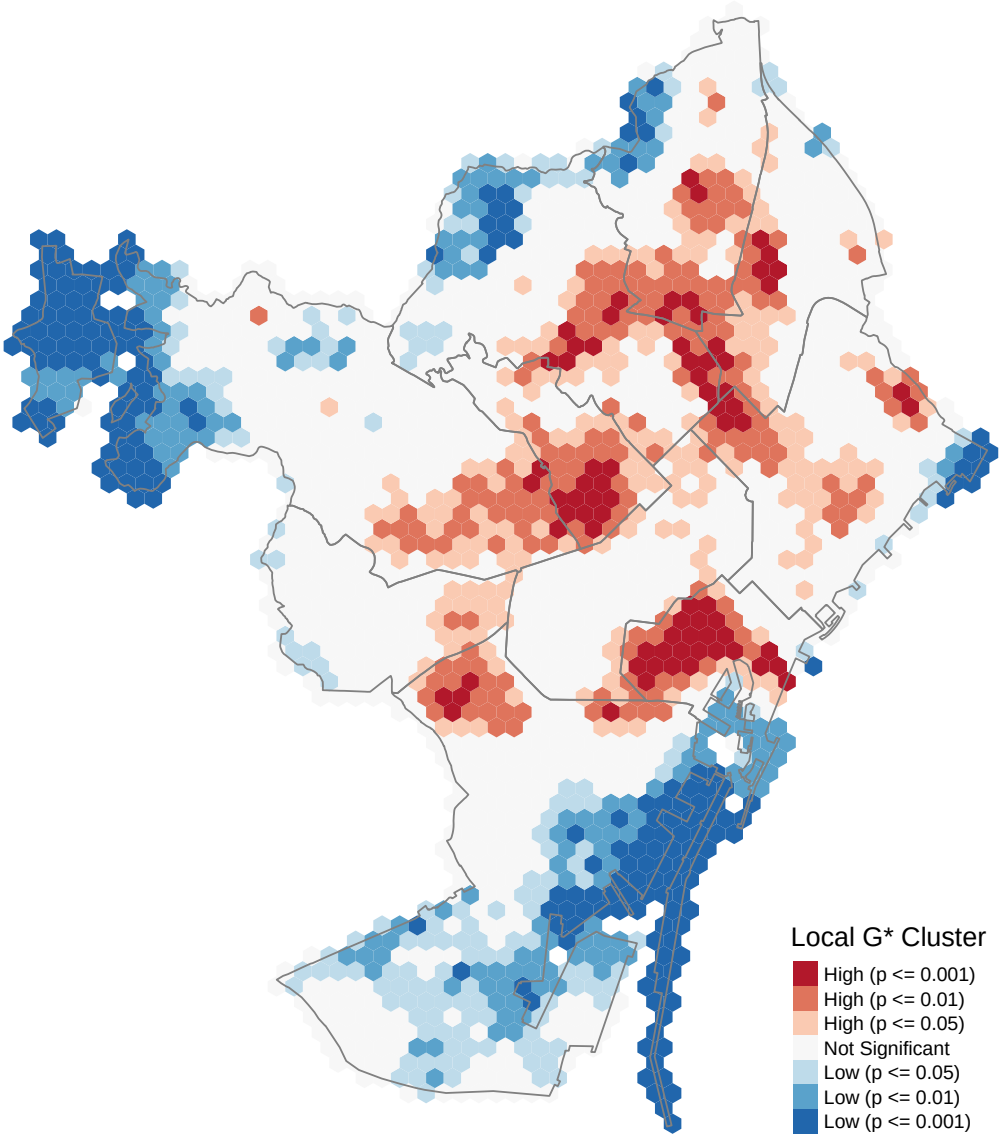


Figure 8.2: Local Getis-Ord G* cluster map of the road density within the limits of Barcelona, computed within a lattice of regular hexagons. High density clusters are shown in shades of red and low density clusters in shades of blue. Higher color saturation corresponds to higher significance, and gray colors to not significant pseudo p-values. The 10 districts of Barcelona are overlaid in gray for reference. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. Road network and district cartography from CartoBCN, under CC BY 3.0.

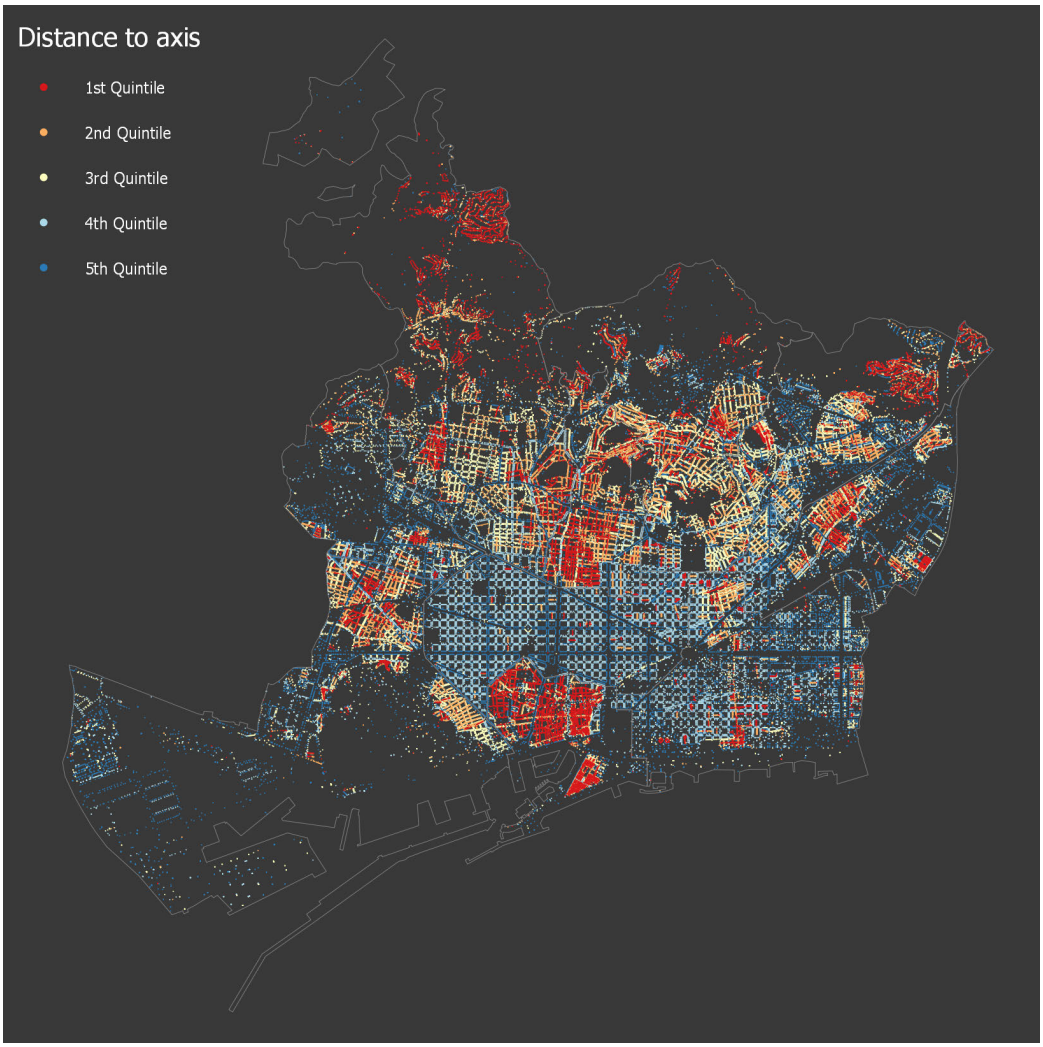


Figure 8.3: Shortest distance of all postal addresses in Barcelonato their corresponding street axes, classified into five quintiles. Map is rotated 44.4° clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. Road network, address and city limits cartography from CartoBCN, under CC BY 3.0.

2. The points were aggregated by road segment, excluding the points that were more than 50 m (the width of Avinguda Diagonal, the widest street in Barcelona) away from any segment.
3. Because the length of segments was variable, the number of pictures of each segment was divided by its length to obtain a measure of linear density, and the result was stored as an attribute for representation.

The process in the first step was computationally expensive because it was necessary to generate a gigantic distance matrix of potentially 1,166,704 x 144,957 elements (almost 85 trillion elements, easily overflowing system memory), each involving a complex square root operation.

Fortunately, the operation was embarrassingly parallel, since the distance of a subset of points to the full set of lines could be partitioned arbitrarily and the partial results joined afterwards, because the distance calculation from any point was independent from all other distance calculations.

Taking advantage of this property, the data was processed in batches[11] for the distance calculation. This approach allowed executing the distance calculation, which would have been otherwise intractable because of memory allocation issues. Despite this, the custom script developed took almost 7 hours for the larger dataset (Flickr geotagged pictures), in a modern Intel Core i7 processor and using the state-of-the-art GEOS C++ libraries (through the rgeos 0.3-23 R package) for the distance calculations. It was therefore necessary to develop a faster methodology for the approach to be practical.

8.3.2 Optimized Address-Based Methodology

An alternative methodology was developed using the SQLite[12] embedded relational database engine, with the spatialite[13] OGC[14] compliant spatial extension. The methodology reduced the number of candidates that had to be considered for the nearest distance calculation significantly, accelerating the aggregation of events into segments process, in the following stages:

1. Computation the road segment each address belonged to (discussed in section 8.3.3).
2. Aggregation of the geotagged picture locations into the nearest address, within a threshold distance (discussed in section 8.3.4).

[11] In this case, batches of 1,000 points each.
[12] The SQLite project is available at *http://sqlite.org/* at the time of writing.
[13] The spatialite project is available at *http://www.gaia-gis.it/gaia-sins/* at the time of writing.
[14] The Open Geospatial Consortium (OGC) website is available at *http://www.opengeospatial.org/* at the time of writing.

3. Transfer the sum of aggregated pictures from the addresses to the corresponding segment, and compute the linear density with the method discussed in section 8.3.1.

8.3.3 Linking Addresses to Street Segments

All segments had an official unique identifier assigned by the cartography producer, maintained and approved by the municipal authorities[15], as well as the codes corresponding to the streets at either side —left and right— of the segment, although for the majority of segments both left and right codes were the same.

In contrast, the addresses (points) only had the street code as an attribute —using the same codification as the left and right street codes of each segment—, but lacked the segment code. It was therefore necessary to develop a method to match a single segment to each address, among all the segments of the corresponding street.

Programmed in SQL using the spatial extensions that provide the required *ST_Distance* function to return the cartesian minimum distance between two projected geometries[16], the result (Fig. 8.4) paired the addresses to their corresponding segments using a computationally efficient process (about 15 seconds):

1. Pairing each address with all the segments that share the same street code in either the left or the right side.
2. Computing the distances between each address and their corresponding subset of segments, instead of all the possible combinations, an almost three orders of magnitude[17] reduction.
3. Assigning as an attribute the closest segment identifier to each address, effectively linking addresses to segments.

8.3.4 Linking Locations to Addresses

After matching addresses to their corresponding street segments, the next step required linking the geotagged locations to their nearest addresses —which acted as "anchors" or "magnets"—, allowing the aggregation of picture counts per street segment through them.

However, this operation was computationally expensive because it involved calculating the distance between two sets of points with a large number of elements

[15] According to the metadata in the file "ETS89_Info_BCN_Adreces_CA.pdf".

[16] Details on the *ST_Distance* function are available at *http://www.gaia-gis.it/gaia-sins/spatialite-sql-4.3.0.html* at the time of writing.

[17] A total of 2,942,423 (1,474,908 + 1,467,515) pairs instead of 2,148,407,697 (144,957 * 14,821).

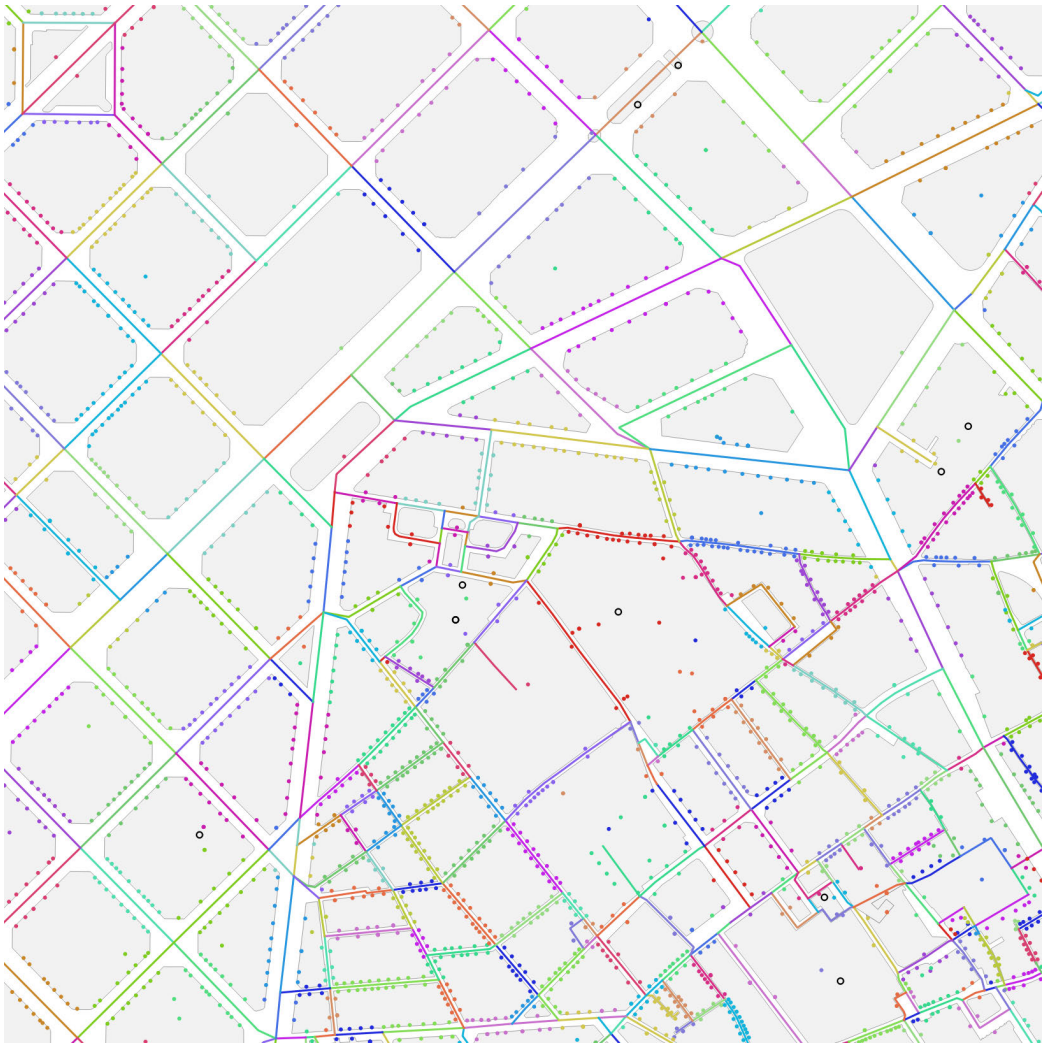


Figure 8.4: Addresses matched to the closest street segment that shares the same street code (left or right sides). Addresses are colored with the same color as the segment they are assigned to. Unmatched addresses appear as black empty circles. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. Road network and city block cartography from CartoBCN, under CC BY 3.0.

each, despite being a much simpler operation than the calculation of the distance between a set of points and a set of line segments.

Fortunately, spatially enabled databases allow using a R-tree, a data structure designed to speed up spatial queries through the use of a spatial index. Although its implementation in spatialite does not work automatically like in other spatial databases such as PostGIS[18], it is possible to manually implement an effective algorithm manually using SQL.

An added advantage of manually specifying the spatial query was that radius parameter in the function *BuildCircleMbr* indirectly restricted the search radius around the addresses[19], effectively discarding the locations outside it.

The custom query finished execution under 6 seconds while the "naive" or "brute-force" approach (discussed in section 8.3.1) with equivalent functionality required 52 minutes (about a 520-fold speedup).

Beyond the computation speedup advantages, this approach allowed to successfully link pictures to street segments even when the streets had very different widths (Fig. 8.5), as the methodology linked each location to the corresponding address, which acted as a proxy to relate the location to the corresponding street segment[20].

8.3.5 Methodology Results Comparison

The results of both approaches were very similar, and when comparing the resulting maps very small differences can be observed (Fig. 8.6), especially in zones where the address distribution is sparse and therefore it is not possible to use the address locations as an "magnet" to capture the surrounding geotagged pictures.

Therefore, the address-based approach was preferred in detail maps of areas with abundance of addresses, while the more computationally expensive distance-only approach was used for overview maps, which covered areas with uneven address densities.

Nevertheless, both approaches were in many cases superseded by another approach (discussed in section 8.5) which, while being even harder to compute, produced more accurate results.

Each methodology resulted in different number of geotagged pictures assigned to the street segments in Barcelona (Table 8.1), in part because they used different

[18]The PostGIS project is available at *http://postgis.net/* at the time of writing.

[19]Details on the *BuildCircleMbr* function are available at *http://www.gaia-gis.it/gaia-sins/spatialite-sql-4.3.0.html* at the time of writing.

[20]In practical terms, the addresses partitioned the plane in a Voronoi tessellation.

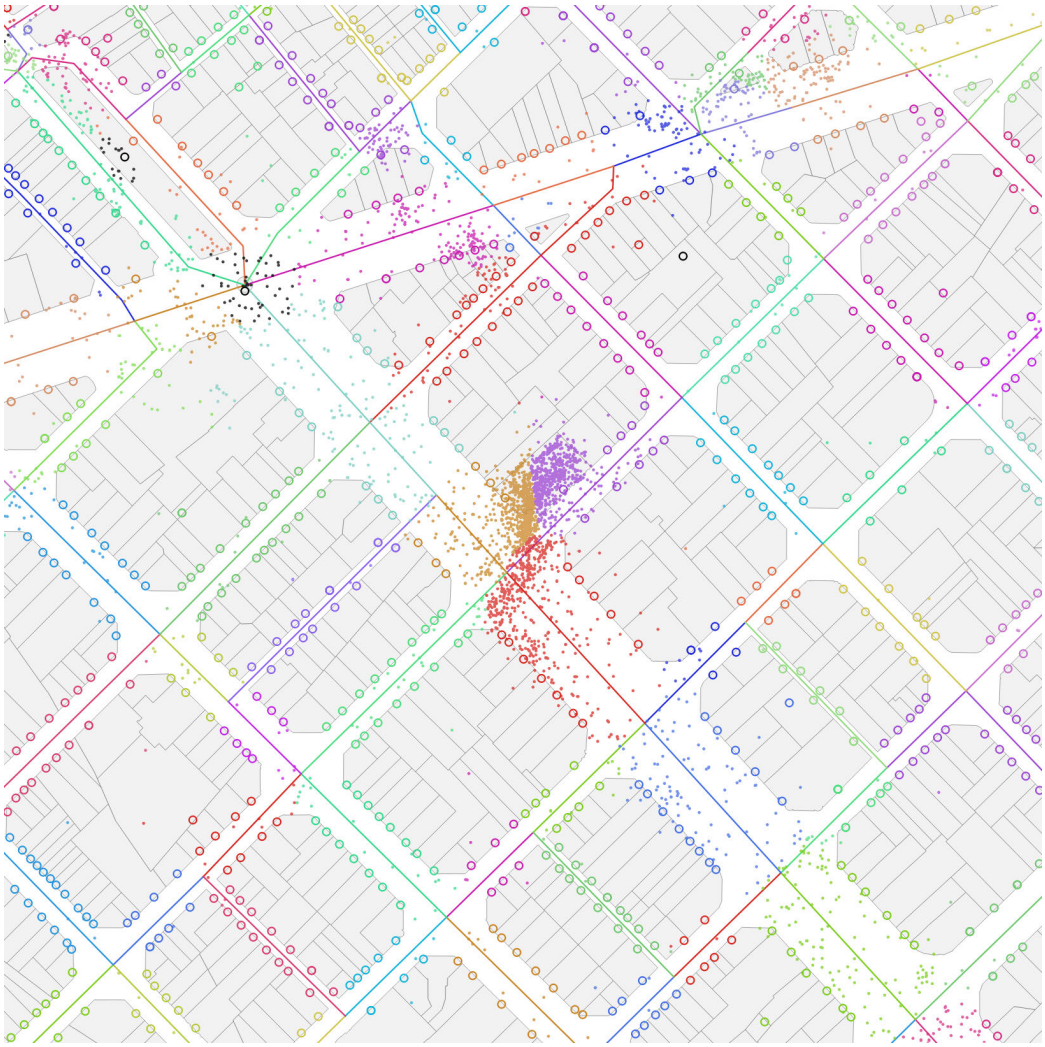


Figure 8.5: Picture locations anchored through their nearest address to the corresponding street segment. Picture locations and addresses are colored with the same color as the segment they are assigned to according to the methodology. Unmatched pictures and addresses appear in black. Map uses the ETRS89 / UTM zone 31N (EPSG:25831) projection. Road network and parcel cartography from CartoBCN, under CC BY 3.0.

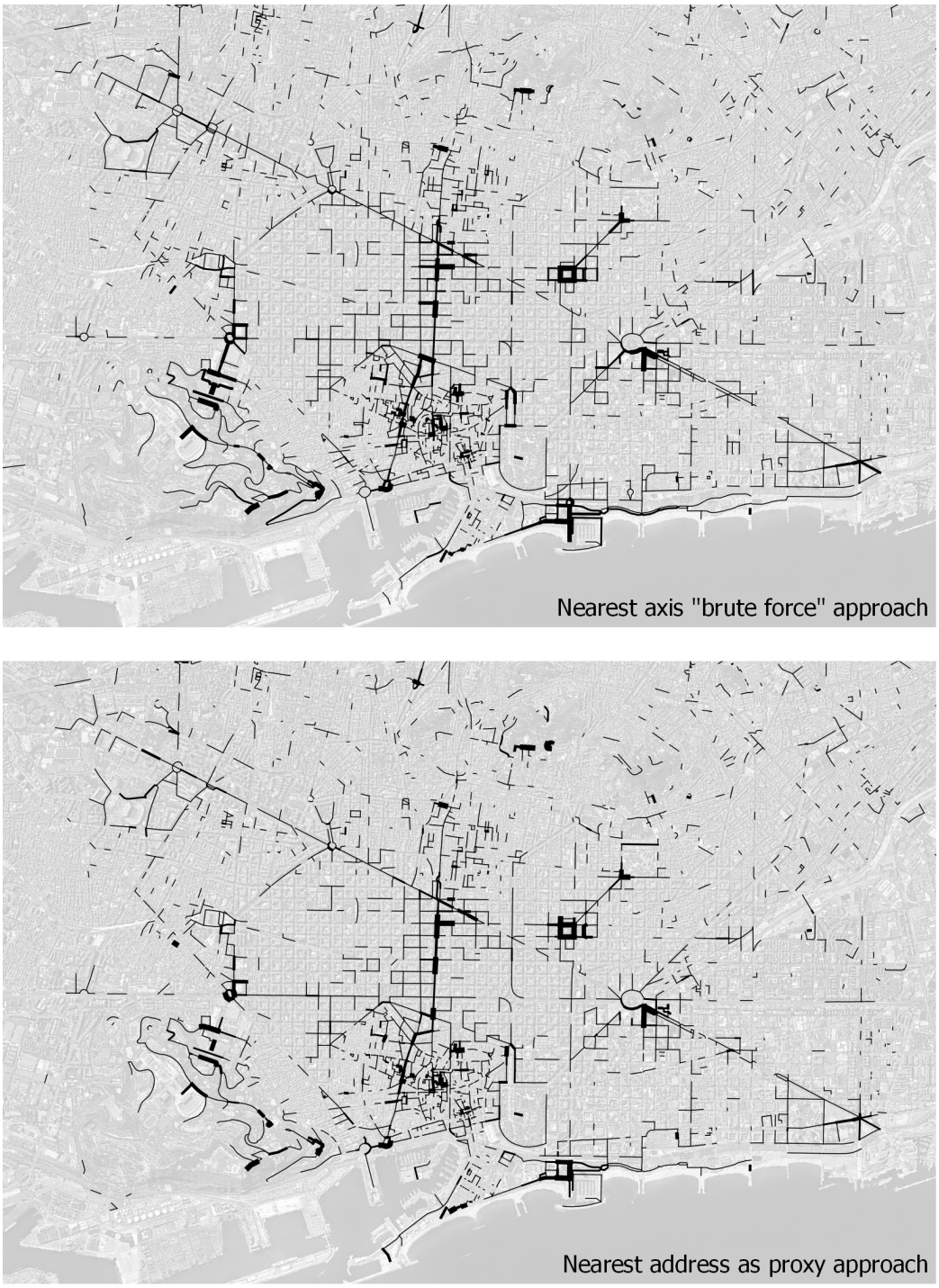


Figure 8.6: Comparison of the results of address-based and distance-only approaches on location counts per segment of the geotagged pictures collected from Panoramio, using the same classification criteria. Map is rotated 44.4º clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. Map scale is 1:75,000. Road network cartography from CartoBCN, under CC BY 3.0, ortophoto from Institut Cartogràfic i Geològic de Catalunya (ICGC), under CC BY 4.0.

Table 8.1: Number of geotagged locations from Panoramio, Flickr and Twitter assigned to street segments, according to the methodologies discussed. Note that Instagram data was dropped because it contained an insufficient number of samples.

Source	Retrieved records	Nearest segment	Address proxy
Panoramio	80,459	57,052	55,964
Flickr	1,166,704	852,950	913,299
Twitter	519,664	3794,83	388,419

strategies to reject locations too far away from the street network, but the differences in number were not significant and were generally concentrated in areas with fewer addresses.

8.3.6 Determining the Most Photographed Streets

Since the street segment data did not include the name of the street, but the names of the street in the left and right sides of the road axis instead (although they were coincident in 96% of the segments), the following methodology was used to count the number of locations corresponding to each street name (tables 8.3, 8.5 and 8.7):

1. The number of locations on each segment were split evenly between its left and right sides, regardless of whether they shared the same street name.
2. A list was compiled from the street or streets any segment belonged to, as the union of three possible cases:
    - The left side was the same as the right side.
    - The left side when the sides were not the same.
    - The right side when the sides were not the same.
3. The corresponding length was computed from the geometry to obtain a density per linear unit.
4. The pictures were summarized and assigned to the corresponding street according to the previous list.

However, multiple factors could make the results of this metric misleading, and it was therefore necessary to develop another metric that reduced these issues, while allowing the identification of the most visited streets. In this case, the computed density (pictures per unit of street length) allowed this classification (tables 8.4, 8.6 and 8.8), although streets (or squares) with lengths of less than 50 meters were excluded to avoid distorting the densities with the presence of small denominators in the quotient. Some of the identified issues were the following:

- Some streets had a high picture count because their extension allowed collecting more pictures (e.g. Avinguda Diagonal).
- Some streets had a single segment with a unusually high number of pictures (e.g. Carrer Provença in front of La Pedrera and Carrer Marina next to both La Sagrada Família and Port Olímpic).
- Some streets were very short but next to the access to an important landmark (e.g. Carrer Olot and the entrance to Park Güell).

However, in contrast with maps, count and density metrics were heavily dependent on some aspects of the street network that made them unsuitable as a general overview of the city:

- Streets can change names arbitrarily while being conceptually the same.
- The street length distribution is very variable.
- Density can be very variable along the length of a street.
- The density metric is sensitive to the street length in the denominator.
- Locating streets by name is difficult if the city is not well known by the intended audience.

8.4 Linear Density per Street Segment

8.4.1 Overview

The individual geotagged locations collected from three different networks —Panoramio (discussed in section 8.4.2), Flickr (discussed in section 8.4.3) and Twitter (discussed in section 8.4.4)— were assigned to their nearest road segment using two different methodologies (section 8.3):

- Searching the nearest road segment within a 50 m radius (discussed in section 8.3.1). This approach was used to produce the corresponding overview maps for each source, centered on the Eixample.
- Searching the nearest address within a 50 m radius, which was used as an "anchor" to link the location to the corresponding road segment (discussed in section 8.3.4). This approach was used to produce detail maps of the old quarter (Ciutat Vella) for each source.

Regardless of the methodology, the picture counts were divided by the length of the corresponding segment to obtain a linear density. As the distributions of these linear densities were very skewed, all results were log-transformed [380]. Finally, the values were standardized as the z-score permitted comparisons across very different sources. Values with z-scores below were dropped and values beyond 10 standard deviations from the mean were placed in the same "overflow" category (Table 8.2).

Table 8.2: Raw and transformed ranges per street segment for the three sources analyzed.

Source	Non-zero	Range	Density (d)	log(d + 1)	z-score
Panoramio	42.7 %	0 – 1,716	0 – 24.90	0 – 3.25	-0.28 – 29.78
Flickr	72.2 %	0 – 34,156	0 – 282.39	0 – 5.65	-0.50 – 10.90
Twitter	47.8 %	0 – 77,033	0 – 1,117.82	0 – 7.02	-0.36 – 22.18

To visualize the spatial distribution of the results, two maps were produced for each of the three sources: an overview map (figures 8.7, 8.9 and 8.4.4) and a detail map (figures 8.8, 8.10 and 8.12), showing prominent features that could informally be classified as follows:

- Street segments that are connected along their axes and therefore can be read as a single street (e.g. Passeig de Gràcia, Les Rambles, Avinguda de la Reina Maria Cristina, Avinguda Gaudí).
- High local density of some street segments in a small area, much higher than their neighbors, corresponding to specific landmarks (e.g. Sagrada Família, Park Güell, Torre Glòries, Monument a Colón, Arc de Triomf).
- Connected networks of streets around specific streets (e.g. parallel streets of Passeig de Gràcia and Rambla de Catalunya, Les Rambles) or landmarks (e.g. streets around the Cathedral, Santa Maria del Mar, Palau de la Música).

Both sets of maps display the linear density (pictures per meter) as the width of the segment itself. The density values are log-10 transformed, and therefore a unit increase corresponds to a 10-fold increase in density. These transformed values were standardized for classification.

8.4.2 Panoramio Pictures per Segment

While the Panoramio dataset had fewer locations than the Flickr and Twitter data (Table 8.1), the focus of its service on geotagged pictures of landmarks provided an opportunity to explore the attractiveness of the points of interest in the city (as discussed in section 3.3).

As a first approximation to the streets with more pictures, the picture count was aggregated per street from the street segment data (as discussed in section 8.3.6). From the computed counts, the list of the streets with more pictures (Table 8.3) contained a mix of streets with strong tourist attraction (Passeig de Gràcia), streets with an unusual concentration of pictures in a particular segment (Carrer Provença and Carrer Marina), very long streets (Avinguda Diagonal) and very short streets next to important landmarks (Plaça Sagrada Família).

Table 8.3: Top streets with the largest number of geotagged pictures collected from Panoramio.

Street Name	Length (m)	Pictures
Carrer Provença	4,476	2,436
Avinguda Diagonal	11,929	1,952
Plaça Sagrada Família	534	1,796
Passeig Gràcia	1,745	1,785
Carrer Marina	3,973	1,774

Table 8.4: Top streets with the highest linear density of geotagged pictures collected from Panoramio, excluding streets of less than 50 m long.

Street Name	Length (m)	Density (pics/m)
Carrer Olot	191	6.34
Plaça Sagrada Família	534	3.37
Plaça Josep Puig i Cadafalch	79	2.73
Plaça Nemesi Ponsati	148	1.68
Plaça Portal de la Pau	330	1.64

Focusing on linear picture density instead of counts (Table 8.4), the most intensely photographed streets in Barcelona corresponded tourist attractions: Carrer Olot (Park Güell), Plaça Sagrada Família, Plaça Josep Puig i Cadafalch (Font Màgica), Plaça Nemesi Ponsati (Estadi Olímpic) and Plaça Portal de la Pau (Monument a Colom).

The overview map (Fig. 8.7) confirmed the bias of the Panoramio data source towards pictures of landmarks, as most of the street segments with outstanding picture density corresponded to important points of interest visited by tourists.

The observed macro pattern provided a valuable picture of the most photographed landmarks in Barcelona, overlaid on the pattern of the urban fabric. In this structure, the axis corresponding to Passeig de Gràcia is dominant —overflowing into Rambla de Catalunya and the neighboring "tributary" street segments—, and contains two of the most highly photographed segments, corresponding to La Pedrera and Casa Batlló. This axis extends into Carrer Gran de Gràcia where it branches into the structure of the old town.

Additional landmarks appear farther away as "satellites" —creating smaller structures that in most occasions overflow into the neighboring streets— at a smaller scale. Some of these areas are centered around a single landmark but others encompass larger areas. In this structure, Ciutat Vella is the most complex and

prominently photographed feature, and as a matter of fact Passeig de Gràcia can be considered its extension, as it has been historically.

Focusing on Ciutat Vella (old quarter) (Fig. 8.8), the structure replicates most of the features observed in the city at a different scale, with streets that attract most of the visitors and overflow into the nearest street segments around them, while additional clusters of interest punctuate specific areas, most of them with a small network of highly photographed streets around them.

8.4.3 Flickr Pictures per Segment

While Panoramio users were more focused on sharing geotagged pictures of landmarks across the world, the objectives of Flickr users were not as clear-cut and much more diverse (as discussed in section 3.4).

First, the geotagging aspect in Flickr is generally a consequence of the GPS capabilities of smartphones and the automatic tagging of geolocation information in the EXIF[21] metadata of the pictures taken by users, not an intentional action of the users.

Second, the subjects photographed are not limited to architecture and landscapes; while these themes are well represented, other themes much more mundane can be found. This was especially true in the period when there Flickr had no competition from other services such as Facebook —or more recently Twitter or Instagram— as a photo storage and sharing platform.

Despite these issues, which make the information from Filckr much "noisier" than the retrieved from Panoramio, the sheer amount of locations (Table 8.1), and the rich metadata provided by the service makes Flickr a valuable resource.

The most photographed streets (Table 8.5) seem to support this hypothesis, and the expected tourist attractions appear mixed with destinations less linked to visitor activity and more with the daily activities of the residents in Barcelona or its area of influence (the origin of Flickr users is further discussed in chapter 5, especially in sections 5.3 and 5.4).

The same effect can be observed in the list of the most densely photographed streets (Table 8.6), which in this case is not dominated exclusively by tourist destinations, and incorporates locations associated to leisure activities destined to locals.

The same effect can be appreciated in the overview map centered on the Eixample district (Fig. 8.9), where the pattern observed in the Panoramio dataset (Fig. 8.7)

[21] At the time of writing, the Exchangeable Image File Format (EXIF) reference website is available at *http://www.exif.org/* while the standard is maintained by the Japan Electronics and Information Technology Industries Association (JEITA) available at *http://www.jeita.or.jp/english/*.

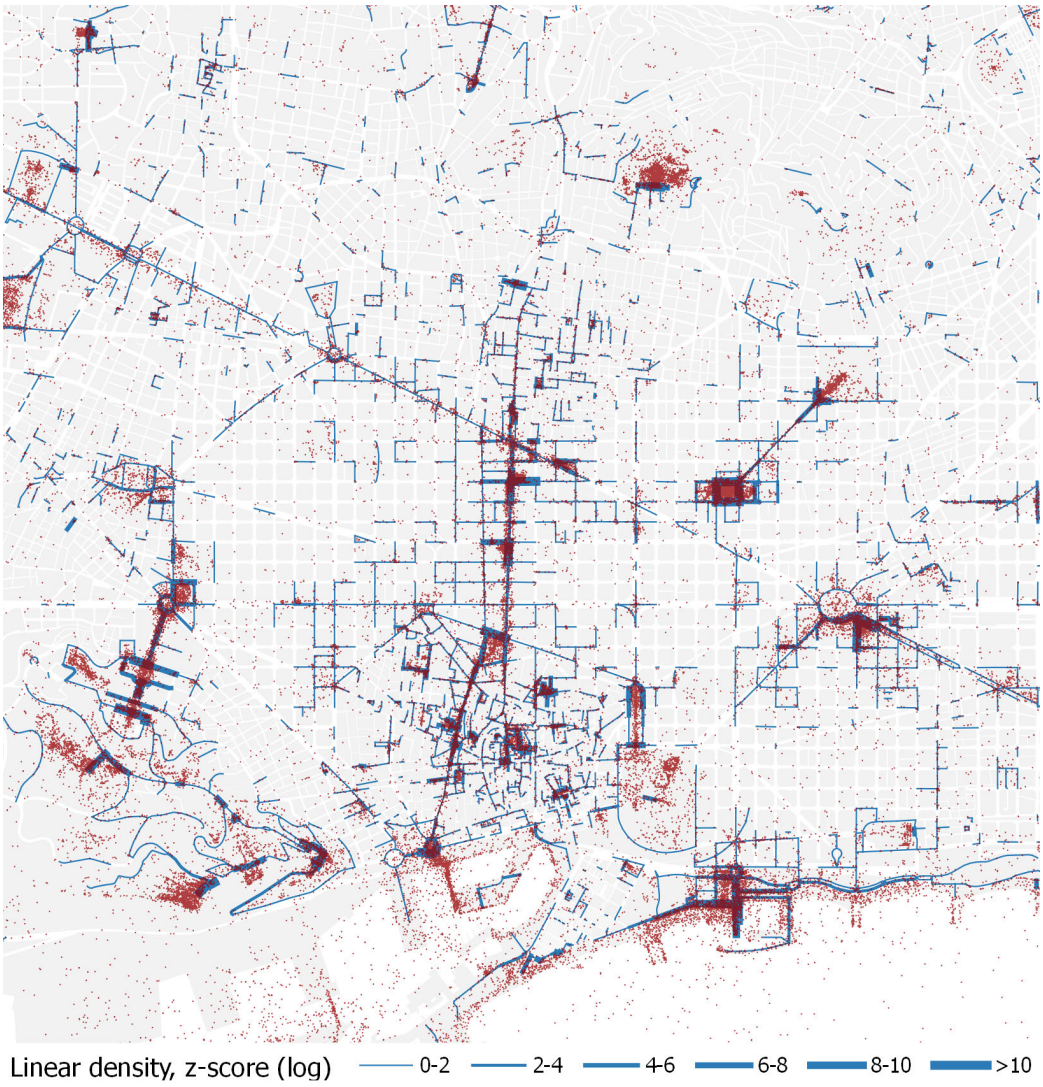


Figure 8.7: Overview map of the linear density of the geotagged pictures collected from Panoramio per street segment in central Barcelona. Magnitudes are log-transformed. Segments are classified according to their z-score on the transformed values, and segments whose values are below the mean are excluded. Picture locations are shown in translucent red. Map is rotated 44.4° clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. Map scale is 1:50,000. Road network and city block cartography from CartoBCN, under CC BY 3.0.

Figure 8.8: Detail map of the linear density of the geotagged pictures collected from Panoramio per street segment in the Barcelona old quarter (Ciutat Vella). Magnitudes are log-transformed. Segments are classified according to their z-score on the transformed values, and segments whose values are below the mean are excluded. Map is rotated 44.4° clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. Map scale is 1:15,000. Road network and sub-parcel cartography from CartoBCN, under CC BY 3.0.

Table 8.5: Top streets with the largest number of geotagged pictures collected from Flickr.

Street Name	Length (m)	Pictures
Carrer Nou de la Rambla	1,288	39,658
Carrer Provença	4,476	37,366
Passeig Gràcia	1,745	29,109
Carrer Aribau	2,208	19,690
La Rambla	1,186	17,739

Table 8.6: Top streets with the highest linear density of geotagged pictures collected from Flickr, excluding streets of less than 50 m long.

Street Name	Length (m)	Density (pics/m)
Carrer Comtes	88	64.80
Carrer Olot	191	63.93
Carrer Canuda	264	57.70
Plaça Sant Josep	258	50.04
Plaça Reial	277	37.78

is also visible. While both maps use the same scale based on standardized units, the larger Flickr dataset allows assigning pictures to more segments (Table 8.2) and because the counts cannot be below zero, more streets pass the threshold fixed on the mean, as a consequence of the Chebyshev's inequality (which limits the fraction of values that can be a certain number of standard deviations away from the mean).

In the detail map of the Ciutat Vella district (Fig. 8.10), a pattern similar to the corresponding map of the Panoramio data set (Fig. 8.8) appears, although the larger number of collected locations from Flickr makes the result richer, with the drawback of being noisier. In this case, the map shows more clearly the streets that connect the areas that appeared disconnected in the locations collected from Panoramio, revealing a wider network of popular streets and —perhaps more importantly— streets that for some reason are less visited.

8.4.4 Twitter Messages per Segment

Twitter messages are not as inextricably linked to their recorded locations as pictures are, but its pervasive use can be a good indicator of the overall activity of its users. Because of its popularity, Twitter can be a valuable resource to record

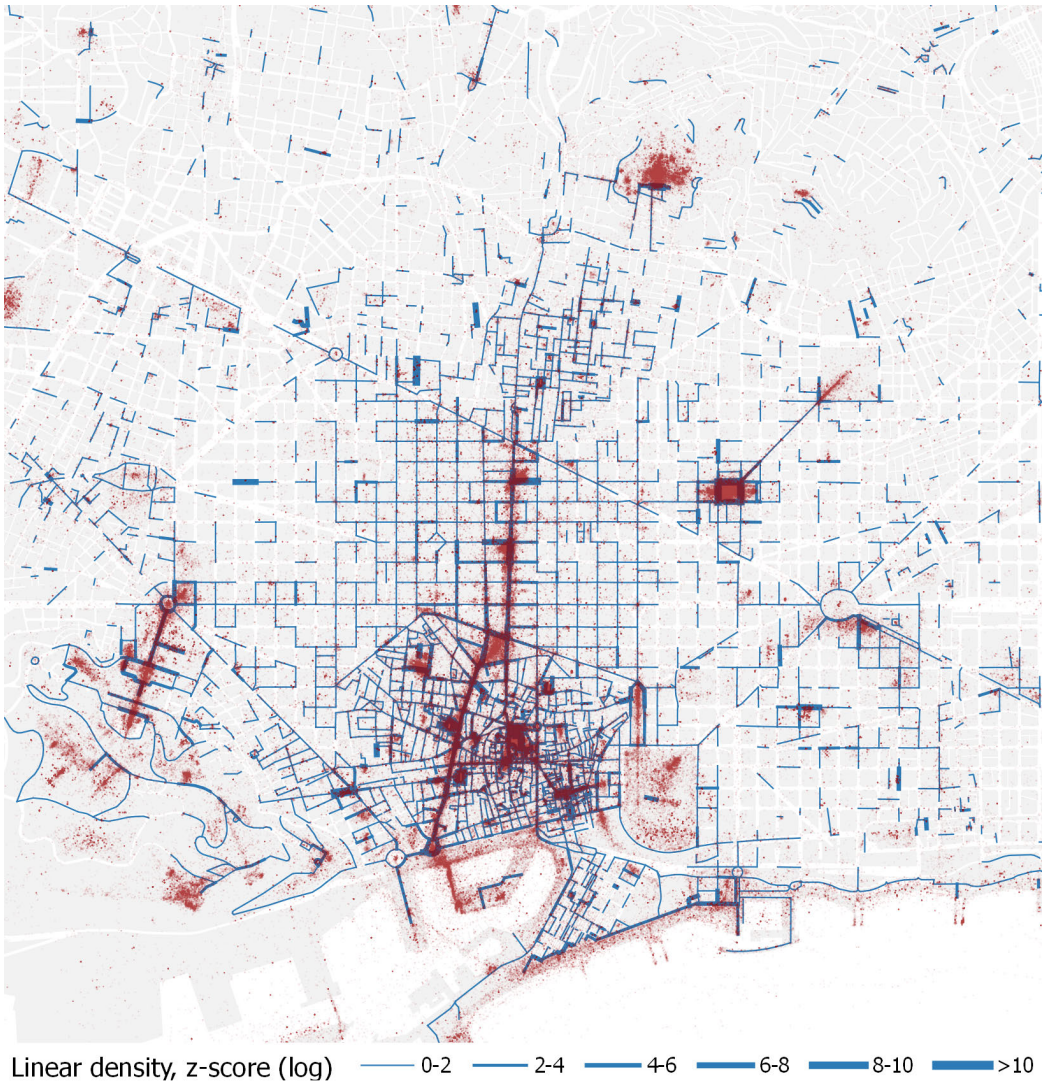


Figure 8.9: Overview map of the linear density of the geotagged pictures collected from Flickr per street segment in central Barcelona. Magnitudes are log-transformed. Segments are classified according to their z-score on the transformed values, and segments whose values are below the mean are excluded. Picture locations are shown in red (90 % transparency). Map is rotated 44.4° clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. Map scale is 1:50,000. Road network and city block cartography from CartoBCN, under CC BY 3.0.

Figure 8.10: Detail map of the linear density of the geotagged pictures collected from Panoramio per street segment in the Barcelona old quarter (Ciutat Vella). Magnitudes are log-transformed. Segments are classified according to their z-score on the transformed values, and segments whose values are below the mean are excluded. Map is rotated 44.4° clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. Map scale is 1:15,000. Road network and sub-parcel cartography from CartoBCN, under CC BY 3.0.

Table 8.7: Top streets with the largest number of geotagged messages collected from Twitter.

Street Name	Length (m)	Messages
Pla Palau	630	77483
Plaça Catalunya	766	15441
Carrer Marina	3,973	11744
Passeig Gràcia	1,745	9544
Carrer Olot	191	7841

the aggregate the spatio-temporal behavior of its users, while providing a plethora of textual content and metadata.

In this research, only the tweets linked to Instagram or Foursquare accounts were collected, the first because it included an image and the second because it was strongly related to the location through its mobile app (as discussed in section 3.6). A complete year (Table 3.4) was retrieved weekly (Table 3.3) —because of the limitations of the (free) Twitter API— to provide a dataset that covered a temporal span long enough to study seasonal changes (discussed in chapter 9).

As a source of location data, Twitter appeared at the other end of the spectrum than Panoramio, providing a rough indirect measure of the location of its users across the collected year instead of a sharp image of the attractiveness of landmarks (discussed in section 6.2).

Despite these limitations, Twitter data provided an opportunity to study not only the high activity areas (hot spots) but also the areas with unusually low activity (cold spots). The identification of these "deserts" should be very valuable to provide insight on areas where the activity is abnormally low.

The streets where the majority of tweets originate (Table 8.7) follow a similar pattern as in the Flickr dataset, including important landmarks but also places often visited by locals, providing a measure of the widespread use of the service, used by visitors and residents alike.

Homogenizing the streets computing their linear density (Table 8.8) also produced similar results, but in this case the imprecision of the location data obtained from Twitter made some short streets to appear more prominent, either by random chance or because of biases in the geolocation algorithms.

The resulting overview map of the segments centered in the Eixample district (Fig. 8.11) shows an even larger amount of segments using the same classification criteria (discussed in section 8.4.1) used in the corresponding Panoramio (Fig. 8.7) and Flickr (Fig. 8.9) maps; while the same streets segments with a large amount of retrieved locations observed in the Panoramio and Flickr maps are strongly

Table 8.8: Top streets with the highest linear density of geotagged messages collected from Twitter, excluding streets of less than 50 m long.

Street Name	Length (m)	Density (msgs/m)
Pla Palau	630	123.03
Carrer Olot	191	41.15
Plaça Catalunya	766	20.15
Plaça Sant Jaume	213	17.37
Carrer Pietat	111	16.01

represented, many streets with less intensity but with a significant amount of activity are also present.

The detail map of the tweet density in Ciutat Vella (Fig. 8.12) exhibits pattern similar to the one observed in the overview map (Fig. 8.11), where the more homogeneous spatial distribution of the collected messages allows a significant number of street segments that were filtered out in the corresponding Panoramio (Fig. 8.8) and Flickr (Fig. 8.10) maps to pass the established threshold. The most visited streets and their adjacent street segments are clearly visible, as well as the areas around some landmarks, but in this case the variations in intensity are more gradual.

The results suggest that Twitter data is more representative of the true spatial behavior of users, and that it can be very valuable to understand the spatial patterns of human behavior in urban settings. In addition, the semantic aspect of tweets can be further analyzed using text mining tools [143], along with other metadata from the Twitter API, opening a new avenue of research into the identification of "deserts" (areas without activity) inside the city, especially when combined with temporal data.

8.5 Network-constrained Kernel Density

8.5.1 Background

The Kernel Density Estimation (KDE) applied to spatial analysis (discussed in section 6.4) is a widely used technique to study the distribution of events in euclidean (geographic) space. However, KDE assumes a homogeneous space, which is not always applicable in situations where this approximation does not hold true —such as a city near a lake—, because it cannot capture the impossibility of an event (e.g. a traffic accident) to occur in the space occupied by the lake.

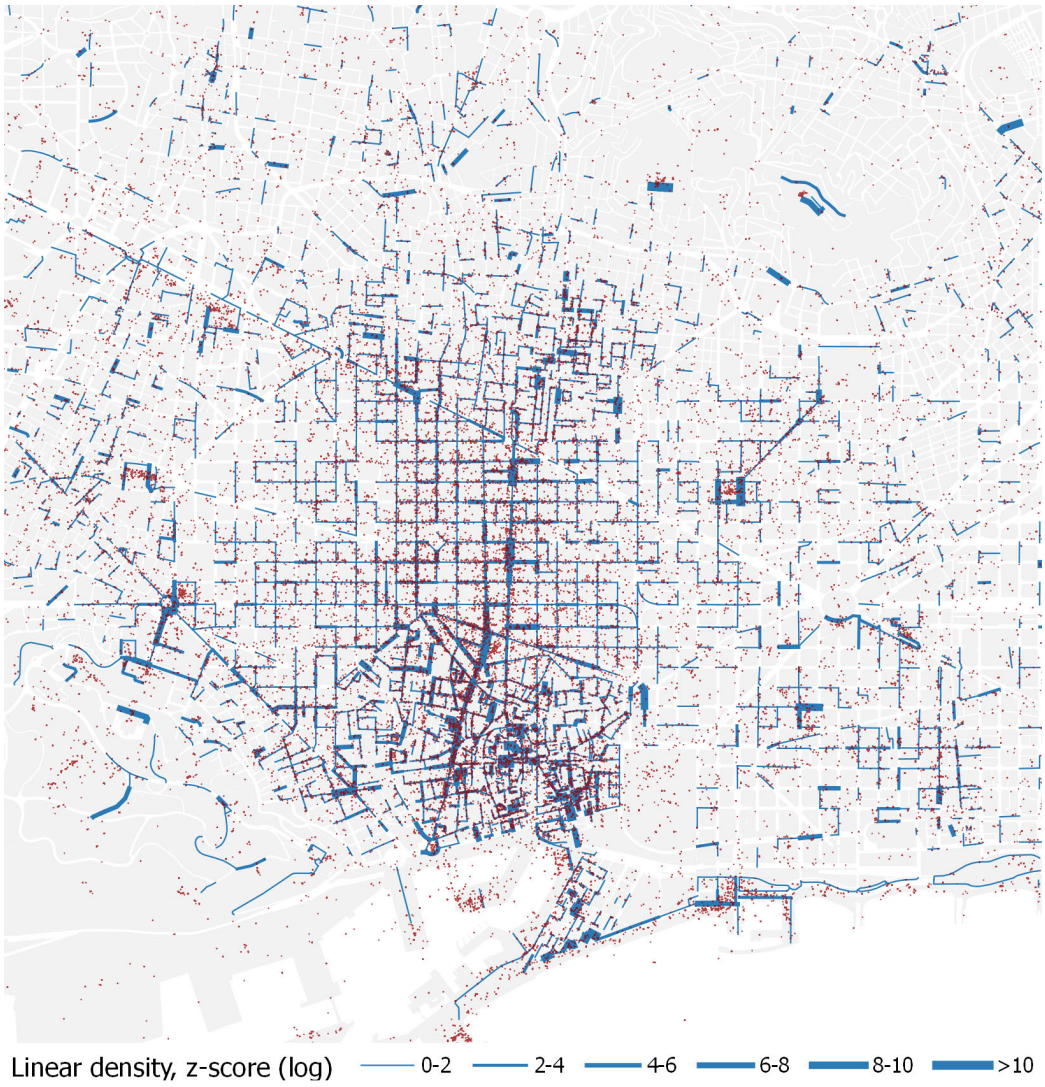


Figure 8.11: Overview map of the linear density of one year of geotagged messages collected from Twitter per street segment in central Barcelona. Magnitudes are log-transformed. Segments are classified according to their z-score on the transformed values, and segments whose values are below the mean are excluded. Picture locations are shown in translucent red. Map is rotated 44.4° clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. Map scale is 1:50,000. Road network and city block cartography from CartoBCN, under CC BY 3.0.

Figure 8.12: Detail map of the linear density of one year of geotagged messages collected from Twitter per street segment in central Barcelona. Magnitudes are log-transformed. Segments are classified according to their z-score on the transformed values, and segments whose values are below the mean are excluded. Map is rotated 44.4° clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. Map scale is 1:15,000. Road network and sub-parcel cartography from CartoBCN, under CC BY 3.0.

Furthermore, when computing KDE for events that happen in the street network, the results can be misleading —specially in detailed analyses— because the events are considered in a planar space when they really are confined to a network, which demand a different analytical approach [381].

This is reflected in the *de facto* coordinate system used to navigate the cities, which is not cartesian but based the network topology of the streets, and composed of two pieces of information: the street name and the postal number (edges and nodes).

Therefore, to improve the accuracy of the developed methodology that computed the density per street segment, which was adequate for urban scales (discussed in section 8.3), another approach —more suitable for detailed analyses— was tested, estimating the kernel density constrained to the street network.

8.5.2 Software Implementation

The computation of the network-constrained kernel density is technically complex, especially for large data sets. There are multiple software implementations capable of calculating the kernel density for networks, based on the algorithm developed by Okabe [382]:

SANET named after the acronym for Spatial Analysis Along Networks, the original toolbox implementing the algorithm developed by Okabe and his team at the University of Tokyo[22], who were very kind providing a license of their software.

PySAL through the module of Network Constrained Analysis[23] (pysal.network), included in the Python Spatial Analysis Library [316] developed by Rey and Anselin.

GRASS through the module v.kernel[24] developed by Menegon and Blazek.

spatstat through the function density.lpp[25] of this R package[26] that includes tools analyze spatial point patterns [327, 326] developed by Baddeley and McSwiggan [383].

[22]SANET is freely available for academic and educational purposes at *http://sanet.csis.u-tokyo.ac.jp/* at the time of writing.

[23]The PySAL Network Constrained Analysis documentation is available at *http://pysal.readthedocs.io/en/latest/library/network/network.html* at the time of writing.

[24]The GRASS v.kernel command documentation is available at *http://grass.osgeo.org/grass74/manuals/v.kernel.html* at the time of writing.

[25]Details of the spatstat implementation of the density.lpp function (Kernel Estimate of Intensity on a Linear Network) are available at *http://rdrr.io/cran/spatstat/man/density.lpp.html* at the time of writing.

[26]The spatstat website is available at *http://spatstat.org/* at the time of writing.

Table 8.9: Computation times for the three developed methodologies, measured on the same computer.

Source	Records	Address-based	Distance-based	Kernel density
Panoramio	80,459	7.1 s	0:34 h	0:14 h
Flickr	1,166,704	47.3 s	6:45 h	2:32 h
Twitter	519,664	34.6 s	3:17 h	1:39 h

After testing the alternatives, the software of choice was GRASS, because of the features it offered and the speed of its implementation, but also the flexibility of its GIS functionality regarding import and export capabilities and handling of spatial references.

The network-constrained kernel density approach was much slower (Table 8.1) than the optimized address-as-anchor method (discussed in section 8.3.4), but about twice as fast than the "brute-force" distance-based method (discussed in section 8.3.1).

For the calculation, the segments corresponding to the street axes were split into pieces not exceeding 5 m. The kernel radius was fixed at 25 m using the default Gaussian kernel[27], with the standard deviation of the Gaussian function set to 1/4 of the radius, using the equal spit method [382] to produce unbiased density estimates. To reduce the computation time, a maximum distance threshold was set to 50 m, beyond which points were not considered.

8.5.3 Application to Geotagged Picture Locations

After processing the data with the GRASS implementation of the algorithm, the resulting street segment fragments were represented in a map, with the computed values of the network-constrained kernel density estimation as their line width, and the picture locations overlaid as reference.

Because this approach was suitable for detailed scales, which are not possible to show in their entirety because the size limitations of an A4 sheet of paper, the example figures for the results from Panoramio (Fig. 8.13) and Flickr (Fig. 8.14) data consist on an overview map (top) and three detail maps (bottom) focusing on three enlarged areas of this map, from left to right:

1. A segment of Passeig de Gràcia showing the local picture intensity in front of two landmarks (La Pedrera and Casa Batlló), and the influence on adjacent streets (Provença and Aragó).

[27] In the GRASS implementation, the following kernel density functions were also available: uniform, triangular, epanechnikov, quartic, triweight, and cosine.

Table 8.10: Descriptive statistics of the distribution resulting from the computation of the network-constrained kernel density estimation.

Source	Range	Mean	Std. Dev.	Skewness	Kurtosis
Panoramio	0 – 58.65	0.11	0.69	38.34 >> 0	2,016.67 >> 3
Flickr	0 – 1,220.23	0.97	9.66	54.07 >> 0	4,382.23 >> 3
Twitter	0 – 1,729.38	0.61	8.56	116.45 >> 0	19,352.80 >> 3

2. An area of Ciutat Vella including Las Ramblas between Mercat de la Boqueria and Plaça Reial (west), and the area around the Cathedral and Sant Jaume Square (east).
3. The surroundings of Sagrada Família, including its two adjacent parks and part of Avinguda Gaudí.

However, the distribution of these values for the three sources was very skewed (Table 8.10) and had to be transformed from the range produced by the kernel density function (different for each source) to a new range, interpolating from 0 to 40 meters using an exponential curve[28] of exponent 1/2 (square root). The result of this transformation was assigned as the line width of the corresponding segment fragment.

[28]More details on the transformation are available at *https://github.com/qgis/QGIS/blob/master/resources/function_help/json/scale_exp* at the time of writing.

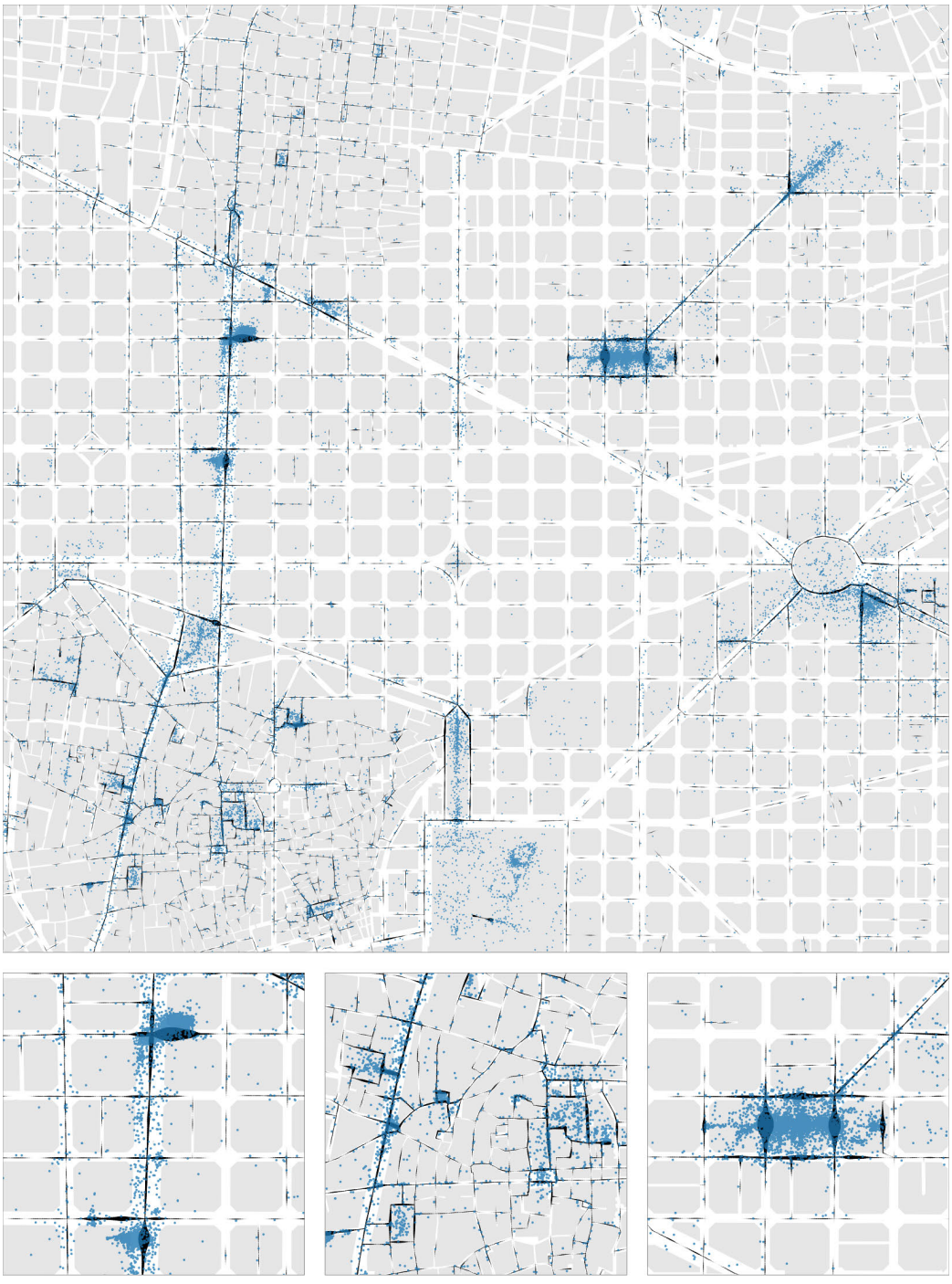


Figure 8.13: Network-constrained kernel density estimation applied to the geotagged pictures of Barcelona collected from Panoramio. Segment fragments are exponentially scaled with an exponent of 1/2. Picture locations are shown in translucent blue. Maps are rotated 44.4° clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. General map scale is 1:25,000 and detail maps scales are 1:15,000. Road network and city block cartography from CartoBCN, under CC BY 3.0.

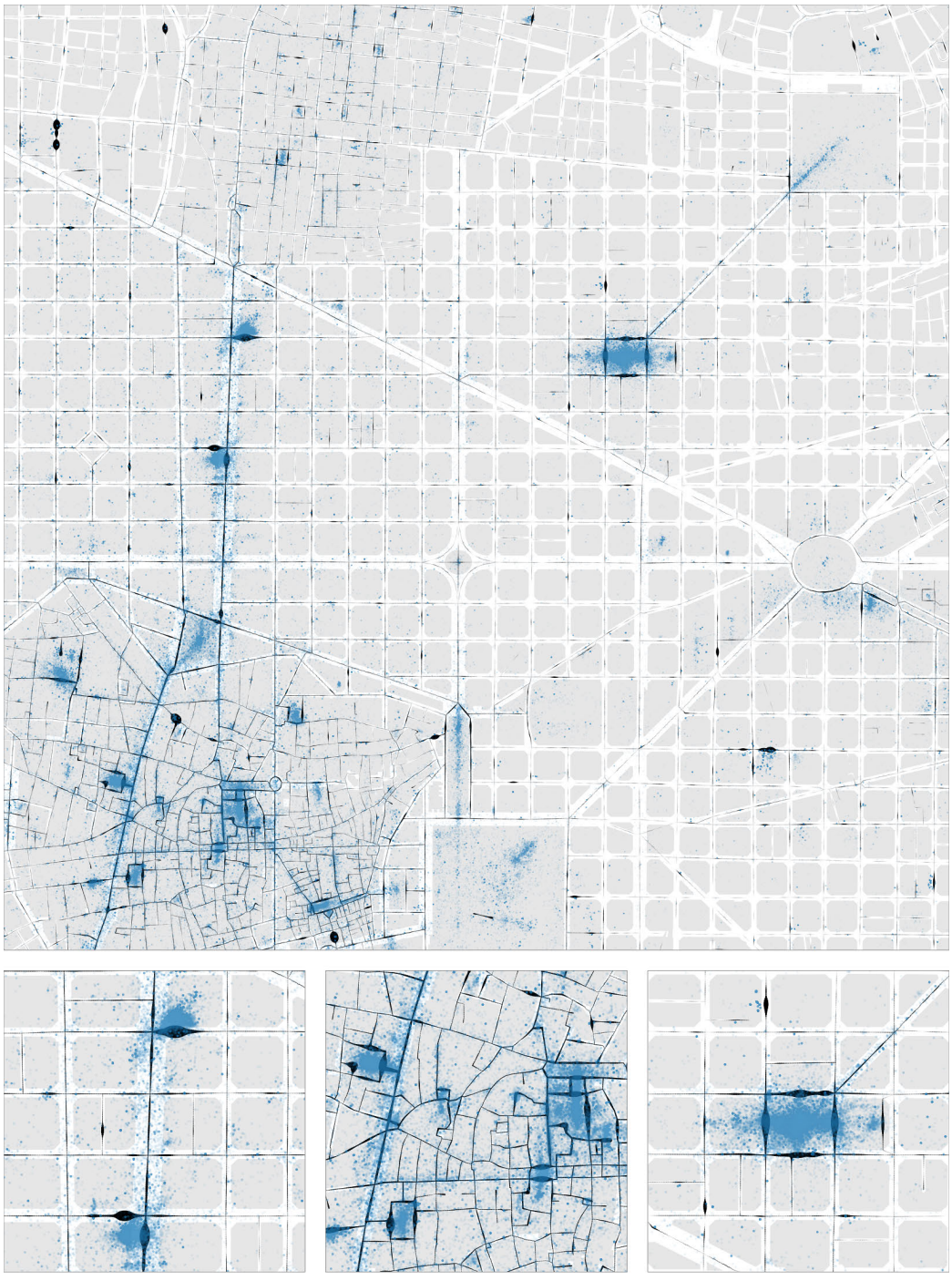


Figure 8.14: Network-constrained kernel density estimation applied to the geotagged pictures of Barcelona collected from Flickr. Segment fragments are exponentially scaled with an exponent of 1/2. Picture locations are shown in translucent blue. Maps are rotated 44.4° clockwise from the orientation corresponding to the ETRS89 / UTM zone 31N (EPSG:25831) projection. General map scale is 1:25,000 and detail maps scales are 1:15,000. Road network and city block cartography from CartoBCN, under CC BY 3.0.

Chapter 9

The Temporal Dimension

> "The only reason for time is so that everything doesn't happen at once."
>
> Albert Einstein

9.1 Introduction

9.1.1 Background

Until recently, most temporal data about cities came from government-sponsored one-off surveys, conducted at specific moments of time, and therefore cross-sectional [138] from the temporal point of view, offering only a series of "snapshots" of data.

Today, the availability of data from new sources—collected through users or automatically generated by sensors— routinely incorporate temporal data [384], providing an opportunity to conduct research on the temporal aspects of the life in the city.

This chapter discusses a methodological approach to *handle* and *visualize* temporal data from these sources, accounting for their challenges (discussed in section 2.3.2) and focusing on the most relevant aspects for urban analysis, while its practical application to a case of study in London has already been published separately [143].

9.1.2 Chapter Outline

While previous chapters have focused on the *spatial* aspects of urban data, this chapter discusses their temporal dimension. In this case, time is analyzed separately from space, and therefore the spatial and temporal interaction is not discussed, because of its complexity [385, 386], but also because of the visualization challenges of representing time and space simultaneously in a printed medium.

This chapter discusses the process of transforming raw temporal data collected from Flickr into a variety of visualization techniques that help understand the temporal dimension of data, adopting and developing existing methodologies, and it is divided into the following sections:

- Section 9.2 discusses the specific challenges and complexity of temporal data.
- Section 9.3 explains required conversion, transformation and filtering processes of the collected data to ensure its accuracy.
- Section 9.4 defines the analyzed temporal cycles (also discussed in section 7.4) and proposes a methodology to adjust the computed frequencies according to the variation of dayttime and nighttime across the year.
- Section 9.5 investigates the necessity to homogenize temporal data (also discussed in section 7.4.4) using time series.
- Section 9.6 introduces the calendar heatmap as a useful summary to visualize an overview of long running time series.
- The next three sections examine in isolation three important cycles in the life of the city: the daily cycle of daytime and nighttime (section 9.7), the weekly cycle of work and leisure (section 9.8) and the yearly cycle of seasons (section 9.9).
- Section 9.10 explores visually the interaction of the cycles discussed in the previous sections using heatmaps, pairing cycles in the horizontal and vertical axes.

9.2 Temporal Data

9.2.1 Properties of Temporal Data

Time has specific properties that make collecting, storing, analyzing and visualizing temporal data challenging [243]. Some of these difficulties arise from the properties of time itself as a dimension in space-time[1], while others arise from

[1]For example, time is apparently irreversible.

cultural or technical backgrounds.

Scale Scientists need to record time from astronomical or geological scales to sub-nanosecond precision, depending on their field of study. In this dissertation, however, the discussed temporal scale is the same as in everyday operations.

Cycles Some cycles are derived from natural phenomena, such as the duration of the year (one revolution of Earth around the Sun), a day (the rotation of Earth around its axis), and to some degree, a month (roughly the duration of the moon cycle). However, other cycles such as weeks[2] are cultural constructs and are not synchronized to the yearly cycle.

Units Time does not use a decimal notation in the division of the day, which has 24 hours, divided into 60 minutes with 60 seconds each. The number of days in a year or a month are not multiples of 10 either (and are variable depending on multiple factors).

Semantics When we refer to time, it can mean an instant (where we consider that the recorded event does not have a duration itself) or an interval (a time span, with a beginning and end). Intervals can be further divided, depending on whether they have consistent lengths (durations) or are anchored in an instant (periods).

9.2.2 Challenges of Temporal Data

Most of the challenges when dealing with temporal data arise from the definition of time used in our everyday activities, which complicates arithmetic with temporal data because the time line is not as reliable as the number line to perform calculations [387].

Calendar Introduced in 1582 as a refinement of the Julian calendar, the Gregorian calendar is the most widely used civil calendar worldwide, but other calendars are used locally. It specifies among other things the duration of the year and the epoch (reference date used as the origin of a particular era).

Irregularities Compared to the number line, the time line has many irregularities: months of different duration (28, 29, 30 or 31 days), leap years and leap seconds.

Time Zones The rotation of Earth requires adjusting the clocks to synchronize the times in different places of the planet to their perceived daily cycle. The tz database[3] is a comprehensive compilation of information on the world's

[2] Although a seven-day week is approximately a quarter of a lunation.

[3] The latest version (2018g) is available from IANA at *https://data.iana.org/time-zones/releases/* at the time of writing.

time zones[4].

Latitude The sunset an sunrise times of any given day within a time zone depend on the latitude of the location.

Daylight Saving Time This practice adjusts the clocks twice a year, advancing the clocks in summer with the objective of obtaining more evening daylight, but delaying sunrise times. The clock is generally returned to standard time in autumn in the northern hemisphere. The adjustment dates are variable across territories.

Localization Many regional and language conventions exist to describe the time beyond using different names for months and days of the week, such as the order of the components of the dates (day, month and year), the first day of the week (Sunday or Monday) or the numbering of the hours within the day (12 or 24 hour clock), which difficult parsing temporal data stored as text strings.

9.3 Handling Temporal Data

9.3.1 Data Sources

The source of the analyzed temporal data was the geotagged pictures of Barcelona collected from the Flickr photo sharing service (discussed in section 3.4). The source data consisted in 1,166,704 picture records produced by 34,283 users (Table 3.5).

Since pictures are rarely uploaded immediately after they are taken —even on mobile devices with high-speed wireless connections— the picture metadata contained three pieces of temporal data[5]:

- The date and time the picture was posted (uploaded) to the service, stored as a UNIX timestamp in Greenwich Mean Time (GMT).
- The date and time the picture was taken, extracted from its EXIF metadata (when available) or the time of upload stored in the MySQL "datetime" format.
- The accuracy or ("granularity") of the instant the picture was taken, offering four possible options, from the most accurate to the least: seconds, months, years or "circa…".

An additional piece of temporal information was the time zone of the user who uploaded the photo to the service, retrieved through the *flickr.people.getInfo* API

[4]A map of the world time zones is available at *http://efele.net/maps/tz/world/* at the time of writing.

[5]More details available at https://www.flickr.com/services/api/misc.dates.html

call, which included the following metadata about his or her timezone (Table 4.1): a time zone label, a unique time zone identifier, and the offset from the Coordinated Universal Time (UTC).

9.3.2 Managing Temporal Data

The complexity of temporal data require standards-based software libraries to parse, store and manipulate temporal data. The present research used the R package lubridate [387], based on the Boost C++ libraries[6] and using the conceptual framework of the Joda-Time Java library[7]. The lubridate package allowed the following operations:

- Parsing and conversion of date-times in various formats into the appropriate classes.
- Time zone conversions[8].
- Extraction of date (year, month, week) and time (hour, minute) components.
- Determine the day of the week of any given date.

9.3.3 Conversion

The first step to handle temporal data was the conversion of the collected data into one of the time classes handled by the R programming language, both inheriting from the POSIXt virtual class.

POSIXct Represents the number of seconds since the Unix epoch (beginning of 1970[9] in the UTC time zone) as a signed 32-bit integer. This format does not include time zone or DST information.

POSIXlt Consists in a named list of vectors representing the components defining a specific time and date. Among other additional data, this format does include time zone or DST information (in contrast with POSIXct).

In the case of the upload date, the format was stored as a Unix time stamp in the UTC time zone, and therefore the conversion was trivial. This time information was then converted to the Barcelona time zone, accounting for DST when applicable.

[6]The Boost C++ libraries are available at *http://www.boost.org/* at the time of writing.
[7]The Joda-Time Java library is available at *http://www.joda.org/joda-time/* at the time of writing.
[8]The R programming language supports 594 time zone definitions (Olson Names).
[9]Corresponding to 00:00:00 of January 1, 1970 (Coordinated Universal Time).

Conversely, the date the picture was taken had to be parsed from its text representation, stored in the MySQL datetime format. However, this string had a crucial piece of information missing to unequivocally refer to a specific moment of time: the time zone.

It was therefore necessary to obtain the corresponding time zone from the corresponding user metadata. Joining the two pieces of information (date-time and UTC offset) produced a time *instant*, parseable in the ISO 8601:2004 format[10].

9.3.4 Discarded Data and Accuracy Issues

For the purposes of the analysis, assuring the quality of time information was crucial. Since time had to be aggregated into intervals —hours, days and months—, the results were very sensitive to inaccuracies in the source data that could potentially distort the results:

Super users Discussed in section 3.7.2, outliers with a large number of pictures who exhibited a temporal pattern very different than the general behavior could distort the results[11], and therefore pictures from users whose picture count exceeded six standard deviations from the sample mean were discarded.

Precision Approximate temporal data is stored in the same format as precise temporal data, setting the unknown time and/or date components to the lowest possible value (e.g. year 2017 is stored as "2017-01-01 00:00:00", and May 4th 2017 is stored as "2017-05-04 00:00:00"), and therefore there is a potential artifact consisting on the accumulation of events at the beginning of the year, month, day and —to a smaller extent— hour. For this reason, the analyses discarded the data capable of distorting the results according to the level of temporal aggregation (Table 9.1).

Metadata In the rare cases when EXIF data was not available, the service assigned the date of upload as the date the picture was taken. Therefore pictures where these times and dates matched were excluded from the analyses.

Time zone For hour and day precision, pictures of users in a time zone whose UTC offset did not match the offset of Barcelona were discarded to ensure

[10]The ISO 8601:2004 format "Data elements and interchange formats – Information interchange – Representation of dates and times" is available at *http://www.iso.org/standard/40874.html* at the time of writing.

[11]For example the pictures of Barcelona nightlife in the weekly Anti-Karaoke sessions at the Apollo venue every Monday, available at *http://www.flickr.com/photos/49603476@N04/* at the time of writing.

Table 9.1: Adopted criteria to discard temporal data because of potential accuracy issues.

Precision	Discarded time stamps	Discarded granularity
Years	None	> 6
Months	Beginning of the year	> 4
Days	Beginning of the month	-
Hours	Beginning of the day	0

the time data was accurate. For months precision this restriction was not necessary.

Data was discarded according to the granularity of the analysis being conducted, to keep the largest amount of valid data regarding the required accuracy, except for the "super-users", which were excluded in all the analyses.

9.4 Probability Adjustments of Cycles

9.4.1 Cycles Definition

The research on the temporal patterns focused mainly on the interaction patterns between the different cycles (Table 9.2) in the daily life of cities. It was therefore necessary to extract the necessary information from the date and time the picture was taken.

Daily cycle Each picture was classified according to whether in the time and date it was taken (in the Barcelona time zone and at the Barcelona latitude) was astronomically day or night. This cycle is discussed in section 9.7.

Weekly cycle The days of the week were classified into work days (Monday to Friday) and weekends (Saturday and Sunday). This cycle is discussed in section 9.8.

Yearly cycle The meteorological seasons were used —as opposed to the astronomical seasons— and months were classified according to groupings of three whole months, as defined since 1780 by the Societas Meteorologica Palatina. This cycle is discussed in section 9.9.

9.4.2 Unbalanced Cycles

To compare the different cycles, their different duration had to be taken into account. While the probability of an event to fall into any of the four seasons in

Table 9.2: Analyzed cycles and their constituent elements.

Periodicity	Measurement units	Cycle
Daily	Hours and/or minutes	Day and night
Weekly	Days	Work and leisure
Yearly	Months	Seasons

the yearly cycle (almost) was the same —all things being equal—, that was not the case in the other two cycles:

- The probability of an event to fall into the leisure category of the weekly cycle was lower, as it only included two days of the seven days of the week.
- The probability of an event to be classified into day or night was the same overall over the course of the year, but changed continuously from one day to another along the seasons.

For most calculations, these probabilities were taken into account and corrected, giving each element in a category the weight corresponding to the inverse of the sum of all the probabilities of the elements in its same category, essentially considering both categories as the two sides of a biased coin.

As a result, if we consider the null hypothesis that the temporal distribution of events is random, the observed events in both categories would tend to be the same. At the same time, if a statistically significant result is observed, the influence of this bias of can be ruled out.

9.4.3 Day and Night Duration throughout the Year

The daily cycle corresponds to the rotation of Earth around its axis, which is perceived by an observer on its surface as the Sun moving through the sky from sunrise to sunset, and disappearing at night.

The axial tilt of the Earth makes the relative daytime and nighttime duration change throughout the year, except for locations situated at the equator. At higher latitudes, this oscillation becomes more dramatic until reaching the polar circles, where the sun does not appear or disappear above or below the horizon at least one day throughout the year.

At middle latitudes such as Barcelona (41° 23' North), the oscillation is as not dramatic but certainly noticeable, and it is further distorted by DST for half the duration of the year (Fig. 9.1). For the research, the daytime and nighttime periods were defined as follows:

Daytime as the period between sunrise and sunset.

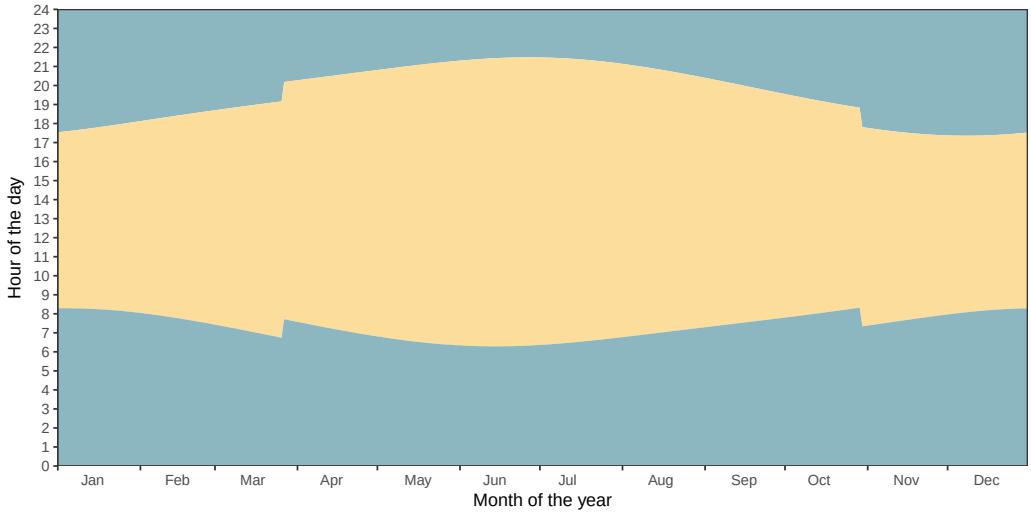


Figure 9.1: Yearly sun graph for the approximate latitude of Barcelona in a typical non-leap year (2017). Discontinuities in March and October are the result of Daylight Saving Time (DST) adjustments in the "Europe/Madrid" time zone. Source: Own work based on the algorithms provided by NOAA through the sun methods in the maptools R package.

Nighttime as the period before sunrise and after sunset.

All daytime and nighttime calculations were computed with the R package maptools 0.9-2 [388], which provided the necessary functions to compute sunrise and sunset times[12] for any given day at any location[13], using the algorithms developed by the National Oceanic & Atmospheric Administration (NOAA).

9.4.4 Daytime and Nighttime Probabilities

While the theoretical probability of a picture —modeled as a random event— to be taken during either the day or the night is the same throughout the year overall (50%), this probability varies continuously according to which day the picture was taken (Fig. 9.2), being exactly 50% only in the March and September equinoxes[14], and most dissimilar in the June (longest day) and December (longest night) solstices[15].

To account for this variability, the corresponding frequencies were weighted

[12]Computation of the times of dawn and dusk is also available.
[13]For latitudes less than ±72 degrees, accuracy is approximately one minute.
[14]In 2017, March 20 at 11:29 and September 22 at 22:02, respectively.
[15]In 2017, June 21 at 06:24 and December 21 at 17:28, respectively.

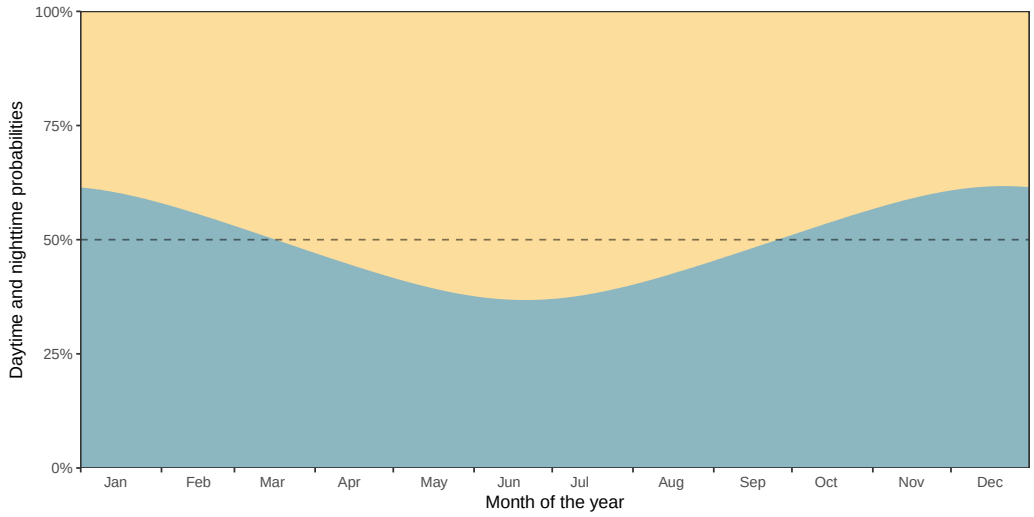


Figure 9.2: Probability that a random event will occur during nighttime or daytime, proportional to the duration of each period at the approximate latitude of Barcelona in a typical non-leap year (2017). Source: Own work based on the algorithms provided by NOAA through the sun methods in the maptools R package.

accordingly, to make probabilities comparable following the calculations discussed in section 9.4.3. Therefore, the figures of the evolution of the daily cycle during the year take into account this variation, as well as the all the necessary time zone and DST corrections applicable, resulting in probability-adjusted frequencies.

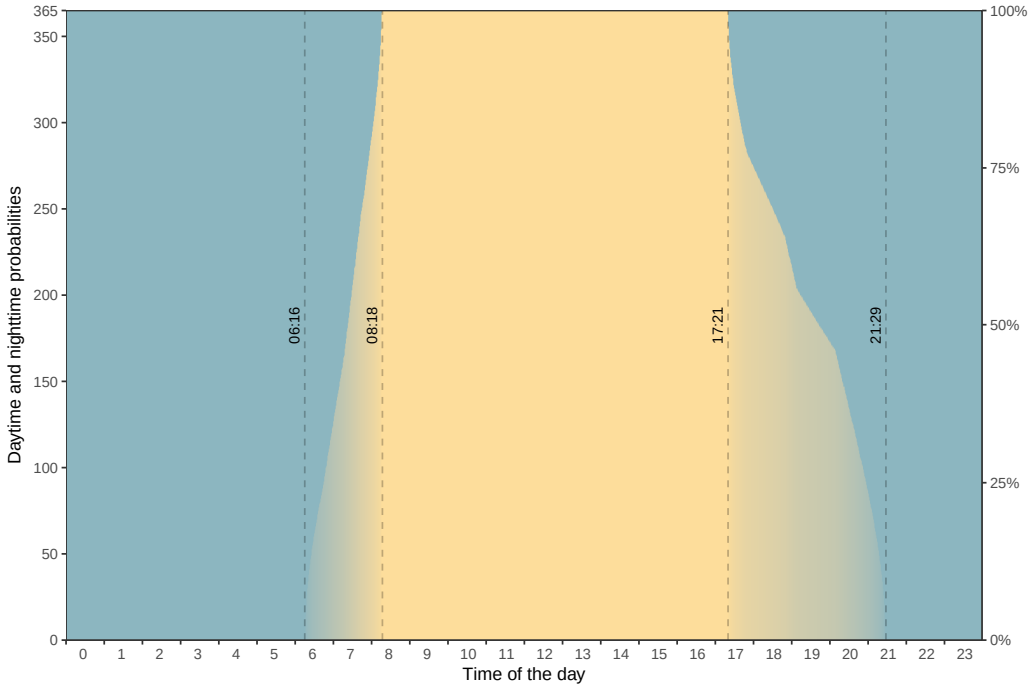
Furthermore, the probabilities also affected the analysis of the daily cycle itself, as each instant of the day did not have the same probability of being day or night throughout the year (Fig. 9.3), and had to be adjusted accordingly.

For the duration of the day, it was possible to compute the number of days —and therefore the probability— that it was daytime or nighttime (Fig. 9.3b). In the example non-leap year 2017, taking into account DST, it was always nighttime before 6:16 and after 21:29, and it was always daytime between 8:18 and 17:21 (Fig. 9.3a).

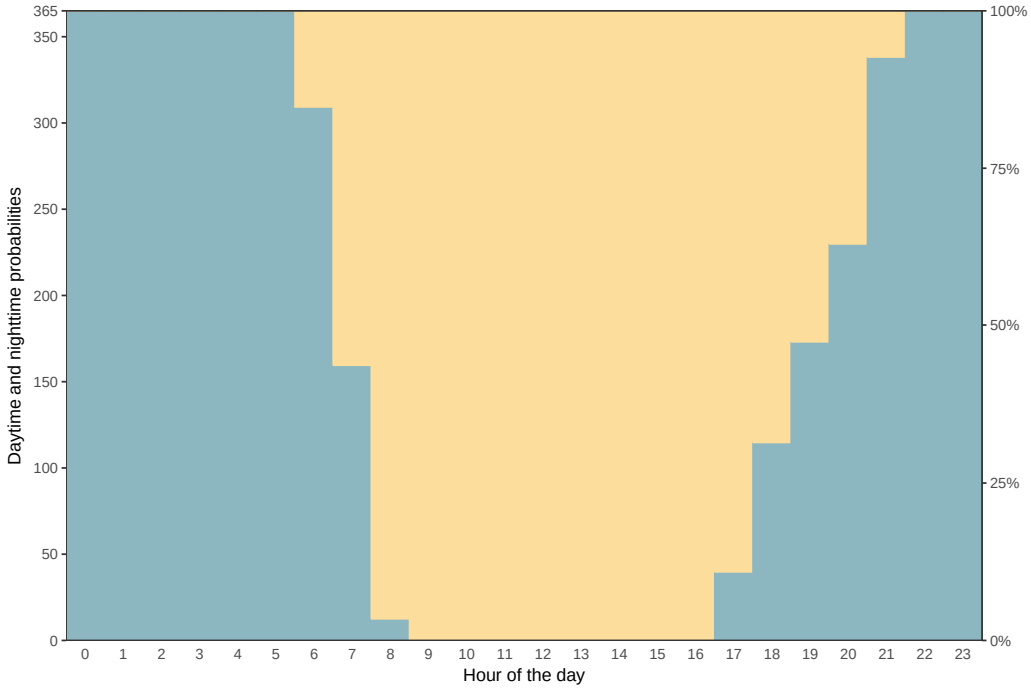
9.5 Time Series Decomposition

9.5.1 Time Series

The popularity of networks changes with the shifting behavior of their user bases (Fig. 3.2). This section investigates the necessity to homogenize temporal



(a) Daytime and nighttime distribution across the minutes of the day.



(b) Daytime and nighttime distribution across the hours of the day.

Figure 9.3: Number of days it was daytime or nighttime for each minute (top) or hour (bottom) of the day at the approximate latitude of Barcelona in a typical non-leap year (2017). The top figure uses a color gradation to denote probability. Source: Own work based on the algorithms provided by NOAA through the sun methods in the maptools R package.

data (also discussed in section 7.4.4) to make comparisons across years without introducing additional biases, using time series analysis.

Time series data consist in a series of measurements ordered in time, generally in equally spaced time intervals (regular time series), but on occasions with unequally spaced intervals (irregular time series).

Time series are generally plotted as line charts, with a line connecting the different measurements of each series across time, conventionally with time in the horizontal axis (with older observations on the left and more recent observations on the right) and the variable or variables of interest in the vertical axis.

Time series data exhibit auto-correlation, because data points are not independent but correlated to one another. For most time series, the best predictor for an event is the previous event or some combination of close events in time[16].

The time series infrastructure relied in the R packages zoo[17] 1.8-0 [389] and xts[18] 0.10-0 [390], which provided the required conversion and manipulation facilities as well as a framework for further analysis.

Both packages are based in a time series object that consists in a matrix (which stores the actual data with time periods as rows and variables as columns), accessed through an index (constructed using a vector of a formal time class).

For the analysis, the time stamps of the geotagged pictures of Barcelona collected from Flickr were converted into a time series object. The process consisted in their aggregation (as count data) into either one-month or one-week bins, resulting in 156 or 626 temporal data points respectively.

9.5.2 Classical Decomposition

The classical seasonal decomposition by moving averages break up a time series into three components: seasonal, trend and irregular, using either an additive or multiplicative model (Fig. 9.4), depending on whether the three components are assumed to be summed or multiplied to produce the observed data:

Trend Using this method, the trend component is first determined using a simple

[16]Similarly, spatial data also are also auto-correlated, as in Tobler's First Law of Geography: "everything is related to everything else, but near things are more related than distant things".

[17]The zoo package is available at CRAN at *http://cran.r-project.org/package=zoo* at the time of writing.

[18]The xts package is available on GitHub at *http://github.com/joshuaulrich/xts* at the time of writing.

moving average[19], with the objective of smoothing short-term variations and highlighting the long-term component of the analyzed data.

Seasonal After removing the trend component from the model, the seasonal component is (optionally) obtained by averaging for each time unit over all periods (which have to be complete), and centering the resulting periodic figure.

Irregular The error component is the (random) remainder of the model compared to the data when predicted using the trend and seasonal components.

When applied to the data, using a seasonal frequency of 12 months, both models show a steady increase in the number of pictures from 2005 until a peak that remains relatively flat for two years (2011-2013), and slowly but steadily decreases until the end of the data collection (2016).

The seasonal component has two peaks in the warmer seasons (spring and autumn), with a small dip in August and a larger cyclic decrease in winter. However, the seasonal component is very small compared to the others.

9.5.3 Additional Seasonal Adjustment Methods

Other decomposition methods beyond the classical time series decomposition were also considered but discarded, because the results were not satisfactory for the Flickr data set:

- Structural Time Series by maximum likelihood, as implemented in the stats 3.4.2 core R package.
- Exponential smoothing state space model, as implemented in the forecast 8.2 R package[20].
- TBATS model (Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components), as implemented in the forecast 8.2 R package.

The seasonal adjustment software X-13ARIMA-SEATS developed by the United States Census Bureau[21], which combines the X-12ARIMA software with the

[19] Essentially a finite impulse response filter.

[20] The forecast package is available on GitHub at *http://github.com/robjhyndman/forecast* at the time of writing.

[21] The X-13ARIMA-SEATS Seasonal Adjustment Program is available at *http://www.census.gov/srd/www/x13as/* at the time of writing.

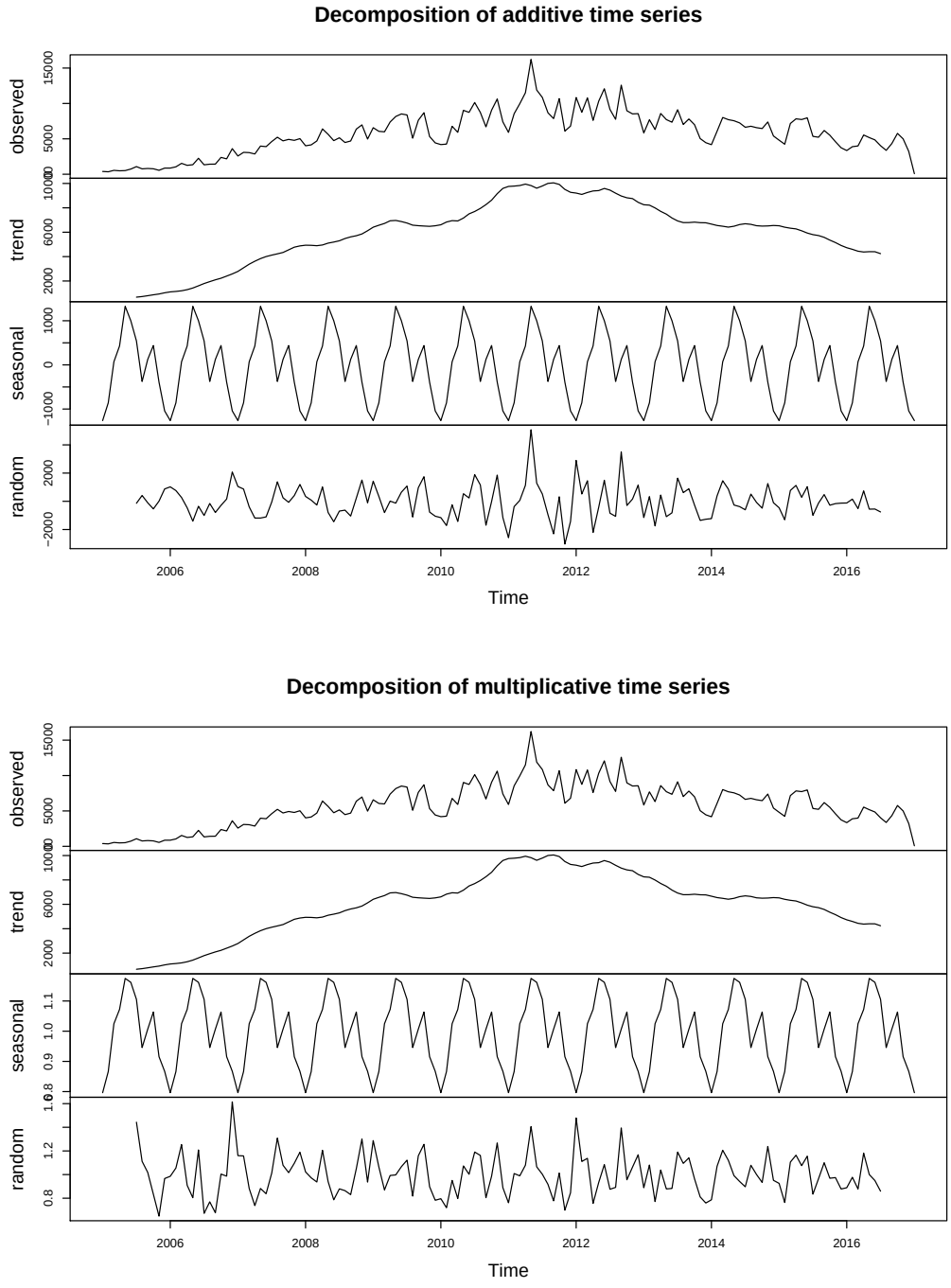


Figure 9.4: Classical time series decomposition of the number of geottagged pictures of Barcelona collected from Flickr in the 2005–2016 period. The aggregated monthly time series is decomposed into three components, using the additive (top) and multiplicative (bottom) methods.

TRAMO-SEATS software developed by the Bank of Spain, was used through the seasonal 1.6.1 R package[22], but the results were also considered unsuitable.

However, the decomposition using the STL method [391], as implemented in the stats 3.4.2 core R package, using loess (local non-parametric regression) was considered adequate and was used to decompose the monthly and weekly aggregated time series (Fig. 9.5). The results of the STL decomposition showed that the trend was smoother and that allowing the seasonal component to change its amplitude resulted in smaller remainder components in the model.

9.6 Calendar Heatmaps

9.6.1 Calendar-Based Graphics

The temporal range of the collected data was large, including data from 12 years in the 2005–2016 period. It was therefore necessary a visualization strategy to display this temporal data in a single chart, with the difficulty of showing 4,383 days in a meaningful way.

Calendar-based graphics allows visualizing an overview of the evolution of a measurement in univariate time-series data, and identifying patterns and trends on multiple time scales (daily, weekly and yearly) using a single chart [392].

Multiple R packages exist that produce this type of graphics: ggTimeSeries[23], ggcal[24], sugrrants[25], and the calendarHeat script[26]. However, the limitations of these packages decided the author to develop a custom solution[27], as it granted more control over the output, using the R package ggplot2 as the underlying layout engine.

This strategy allowed producing any kind of visualization in a flexible way, decoupling data from visualization, allowing the transformation of data as the first step in the visualization pipeline:

[22] The seasonal package is available on GitHub at *http://github.com/christophsax/seasonal* at the time of writing.

[23] The ggTimeSeries package is available on GitHub at *http://github.com/AtherEnergy/ggTimeSeries* at the time of writing.

[24] The ggcal package is available on GitHub at *http://github.com/jayjacobs/ggcal* at the time of writing.

[25] The sugrrants package is available at *http://pkg.earo.me/sugrrants/* at the time of writing.

[26] The calendarHeat script is available on GitHub at *http://github.com/iascchen/VisHealth/blob/master/R/calendarHeat.R* at the time of writing.

[27] Inspired in the Wicklin and Allison poster "Visualizing Domestic Airline Traffic with SAS Software", available at *http://stat-computing.org/dataexpo/2009/posters/wicklin-allison.pdf* at the time of writing.

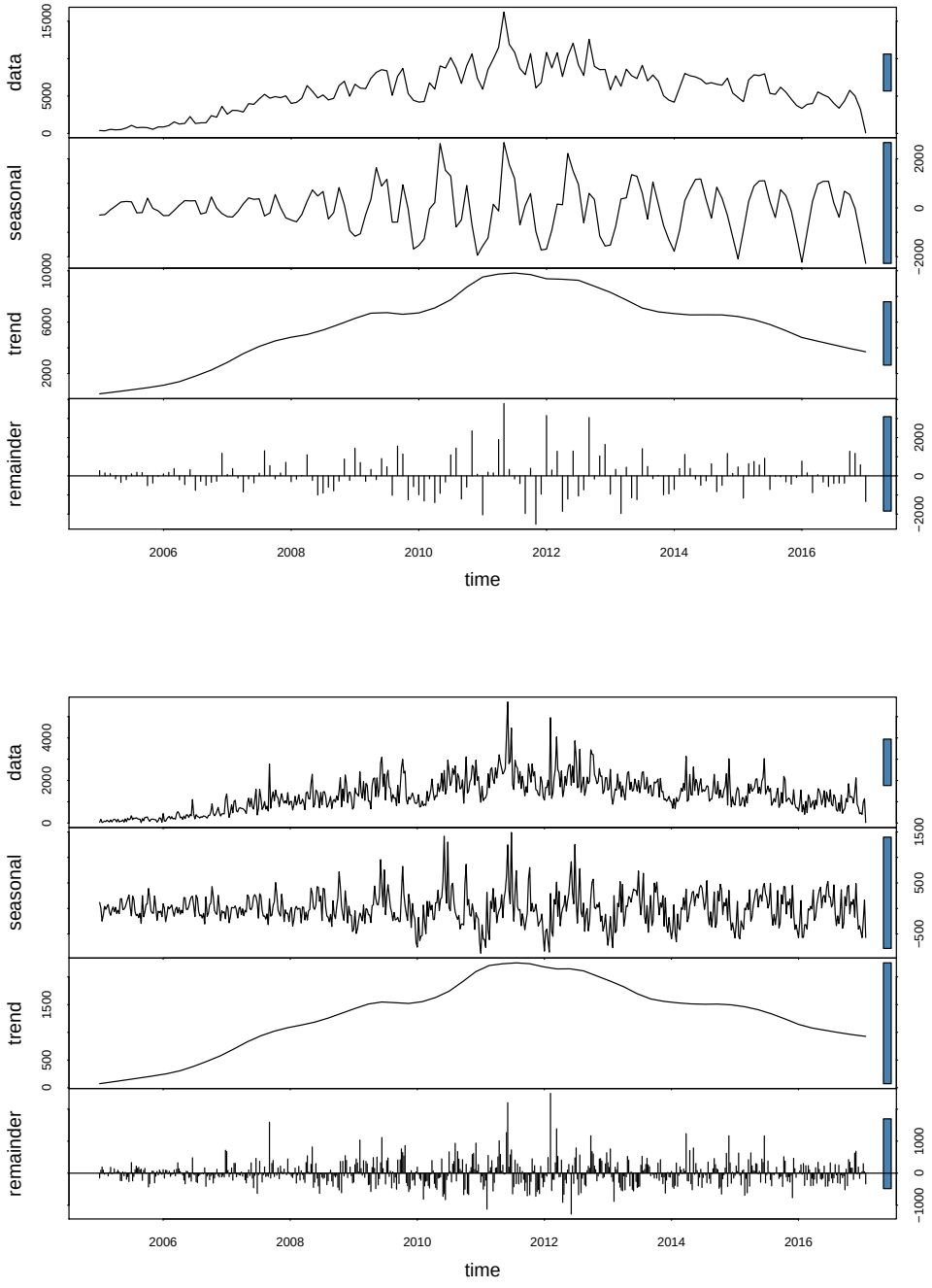


Figure 9.5: STL time series decomposition of the number of geotagged pictures of Barcelona collected from Flickr in the 2005–2016 period. The aggregated monthly (top) and weekly (bottom) time series is decomposed into three components, using loess.

Years are grouped in horizontal strips, running from the earliest at the top to the latest at the bottom.

Months are arranged horizontally from left to right —flipped and rotated from their typical arrangement in a conventional calendar— and month boundaries are visible as white lines.

Weeks of the year run horizontally in each strip, from 1 to 53 (and exceptionally 54), using Monday as the first day of week (the UK convention), where the first Monday of the year would be day one of week one.

Week days run from top to bottom in each strip, beginning in Monday (top) and ending in Sunday (bottom).

Days are tiles placed at the intersection of its corresponding weekday and week of the year, and are filled according to the measured variable in a continuous color scale.

9.6.2 Calendar of Counts

The resulting calendar heatmap covers the complete temporal span of the geotagged pictures retrieved from Flickr (Table 3.4), aggregating the picture counts into one-day bins. The graph allows visualizing multiple patterns and trends in multiple time scales.

The overall shape of the trend (discussed in section 9.5) is visible (Fig. 9.5), with fewer pictures in the first years (until 2007) and a slow but steady decline since 2013.

The peaks of the seasonal patterns are also visible, especially the decline in August, but also the reduced volume during winter and the summer months (discussed in section 9.9).

The increased activity during the weekends is also apparent, but is not as pronounced in the summer days (discussed in section 9.8), and the outliers corresponding to the Monday night karaoke parties are also visible (discussed in section 9.3.4), particularly in 2010 and 2011.

While the daily cycle (discussed in section 9.7) cannot be displayed because of the one-day binning resolution, exceptional peaks in specific days are visible, such as May 27, 2011 (Friday), which correspond to the forceful removal of camped protesters in Catalonia Square[28].

[28] As published in the press, according to El País "Los Mossos cargan contra el 15-M" available at *http://elpais.com/diario/2011/05/28/espana/1306533610_850215.html* at the time of writing.

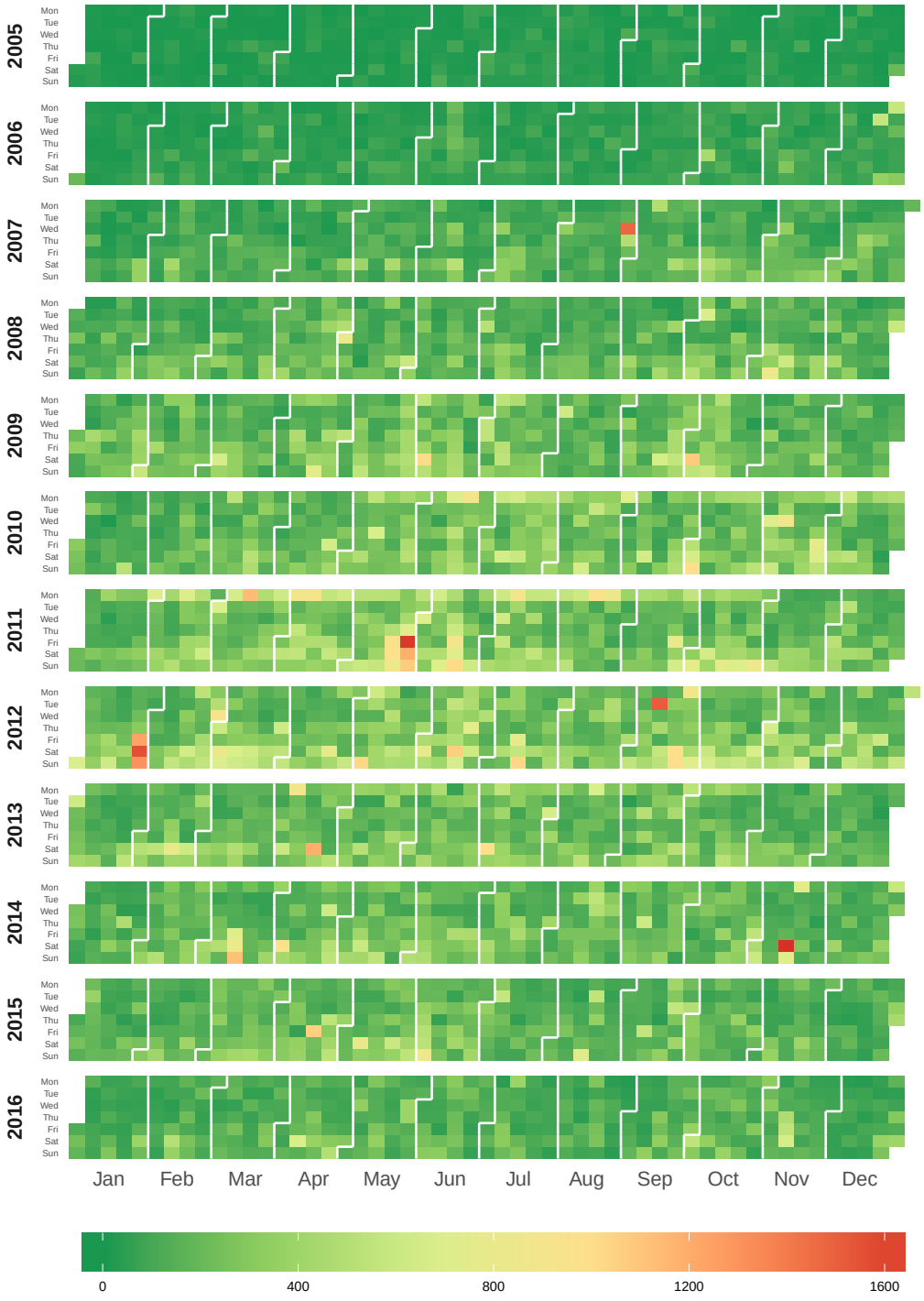


Figure 9.6: Calendar heatmap of the time stamps of the geotagged pictures of Barcelona collected from Flickr in the 2005–2016 period. Color represent the number of pictures taken in each specific day.

9.6.3 Adjusted Calendar

However, plotting the raw counts per day does not take into account the overall trend of the popularity of the Flick service (discussed in section 9.5), which peaked between 2011 and 2013 (Fig. 9.5).

To compensate this effect, visible in the calendar of counts (Fig. 9.6), the number of pictures was normalized dividing the count of each day by the sum of the counts of all days of the same year, to obtain the fraction of pictures per day in the context of the total number of pictures in that year[29].

This approach is similar to the method discussed in section 7.4.4. The resulting calendar equalized the values across years (Fig. 9.7), and avoided two types of distortion: significant peaks in a low-volume year but small in the context of all years[30], and on the contrary very pronounced peaks in the overall distribution but less exceptional in the context of its specific high-volume year[31].

9.7 The Daily Cycle

9.7.1 Daytime and Nighttime

The daily cycle oscillates between daytime and nighttime, which at the latitude of Barcelona changes in length along the year, as discussed in section 9.4. Its spatial distribution per neighborhood is discussed in section 7.4.1, while this section focuses in its temporal aspects only.

Unlike the weekly cycle (discussed in section 9.8), this is a natural cycle followed by all living organisms on Earth that are exposed or depend on sunlight or the absence of thereof, and is governed by the circadian rhythm, a biological process that oscillates with a period of roughly 24 hours.

This section explores the trends in the picture-taking behavior of Flickr users in the city of Barcelona along the daily cycle in thee different dimensions:

- Daytime and Nighttime (section 9.7.2)
- Work and Leisure (section 9.7.3)
- Meteorological seasons (section 9.7.4)

Since the figures binned the temporal data in one-hour intervals, the results were very sensitive to inaccuracies in the source data (discussed in section 9.3.4). To avoid introducing noise —despite reducing the available data—, the source was filtered with the following criteria:

[29]Therefore, the sum of all values across any year totaled 100%.
[30]For example as December 31, 2005 (Saturday).
[31]Such as the forceful removal of camped protesters on May 27, 2011 discussed earlier.

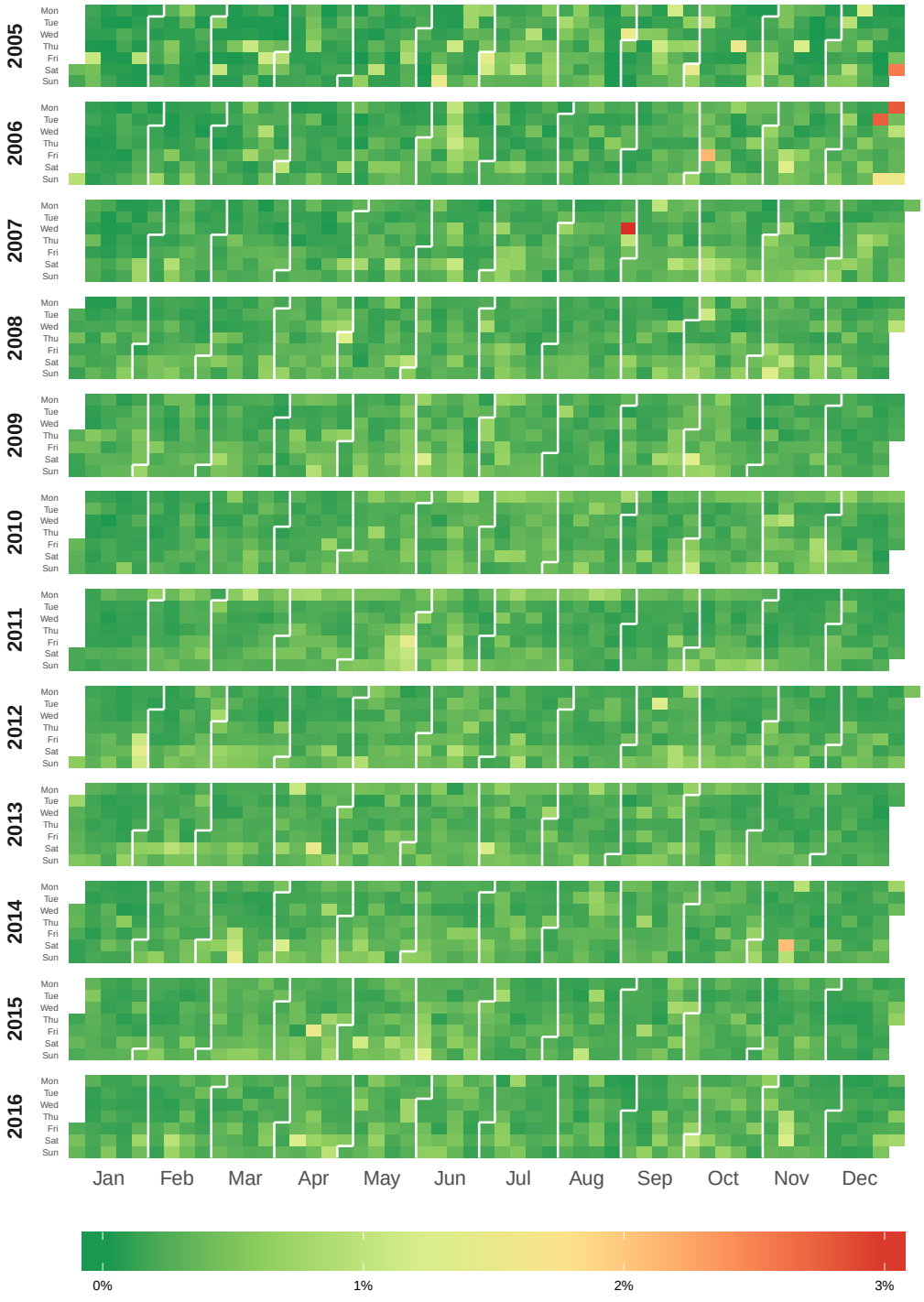


Figure 9.7: Normalized calendar heatmap of the time stamps of the geotagged pictures of Barcelona collected from Flickr in the 2005–2016 period. Color represents the fraction of pictures in the context of the total number of pictures in that year.

- Only the granularity of "0" was used (defined as data with "seconds" accuracy[32]).
- The users who resided in a location with a different UTC offset than Barcelona had their pictures excluded.
- Pictures taken by users whose number of pictures taken exceeded six times the standard deviation of the sample were discarded.
- When the "taken" date and time matched the "upload" metadata the pictures were excluded, as the temporal information was likely derived from the later.

9.7.2 Daytime and Nighttime within the Daily Cycle

For most of the population, this cycle is more or less coincident with their waking and sleeping hours, and on work days is roughly synchronized with the activities of work and study. However, artificial lighting in homes and public spaces has disrupted this once immutable relationship, blurring the lines between day and night.

Despite these changes, streets are less busy during the night hours, especially after midnight, when most of the population is at home getting ready to go to sleep, and therefore it is expected that less pictures will be taken.

In addition, regardless of the advances in camera technology, sufficient lighting is still required to produce good quality pictures —and therefore worthy to be shared in social networks—, especially pictures of exterior spaces.

Finally, it is easier to take pictures in daylight, as nighttime photography is arguably more technically complex, and the range of subjects available in daytime photography is more extensive.

Observing the hourly distribution of the most reliable time stamps (discussed in section 9.7.1), it is not surprising that the results show that the majority of pictures were taken during the day or early in the night (Fig. 9.8), as after midnight the number of pictures is drastically reduced. The distribution includes all the pictures across the 24 hours of the day —disregarding any other variable—, and the intensity seems to quickly peak around 10 AM, with a valley during the lunch hours, only to peak again around 4 PM and decrease more slowly until midnight.

Breaking down the histogram by month 9.9 in a grid allows observing the influence of the yearly cycle —which directly influences the length of day and night— in the distribution of the temporal pattern of the pictures.

[32] According to the Flickr API documentation available at *http://www.flickr.com/services/api/misc.dates.html* at the time of writing.

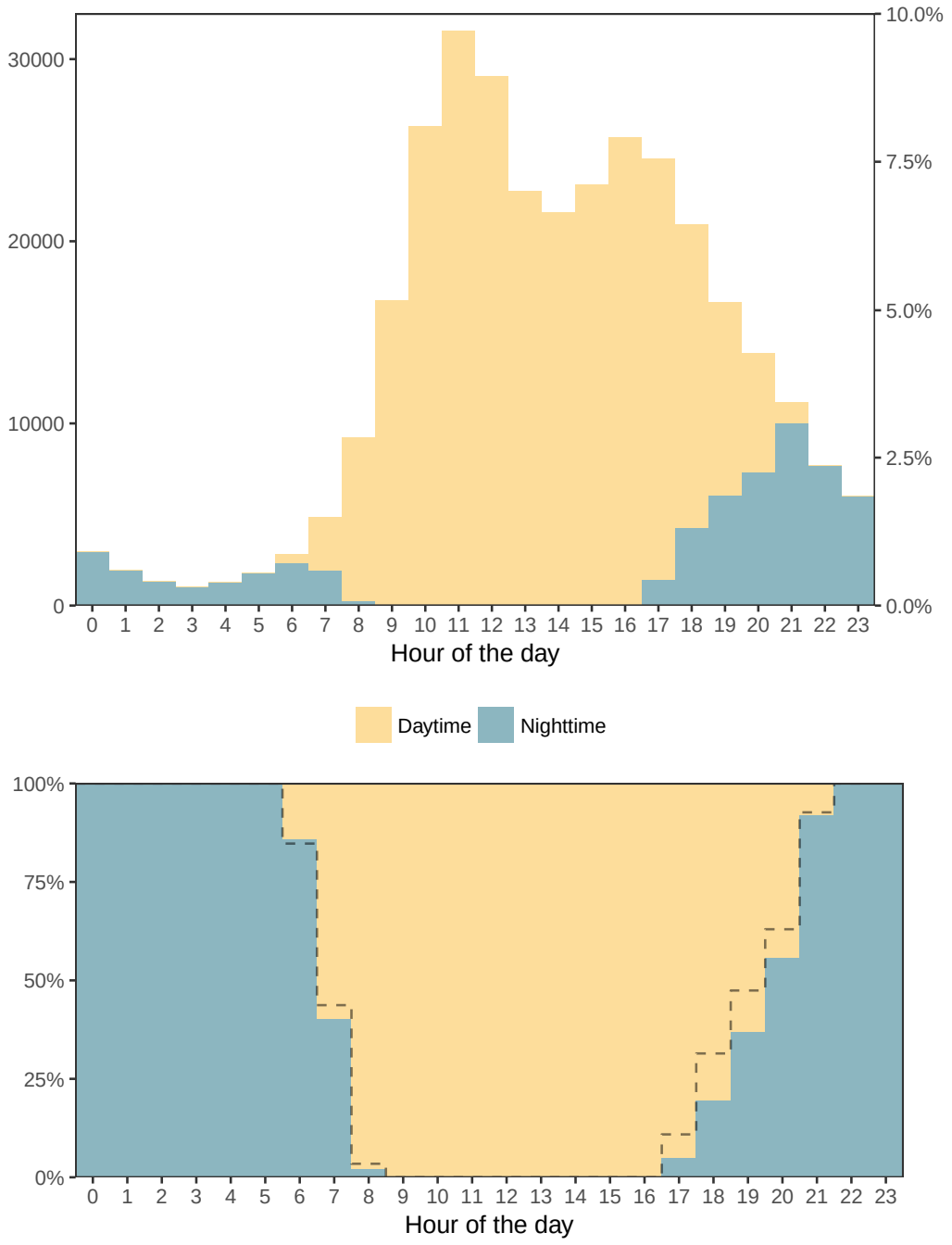


Figure 9.8: Hourly daytime and nighttime distribution of the most reliable time stamps of the geotagged pictures of Barcelona collected from Flickr. Top figure shows the actual distribution and bottom figure the probability adjusted distribution, with a dashed line depicting the theoretical distribution. Daytime and nighttime calculations at the nominal latitude of Barcelona based on the algorithms provided by NOAA through the sun methods in the maptools R package.

In these histograms, the extended available daylight between March and September (equinoxes), or conversely, the reduced daylight between September and March, can be observed. The histograms suggest that the activity is extended by roughly 2 hours after sunset, probably because the limited available dusk light is supplemented by artificial lighting from street lights.

Furthermore, the bimodal distribution observed in the single yearly histogram is not constant over the months, and seems to get more prominent in the warmer seasons, compared to the colder seasons, suggesting that as the second peak gets gradually closer to the nighttime it tends to disappear.

9.7.3 Work and Leisure within the Daily Cycle

As discussed in section 9.8.1, the distribution of workdays and weekends is not balanced, as a purely random process would be expected to result in 5/7 of the events happening on work days and 2/7 on weekends, on the long run.

Despite this probabilistic imbalance, the hourly distribution of the most reliable time stamps (discussed in section 9.7.1) during the daily cycle shows that both distributions are fairly even (Fig. 9.10), and appear to be split evenly throughout the duration of the day.

Therefore, to visually emphasize the differences it was necessary to produce a filled histogram with adjusted probabilities (Fig. 9.10). In this histogram it is visible that on weekends more pictures are taken than what it would be expected by chance, taking into account the length of the weekend compared to the work days, with a comparatively higher proportion in the central hours of the day and after midnight.

Analyzing the histogram across the yearly cycle (Fig. 9.11), the summer months (June to August) seem to reduce the proportion of pictures taken during the weekends, especially in July and August, arguably because most of the population is on summer holidays and therefore their behavior is not as dissimilar across the days of the week as in the rest of the regular months.

Furthermore, a significant part of the resident population spends their weekends elsewhere during the summer months (and in smaller proportion during the ski season), and therefore they take their weekend pictures in places elsewhere other than Barcelona.

9.7.4 Seasons within the Daily Cycle

Finally, to understand the influence of the yearly cycle in the daily distribution of pictures, a histogram broken down according to the month they were taken

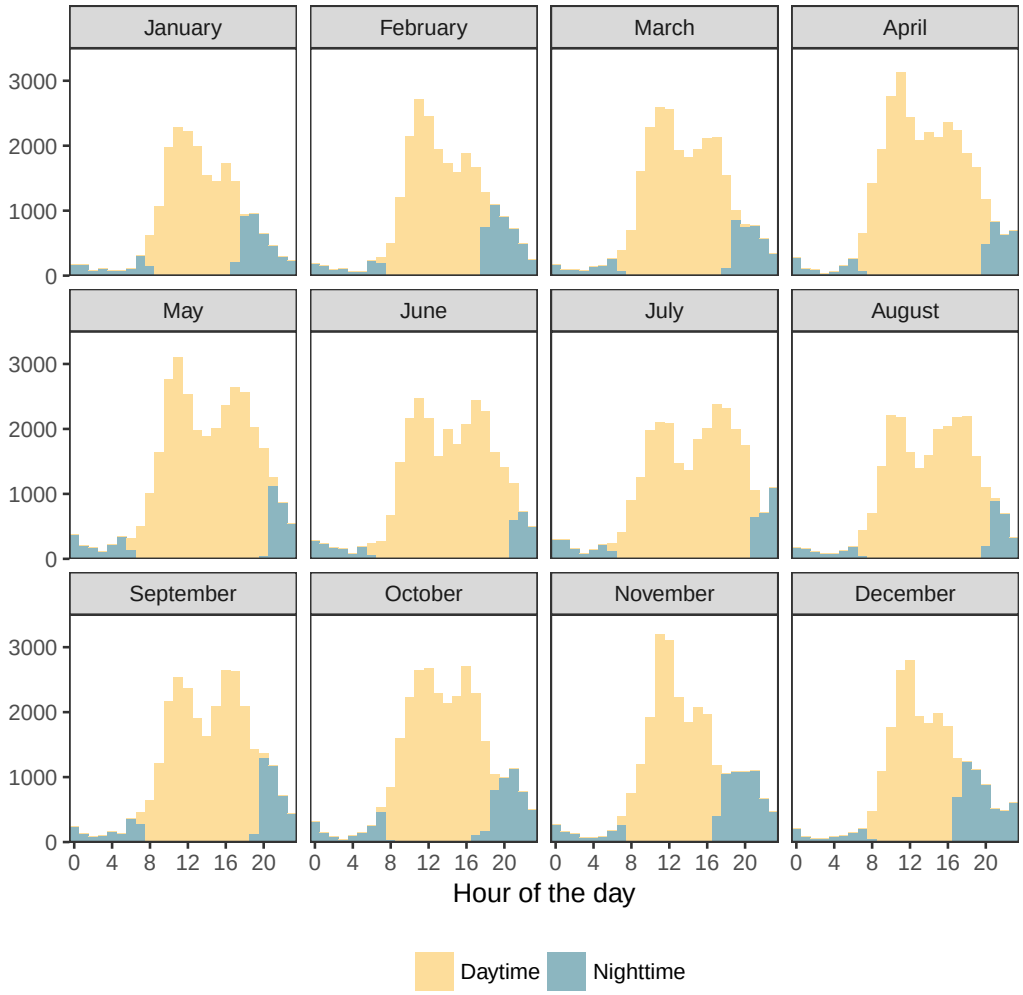


Figure 9.9: Hourly daytime and nighttime distribution per month of the most reliable time stamps of the geotagged pictures of Barcelona collected from Flickr. Daytime and nighttime calculations at the nominal latitude of Barcelona based on the algorithms provided by NOAA through the sun methods in the maptools R package.

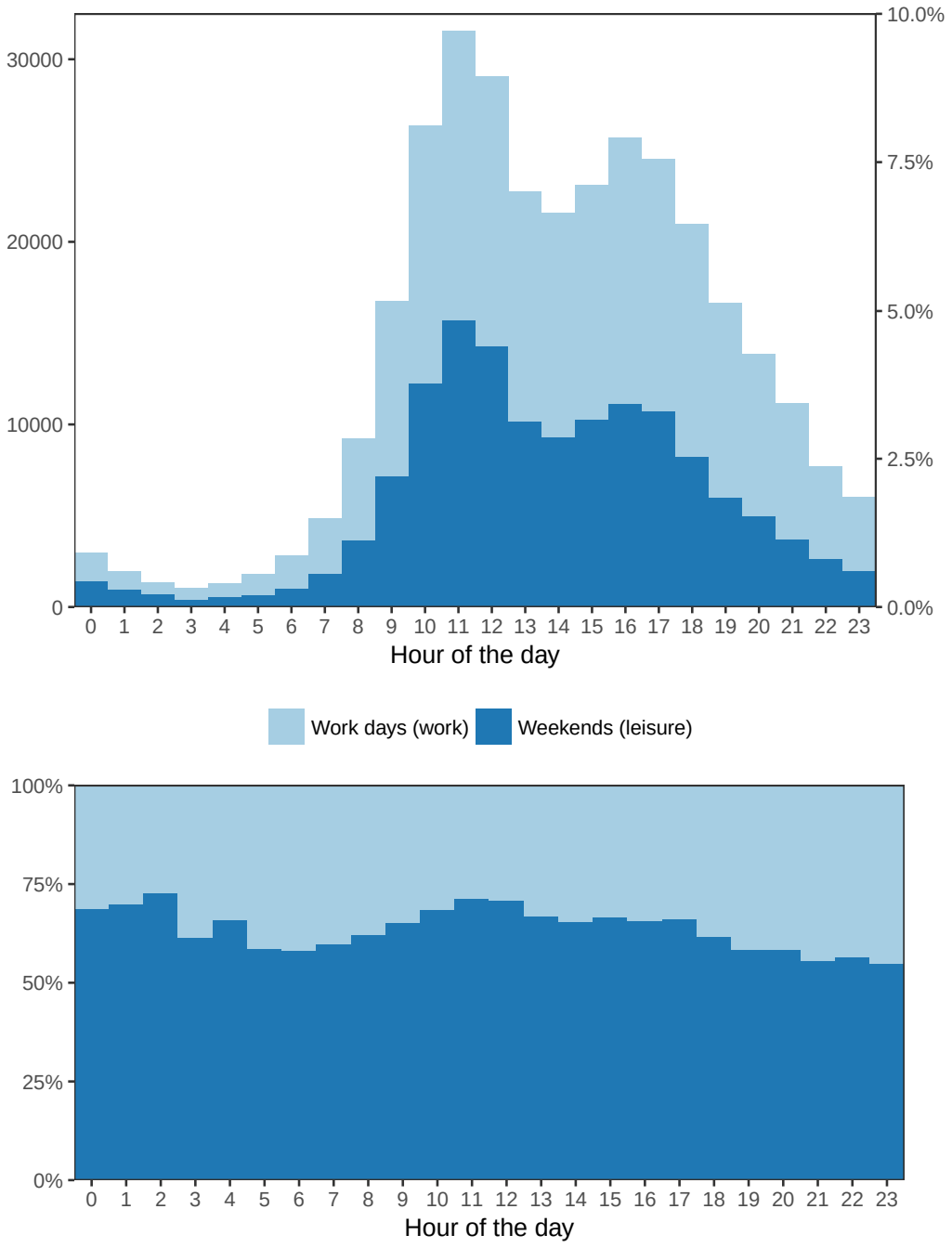


Figure 9.10: Hourly work day and weekend distribution of the most reliable time stamps of the geotagged pictures of Barcelona collected from Flickr. Top figure shows the actual distribution and bottom figure the probability adjusted distribution. Weekday calculations based on the algorithms of the lubridate R package.

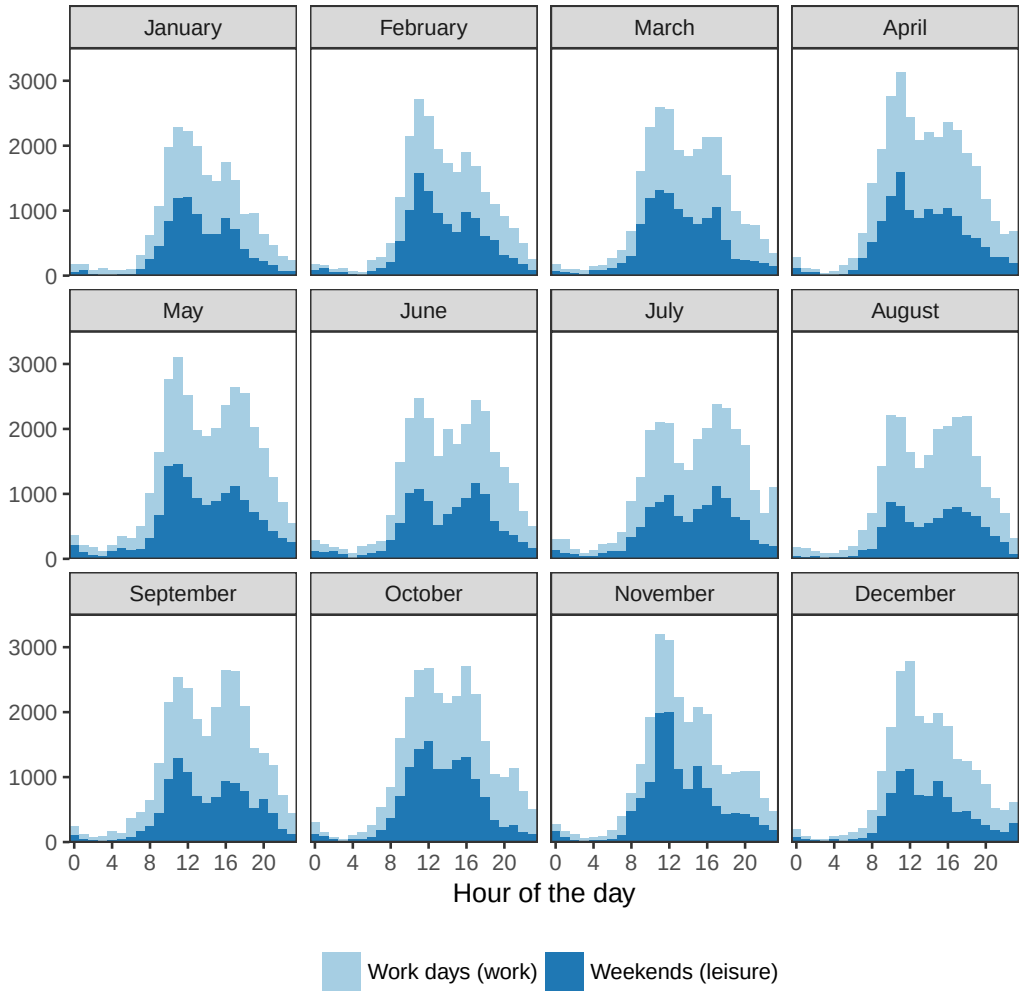


Figure 9.11: Hourly work day and weekend distribution per month of the most reliable time stamps of the geotagged pictures of Barcelona collected from Flickr. Weekday calculations based on the algorithms of the lubridate R package.

was produced (Fig. 9.12), using the same color scheme as figures 9.14 (bottom) and 9.15 to allow easier comparison.

As expected, the number of pictures is smaller in the colder winter months (December, January and February) than in the rest of the months, and gradually increases as the weather gets warmer and the days longer. However, in the summer months, where according to this trend it would be expected that the intensity would be the highest, the peaks of the histograms decrease, suggesting that either there is a smaller picture-taking population in the city or the popularity of other activities overshadows photography in those months.

9.8 The Weekly Cycle

9.8.1 Work and Leisure within the Weekly Cycle

Like the daytime and nighttime distribution in the daily cycle discussed in section 9.7.1, the work and leisure distribution is also biased in the weekly cycle, as more days in the week are work days rather than weekends.

Unlike the daily and yearly cycles, the weekly cycle does not correspond to a natural cycle derived from the apparent position of the Sun from a point in the surface of the Earth, but obeys to the organization of western societies into work days (Monday to Friday) and leisure days (weekends).

In addition, the weekly cycle repeats monotonically and is not synchronized with any other cycle, and therefore over the span of the collected data some irregularities can influence the results of the analysis:

- Over the years, the distribution of days of the week across months is variable, as some of them have 4 weekends and others 5.
- Some non-working days fall in different weekdays over the years, and when falling in Tuesday or Thursday it is customary for many people to take a long weekend including the corresponding Monday of Friday, respectively.
- Some months are paid holidays for most of the population (July and August), and therefore the influence of the work/leisure cycle is diminished.

Regardless of these issues, the collected span of over 13 years was considered sufficiently wide (Fig. 3.3), and weeks are short enough that these differences cancel each other when analyzed globally. Its spatial distribution per neighborhood is discussed in detail in section 7.4.2, while this section focuses in its temporal aspects only.

However, the influence of super-users had the capacity to distort the results, and therefore pictures from users whose picture count exceeded six standard deviations from the sample mean were discarded (as discussed in section 9.3.4).

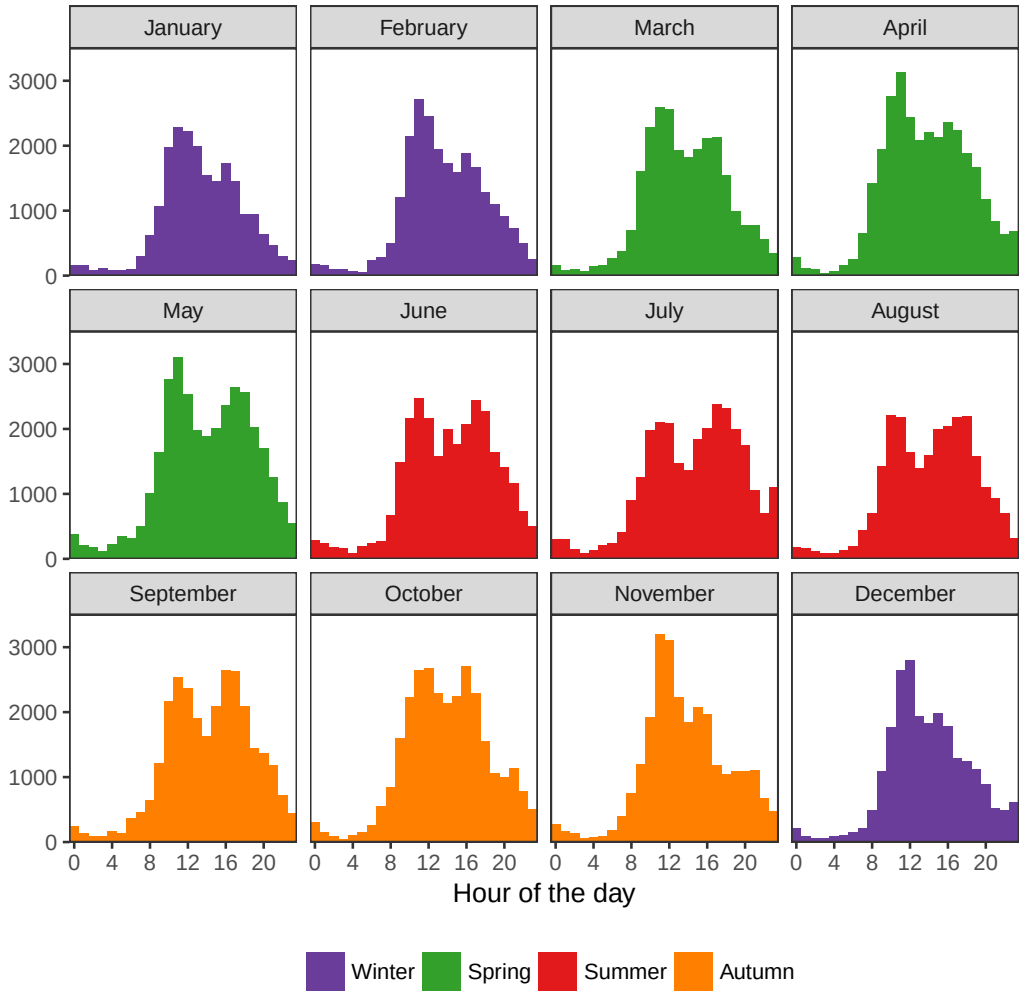


Figure 9.12: Hourly seasonal distribution per month of the most reliable time stamps of the geotagged pictures of Barcelona collected from Flickr. Months are colored according to their meteorological season.

This adjustment diminished the capacity to introduce biases of some users with large picture collections with very specific temporal patterns different from the rest of the users (e.g. taking thousands of pictures every Monday night in the same place), which would not be effectively canceled out by the majority.

The result (Fig. 9.13) shows that the number of pictures taken on weekends is significantly higher, and drops in the work days, which ramp up from the beginning of the week on Monday to the end on Friday, which for many people is the prelude of the weekend and therefore has a leisure aspect, especially in the afternoon.

9.8.2 Daily and Seasonal Cycles within the Weekly Cycle

Since the weekly cycle runs across the whole year and is much smaller (almost two orders of magnitude), and because the sampled period is sufficiently wide to smooth all the possible variations, it was expected that the variation across the daily and seasonal cycles would be negligible (Fig. 9.14).

The results show that the daytime and nighttime distribution of the pictures taken in the course of the 13 years retrieved did not change significantly 9.14a, although it appears that the the proportion of daytime pictures is higher on weekends.

When comparing across seasons the differences are not significant as expected 9.14b, as the proportions of the different seasons remain relatively constant throughout the course of the weekly cycle.

9.9 The Yearly Cycle

9.9.1 Seasons

Because of Earth's axial tilt relative to the ecliptic plane, the earth orbit around the Sun changes —for latitudes other than the equator— the duration of the day and the angle of incidence of the Sun rays. One revolution around the Sun corresponds to one (sidereal) year and during this year cyclic variations of weather occur, according to which the duration of the year is divided into seasons.

At the Barcelona latitude and with a Mediterran Climate Köppen classification there are four distinct seasons: winter, spring, summer and autumn. For the analyses in this research, the meteorological seasons were used, which define seasons as groupings of whole months. Its spatial distribution per neighborhood is discussed in detail in section 7.4.3, while this section focuses in its temporal aspects only.

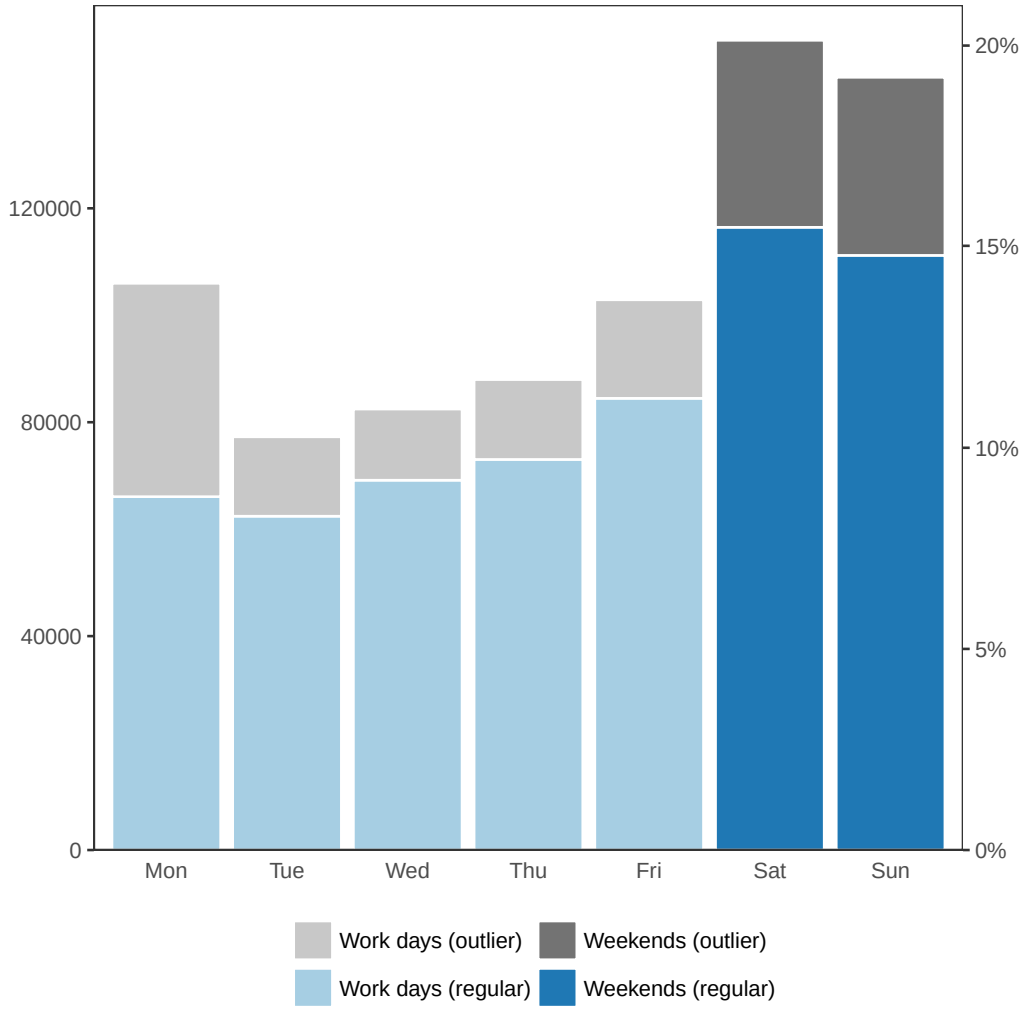
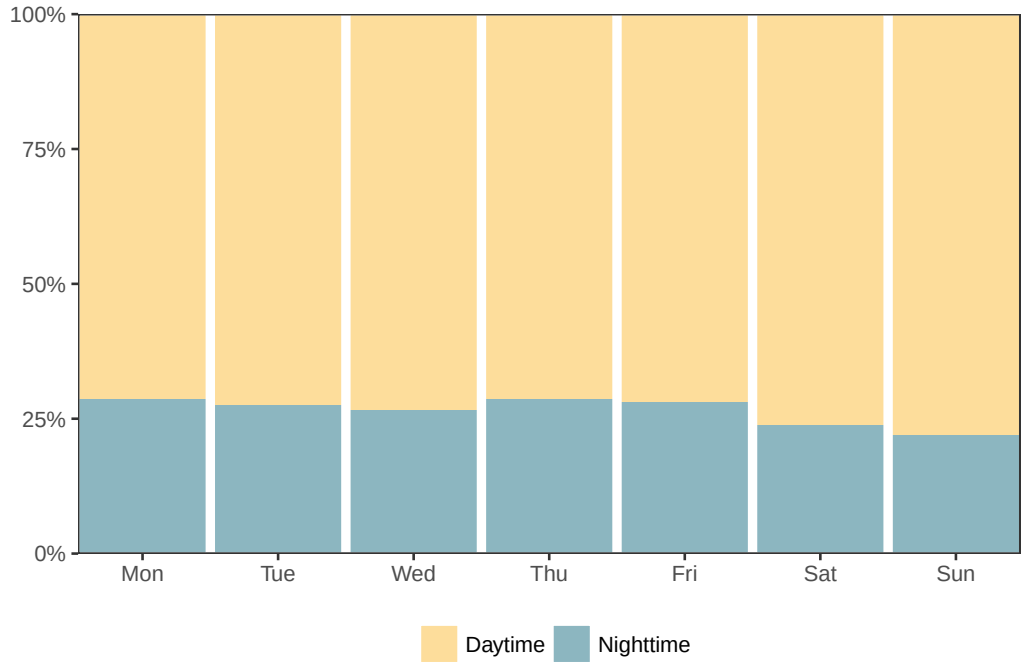
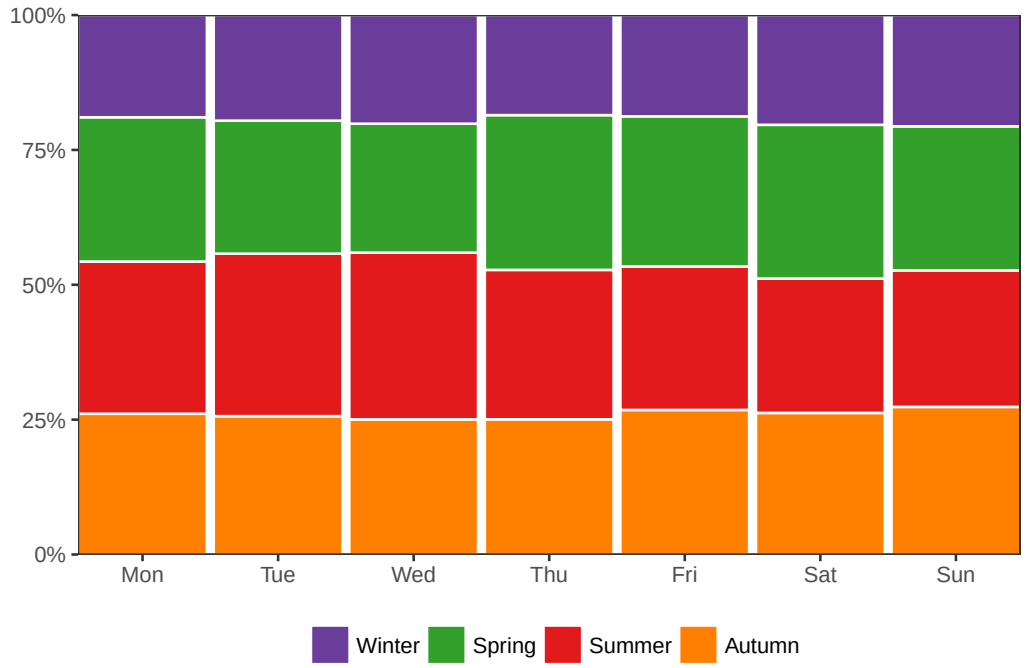


Figure 9.13: Distribution of the time stamps of the geotagged pictures of Barcelona collected from Flickr per week day, classified between leisure days (weekends) and work days (Monday to Friday). Super-user contribution, defined as users whose number of pictures taken exceeded six times the standard deviation of the sample, are represented in gray colors.



(a) Weekly daytime and nighttime distribution, with probability-adjusted frequencies according to the length of the day.



(b) Weekly seasonal distribution.

Figure 9.14: Weekly distribution of the time stamps of the geotagged pictures of Barcelona collected from Flickr, classified according to the daily cycle (top) and the seasonal cycle (bottom). Daytime and nighttime calculations at the nominal latitude of Barcelona based on the algorithms provided by NOAA through the sun methods in the maptools R package.

It was expected that the activity would follow this climate pattern, as the act of taking a photograph is more likely if there is more available light or the day is longer, providing better opportunities to take a picture. The temperature also plays a role as warmer weather encourages outdoor activities.

This pattern was partially confirmed in the analysis, where the intensity rises and falls according to the seasons throughout the year, with the exception of summer days (Fig. 9.15), a pattern also observed in section 9.7.4 and further discussed in section 9.9.2.

9.9.2 Work and Leisure within the Yearly Cycle

This dip in intensity in summer could be explained —as discussed before— because on summer weekends the residents of Barcelona tend to leave the city, and also because in August many locals travel to domestic or abroad destinations, and the tourist influx cannot compensate their picture-taking activities.

This was partially confirmed analyzing the yearly cycle in conjunction with the weekly pattern of work days and weekends (Fig. 9.16), where the weekends in summer have their activity reduced.

Computing the probability-adjusted frequencies, and considering the relative probabilities of work days and weekends, this phenomenon appears more clearly (Fig. 9.17) and a similar effect is visible (but not as noticeable) in the winter vacations in December.

9.9.3 Daytime and Nighttime within the Yearly Cycle

The analysis of the variation of the daily cycle across the year highlights the relevance of weighting the results according to their expected value instead of using raw counts.When using raw counts, the results suggest there is an abundance nighttime photography in the winter days compared to the warmer seasons (Fig. 9.18), but using probability-adjusted frequencies shows a different picture.

In the top figure, a dot is placed according to the relative length of day and night of a typical year for reference. Visually the two proportions seem to run parallel, showing a larger proportion of daytime pictures as discussed in previous sections.

The bottom figure, shows the same data using weighted proportions, where the March and September bars have roughly the same height in both subfigures —because of their respective equinoxes—, resulting in a fairly uniform distribution.

The results suggest that there is not a significant difference in the daytime and nighttime proportions along the year beyond what can be explained solely because

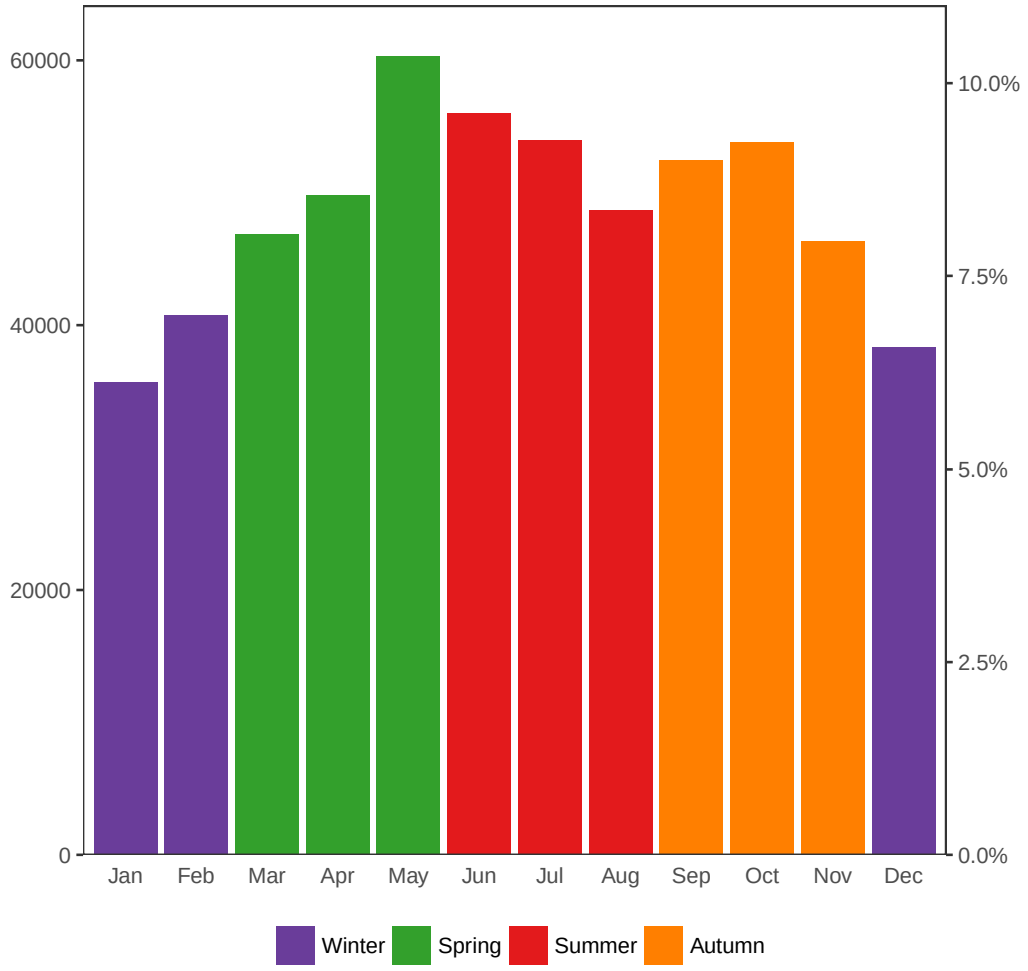


Figure 9.15: Monthly distribution of the time stamps of the geotagged pictures of Barcelona collected from Flickr, classified according to the meteorological season when they were taken.

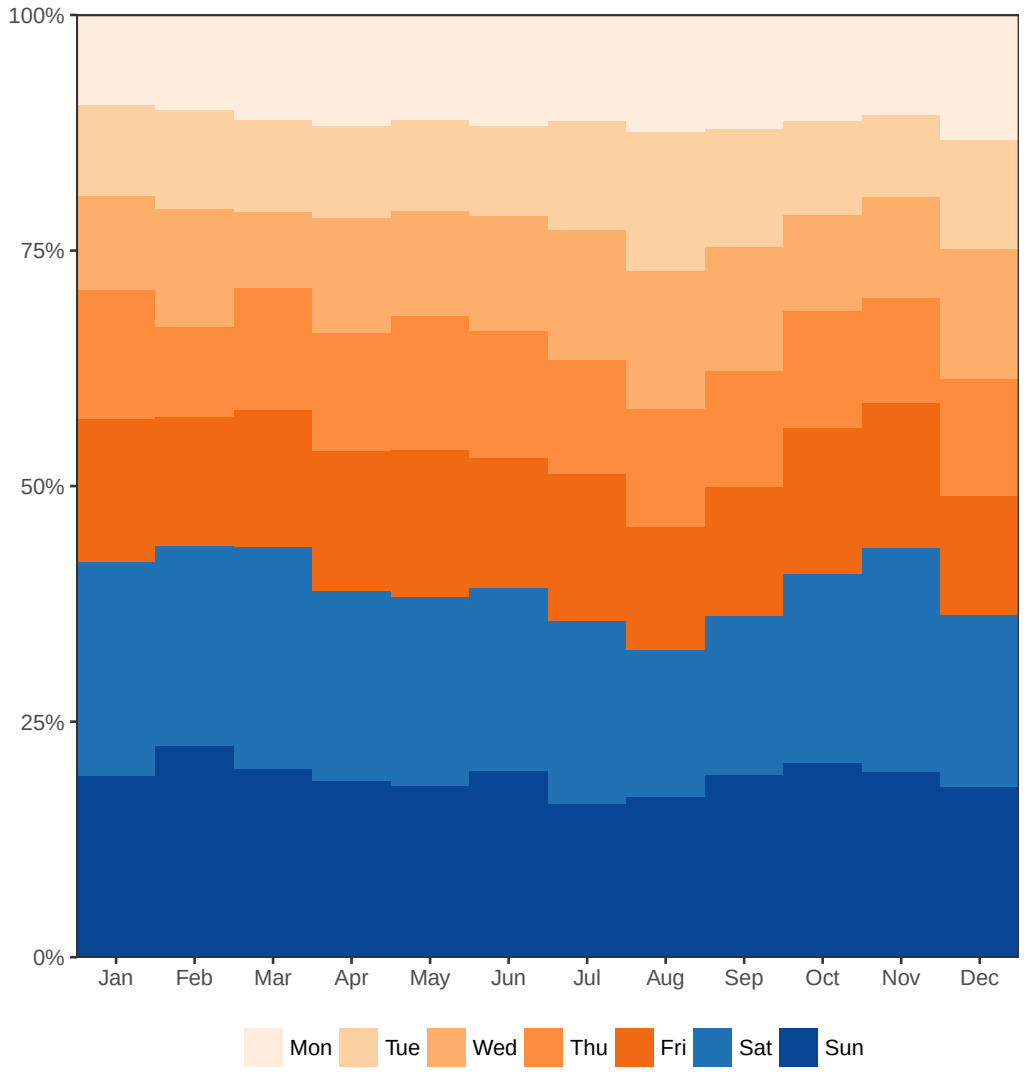


Figure 9.16: Monthly distribution of the time stamps of the geotagged pictures of Barcelona collected from Flickr, classified according to the week day they were taken. Work days are in orange shades and weekends in blue shades.

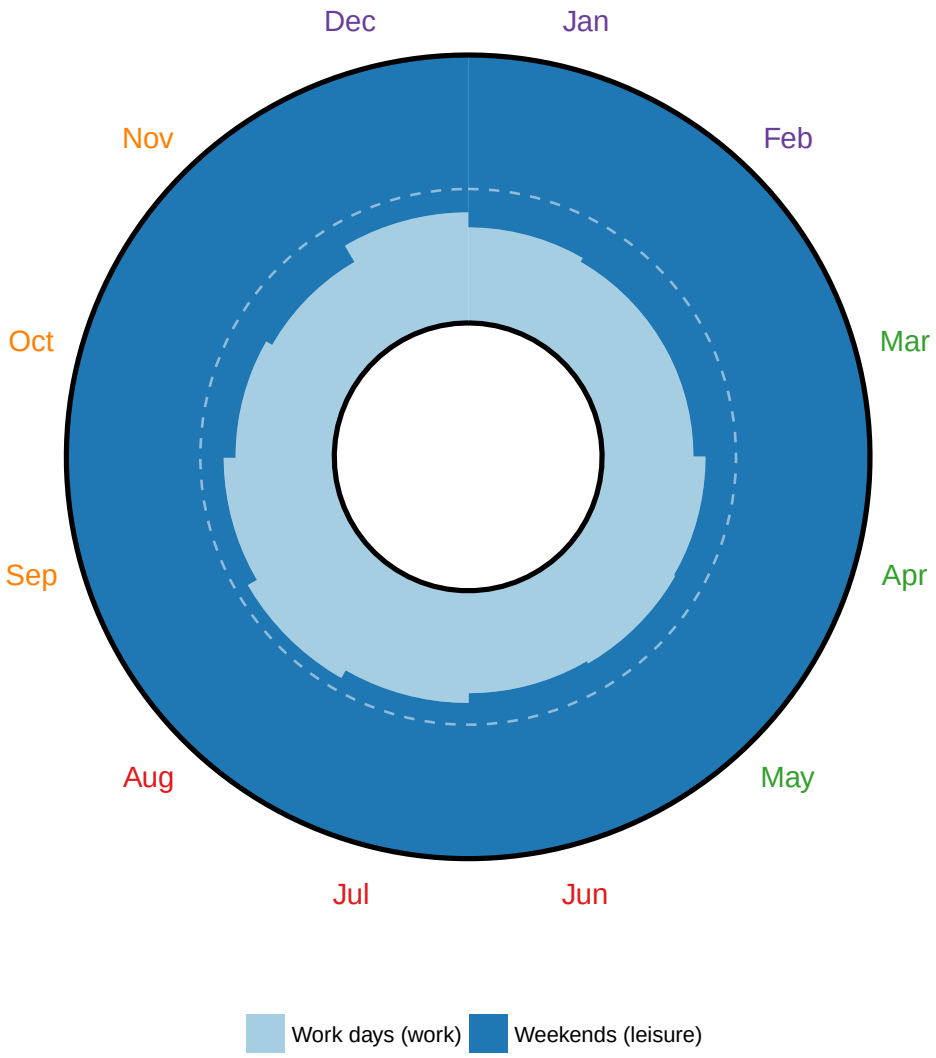


Figure 9.17: Monthly distribution of the time stamps of the geotagged pictures of Barcelona collected from Flickr, classified according the day of the week they were taken as work days (light blue) or weekends (dark blue), using the probability-adjusted frequencies, with the dashed line marking the equal adjusted probability.

of the duration of the days.

9.10 Cycle Interaction

9.10.1 Cycle Heatmaps

The conventional approach to visualizing bivariate distributions is using a scatter plot. However, when the number of coordinate pairs is very high some challenges arise because of overplotting, as many data points are drawn on top of each other, making them indistinguishable (discussed for spatial data in section 6.2.1).

An alternative approach is using heatmaps, which are the generalization of a 1D histogram into the two-dimensional plane. In a heatmap, the data points are binned into squares or rectangles and the number of data points are counted.

In the cycle heatmaps discussed in this chapter, both axes were measured in units of time. In this case binning used the most sensible units available, according to the nature of each cycle:

- 12 months for the yearly cycle.
- 7 week days for the weekly cycle.
- 24 hours for the daily cycle (and on some occasions 1,440 minutes).

Unlike a one-dimensional histogram where the height of the bar can encode the frequency, in a heatmap both the x and y coordinates are used for the position of each tile, and therefore the tile color is generally used to represent the number of points in each bin, using a continuous color ramp.

Therefore, the choice of the color ramp is not trivial, as it should be perceptually uniform, while being perceived by readers with color blindness. For the results shown, the viridis color palette was used, using the R package viridis 0.4.0 [308], a port of the corresponding color ramp in the matplotlib[33] Python plotting library.

9.10.2 Interaction across Days and Months

The interplay of the daily cycle (daytime and nighttime) and the yearly cycle (four seasons in the climate zone of Barcelona) can be visualized using a heatmap, with divisions of the day (hours or minutes) in one axis, and the months of the year in the other (Fig. 9.19).

While the resulting diagram is rectangular, it should be imagined as the development of the surface of a torus, as January and December are topologically next to each other, as well as 23:59 and 00:00 of next day.

[33]The matplotlib library is available at *http://matplotlib.org/* at the time of writing.

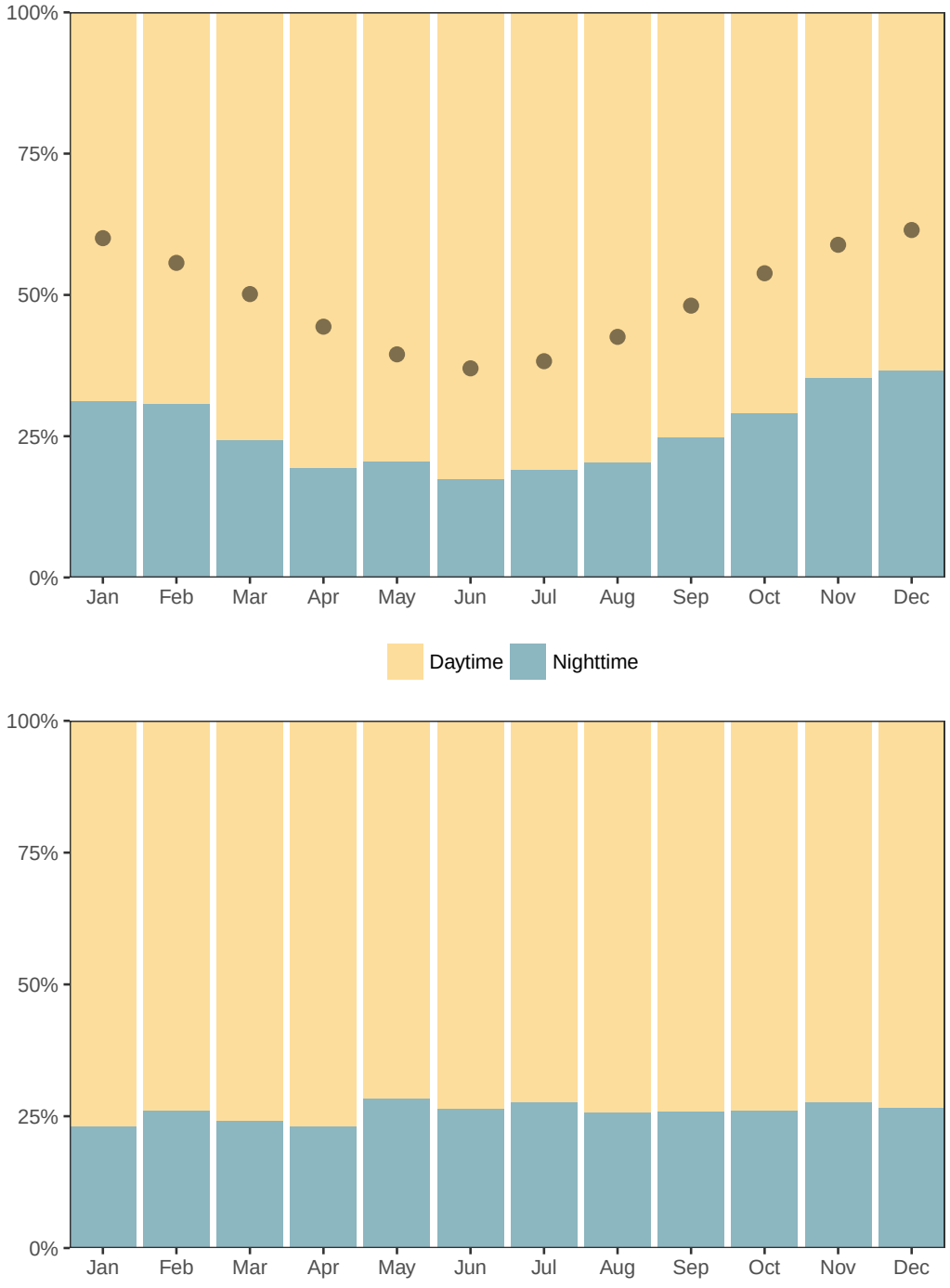


Figure 9.18: Monthly distribution of the time stamps of the geotagged pictures of Barcelona collected from Flickr, classified as daytime or nighttime. Top figure shows the counts and the expected value of a random process (dot), bottom figure shows the same data with adjusted probabilities according to the length of the day they were taken. Daytime and nighttime calculations at the nominal latitude of Barcelona based on the algorithms provided by NOAA through the sun methods in the maptools R package.

Both cycles are intimately related, as the duration of day and night varies across the length of the year. Therefore it would be expected that the heatmap was roughly symmetric along the axes corresponding to either solstice.

The heatmap approximates this expected distribution (Fig. 9.19), and also shows the slightly reduced activity on summer days, as a consequence of the reduced population during this part of the year, especially July and August.

When the resolution is increased from hours (Fig. 9.19a) to minutes (Fig. 9.19b), the outliers are more visible and the heatmap becomes noisier , but the overall pattern is still visible, especially if the count data is square-root transformed.

9.10.3 Interaction across Days and Weeks

While the weekly cycle is not derived from natural phenomena —in contrast with the daily and yearly cycles—, it is somewhat synchronized with the daily cycle, as it consists on a grouping —albeit of arbitrary length— of 7 consecutive days.

Therefore, a heatmap of the interplay of the daily and weekly cycles should allow visualizing the behavioral changes across the week in the temporal patterns throughout the day, grouping the time stamps according to the divisions of the day (hours or minutes) and the seven days of the week (5 work days and 2 leisure days).

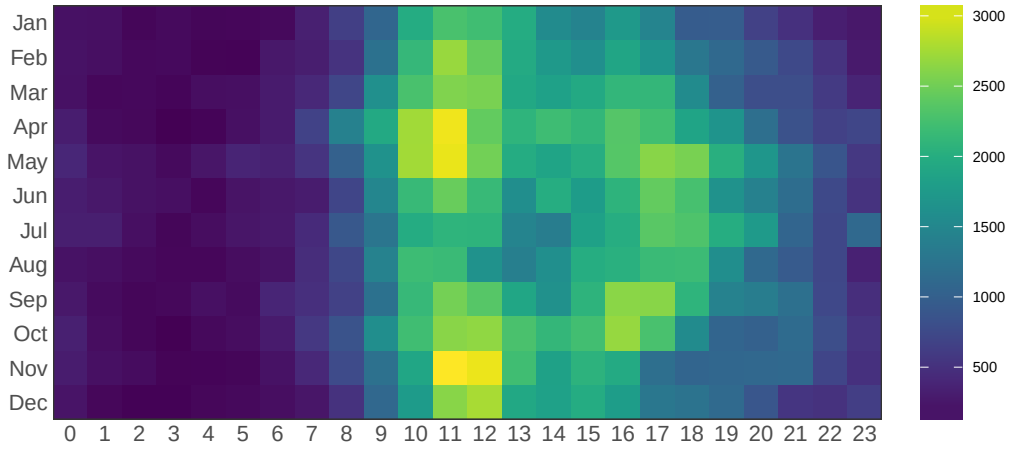
The resulting heatmap (Fig. 9.20) shows that the intensity is significantly higher on weekends, especially in the central part of the day and on the Saturday afternoons. The reduced activity during the lunch hours (Fig. 9.20a) is very visible across all days of the week, as well as the increased activity on Friday afternoon, compared to the rest of the work days.

When square-root transforming the count data aggregated in one minute intervals (Fig. 9.20b), the heatmap becomes noisier, but retains the overall pattern exhibited when aggregating the time stamps in one hour intervals.

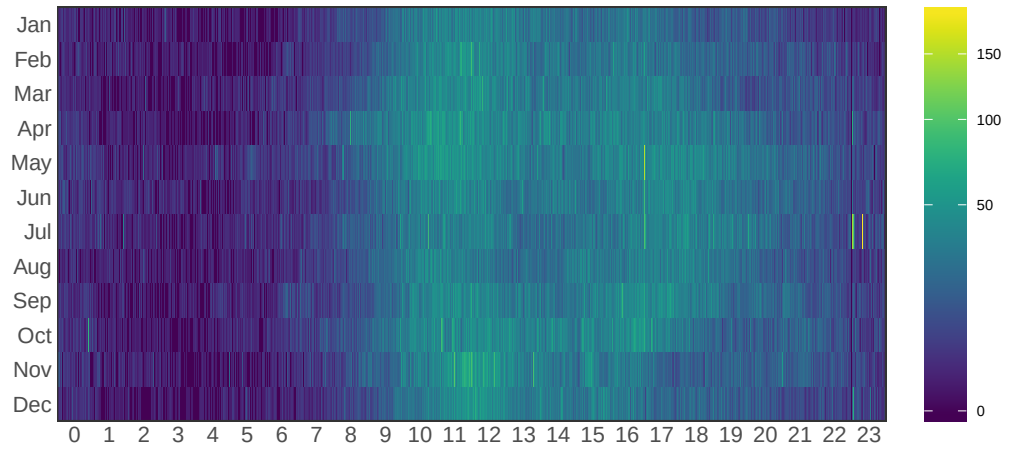
9.10.4 Interaction across Weeks and Months

Plotting a heatmap relating the week days and months allows visualizing the changes of the intensity of the weekly cycle across the duration of the year. Unlike the day/year (section 9.10.2) and day/week (section 9.10.3) interactions discussed before, the result focuses on macro-level variations.

In the resulting heatmap (Fig. 9.21), the variation across the week depending on whether a day is a work day or belongs to the weekend is clearly visible. In the heatmap, the strip that correspond to Fridays becomes a transition boundary as this week day has both work and leisure elements.

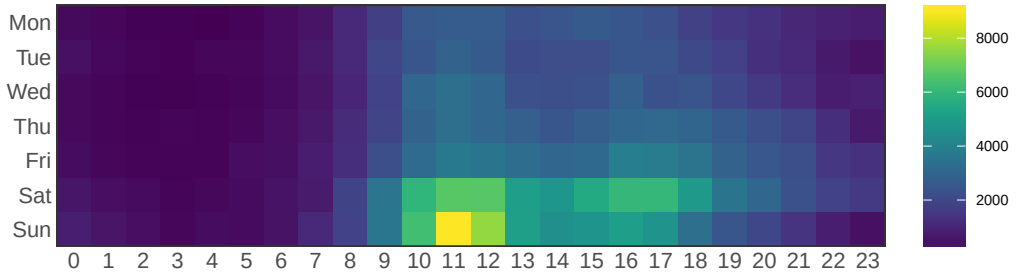


(a) Time stamps binned per month on the vertical axis and per hour on the horizontal axis.

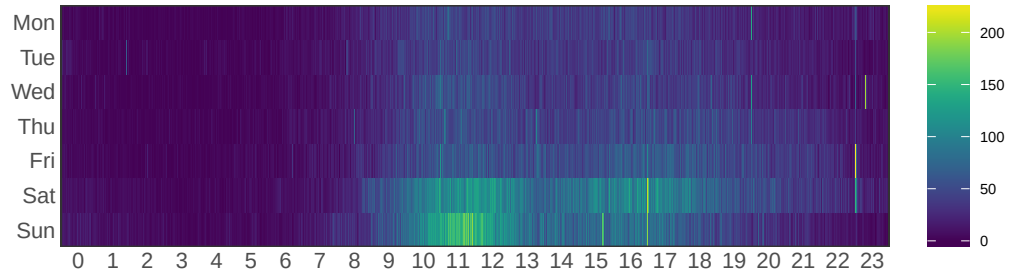


(b) Time stamps binned per month on the vertical axis and per minute on the horizontal axis.

Figure 9.19: Heatmaps of the timestamps of the geotagged pictures of Barcelona collected from Flickr across the daily and yearly cycles. Values in the bottom figure are square root transformed.



(a) Time stamps binned per week day on the vertical axis and per hour on the horizontal axis.



(b) Time stamps binned per week day on the vertical axis and per minute on the horizontal axis.

Figure 9.20: Heatmaps of the timestamps of the geotagged pictures of Barcelona collected from Flickr across the daily and weekly cycles. Values in the bottom figure are square root transformed.

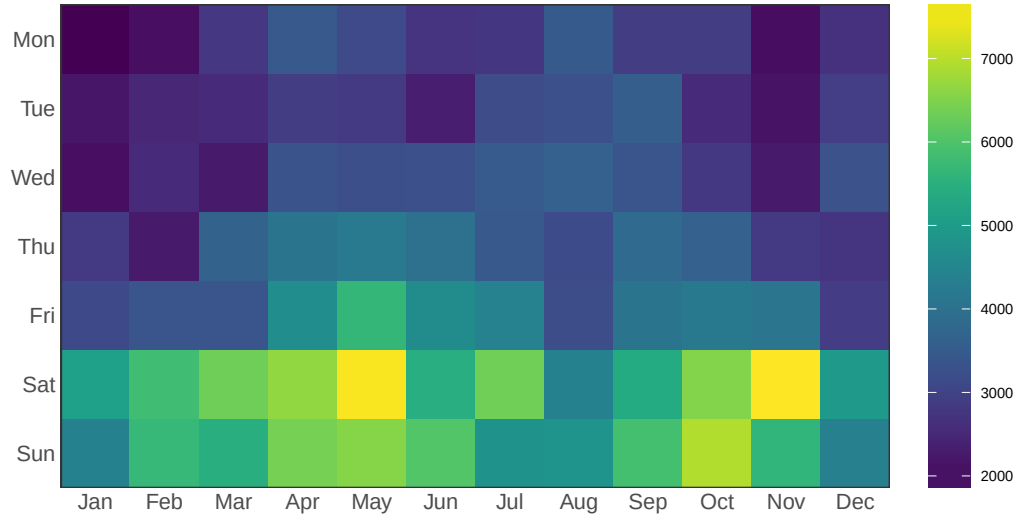


Figure 9.21: Heatmap of the timestamps of the geotagged pictures of Barcelona collected from Flickr across the daily and yearly cycles.

In the figure, the increase in activity around both Spring (March) and Autumn (September) equinoxes, and the activity reduction during the Summer days is also visible (as discussed in section 9.9.1).

Chapter 10

Conclusions

> "We can only see a short distance ahead, but we can see plenty there that needs to be done."
>
> Alan Turing

10.1 Background

This thesis started with the objective of establishing social media data as a new topic in urban research. The research focused in the city of Barcelona as a case of study, and proposed a multifaceted approach to extract knowledge from publicly shared user-generated content. In the introduction, this aim was divided into objectives within four main areas (section 1.2.2):

1. to research the suitability of urban data (section 10.2.1),
2. to develop the corresponding retrieval methodologies (section 10.2.2),
3. to the investigate new methodologies to handle urban data (section 10.2.3),
4. to establish the most appropriate visualization strategies (section 10.2.4).

The focus of the research was on the methodology of the analysis pipeline, from the collection of data to the visualization of the results, discussing the necessary data wrangling required in the process. Therefore, the thesis is more methodological than explanatory and might open more questions than it closes answers, which should be pursued in further research.

10.2 Summary

10.2.1 Researching Urban Data

Urban research has been encumbered by the scarcity of spatial and temporal data appropriate for its scale of analysis. However, the generalization of smartphone use and pervasive communication technologies, along with the production and distribution of user-generated content of Web 2.0 services provide unique opportunities to gather data on user behavior.

One of the aims of the thesis was investigating the suitability of data publicly shared on social networks, using the city of Barcelona as its case of study. The research focused on data with a marked spatial component —primarily picture content, but also short textual messages—, which on many occasions were accompanied with temporal, semantic or multimedia data, along with shared user metadata.

While the production of this user-generated content is definitely not intended to contribute urban data, it provides a unique opportunity for urban research, albeit with important limitations that must be taken into account. One of the chief limitations is the unavoidable self-selection bias of the users, which can be turned into an advantage provided that it is aligned with the object of the research.

Therefore, if we consider users as participants taking part in a natural experiment, large observational studies of behavior can be conducted, within significant temporal spans in broad geographic areas, involving a large number of participants and collecting a substantial quantity of very diverse data, which would have been otherwise impossible or very costly to obtain.

These user-generated data, along with the dissemination of institutional data through open data initiatives, enable new data-driven approaches in urban research. However, they should not substitute but rather complement existing urban research practices, and in the words of Miller and Goodchild [95, p. 460] "we must avoid a data dictatorship: data-driven research should support, not replace, decision-making by intelligent and skeptical humans.".

10.2.2 Collection Methodologies

Urban studies have relied for decades first and foremost on institutional data, which is generally collected using a defined methodology with a specific periodicity. In contrast, social media is produced in real-time, without the requirements of urban researches in mind.

Because of the dynamic and ephemeral nature of the Internet, protocols to access data from online services are continuously evolving, and the corresponding

web APIs and data structures change to accommodate the addition, removal or modification of features according to the business decisions of the data provider.

During the development of the collection methodologies during research, it become apparent that API implementations can oftentimes be poorly documented, with outdated or incomplete details scattered across multiple locations, and as a consequence accessing data from these services can become a moving target, and even in some occasions the services can unexpectedly cease operations as in the case of Panoramio.

Therefore, tools developed to access social media data can be expected break without notice at any time, hindering research —particularly in long-term investigations— and hampering reproducibility because of the difficulty in replicating results. These issues arose during the development of the research and are thoroughly discussed in the thesis, but the solutions provided are unavoidably temporary.

However, the value of the retrieved data makes the effort invested worthwhile, even more as these sources of information become pervasive and are increasingly included in the urban research toolbox, although this situation is expected to improve in the future with the advent of the semantic web [393].

The research evidenced that there is not a one-size-fits-all source of information, as different services serve specific purposes and therefore the type of content publicly shared and their user bases can be very diverse. The implications for urban research, as the selection of the targeted service can introduce bias and provide inadequate results, require the consideration of the most suitable source depending on the research question.

10.2.3 Handling Urban Data

Government data used in urban research was until recently stored in well-documented formats using a relational structure of tables [394], conveniently downloadable from a single portal. During research it quickly became clear that the standard tools used by practitioners of the discipline —GIS and spreadsheets— were not capable of managing social media data effectively. The main issues found can be classified into two categories: the amount of data to handle and how these data are structured.

The most outstanding issue was the number of records to process, invalidating some widely used spreadsheets like Microsoft Excel because of its limitation of roughly one million rows[1]. Moreover, even in cases when the software is capable

[1]Microsoft Excel worksheets cannot exceed 1,048,576 rows by 16,384 columns according to *https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3* available at the time of writing.

of *storing* the data, it still needs to be *processed* efficiently, which is challenging for standard GIS packages[2].

While the first issue can be solved technically —via software or hardware—, these tools still require tabular data. However, most of the collected data was typically non-rectangular, stored in formats that required parsing (e.g. JSON, XML) and had to be converted into nested, difficult to navigate tree-like structures[3].

Because of these reasons, one of the earliest conclusions during research was the necessity to adopt a different workflow, avoiding pre-packaged solutions and developing a pipeline capable of handling the data retrieval and transformation processes programmatically[4], allowing access to more powerful data cleaning and wrangling methodologies not available in off-the-shelf tools.

This paradigm shift based in data science best practices had additional advantages making possible much more complex analyses: code reuse, easier debugging of errors, protection from the propagation of mistakes early in the pipeline, improved reproducibility and better documentation of every process.

The advantages afforded by adopting a more script-based and less GUI-based workflow[5] cannot be neglected, and some exposure to these tools should be very beneficial in urban research curricula, as they have the capacity to significantly improve the productivity of researchers and the quality of their analyses.

10.2.4 Effective Visualization

Very early in the research process it become obvious that the abundance and complexity of data could lead to an information overload, and it would be necessary to develop a set of representation strategies to visualize the obtained results.

Data visualization is a powerful strategy to understand patterns in data, using the visual system as an interface to the brain of the analyst. Effective visualization is also crucial to communicate the results from experts to a wider audience alike. Good data representation should therefore provide a better understanding of the city, allow a better communication of the results, and more importantly, suggest new research questions as a powerful exploratory analysis tool.

However, the volume of the collected data and the complexity of urban phenomena required the development of multiple visualization approaches, to make

[2]Without resorting to specialized SQL implementations.

[3]Even when data is rectangular, SQL lacks the expressiveness to conduct very complex analyses, because its declarative paradigm precludes imperative programming.

[4]The development process produced 33,031 lines of code, and the 3,780 commits of the version control system were hosted in 15 repositories on GitHub.

[5]As discussed in Hadley Wickham's provocatively titled talk "You Can't Do Data Science in a GUI" available at *https://www.meetup.com/acm-chicago/events/248060005/* at the time of writing.

sense of the patterns of millions of data points in different scales (from global to neighborhood), topologies (Euclidean and network) and dimensions (spatial and temporal).

The thesis develops a battery of visualization strategies tailored to the variety of representation challenges encountered, after a lengthy research process that discarded many standard techniques that were not effective in the context of Big data, adapting existing methods to the requirements of Big Data in urban research.

These representation challenges should share three common objectives: amplifying the signal and filtering the noise (improving the signal-to-noise ratio), providing the most objective representation of the underlying phenomena while avoiding subjective biases, and communicating the results effectively to the widest audience.

However, these goals are difficult to achieve using traditional maps or diagrams printed on paper, and it seems reasonable that with Big Data should be accompanied with its corresponding "Big Visualization", incorporating new elements such as interactivity, customization or storytelling, to name a few possibilities. This approach can be realized with tools such as the shiny R package in combination with the leaflet library[6] (Fig. 10.1).

10.3 Future Directions

10.3.1 Future Research

Future avenues research should focus on the analysis of the written content of the retrieved data (picture captions and messages), using text mining tools to extract semantic information —already published separately [143]—, as well as on multimedia content, using computer vision to classify the images using deep learning neural networks.

Improvements to the methodology should focus on alternative statistical methods; in particular Bayesian statistics should be investigated as a better inferential foundation than classical (frequentist) approaches used in the research [395], capable of providing more intuitive interpretation of the results using credible intervals or Bayes factors.

[6]The leaflet open-source JavaScript library is available at *http://leafletjs.com/* at the time of writing.

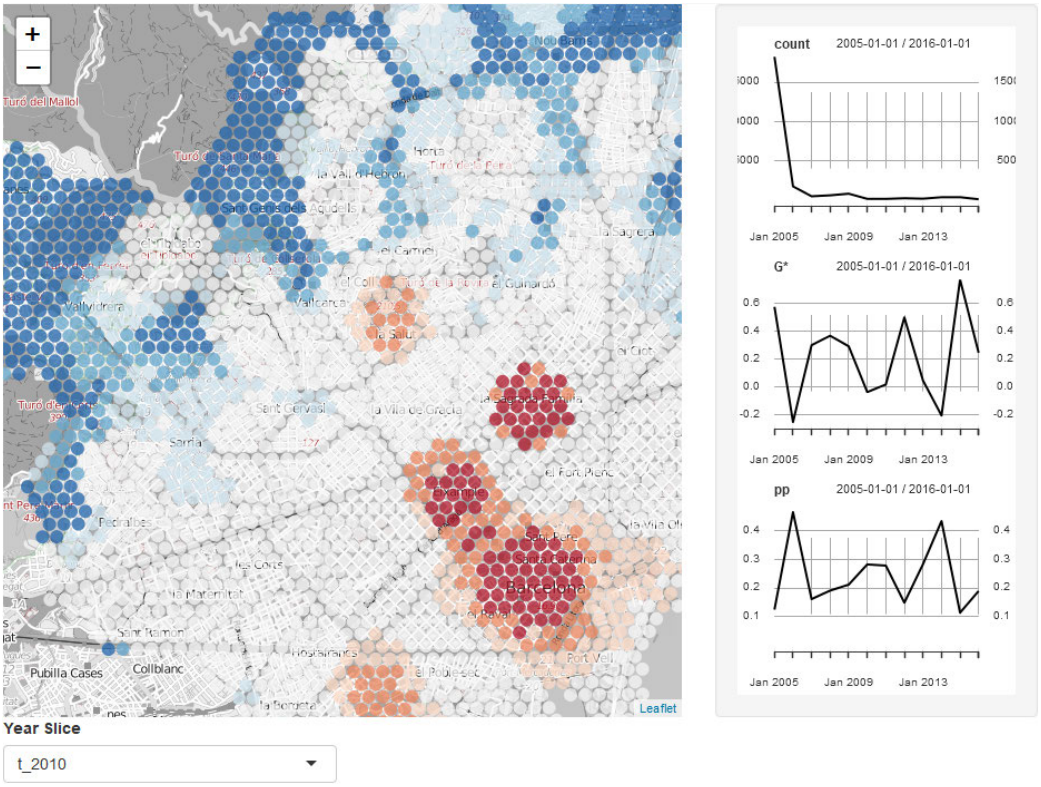


Figure 10.1: Work in progress of an interactive visualization prototype showing a dashboard developed using the shiny R package, that allows zooming and panning any yearly time slice of one of the maps discussed in section 6.3, and clicking on any of the tiles to visualize the corresponding time series.

10.3.2 Future Applications

The research should open new opportunities for all stakeholders in the decision-making process based on facts about the behavior of people in cities, especially if coupled with urban, transportation or pedestrian simulation tools, resulting in better planning policies and designs more suited to the needs of its inhabitants and visitors. An pilot is currently being developed in this context and is currently in the latest stages of development.

Another facet in development is the educational value of these data in the teaching of architecture, which have been already used to improve student engagement [143]. It is expected to be included in a GIS elective to provide architecture students an attractive case of study more aligned with their experience with technology.

Finally, the methodology developed in this thesis is being tested to research the 20 cities besides Barcelona (Table 10.1) in the Mastercard Destination Cities Index [9] as comparison benchmark.

10.3.3 Final Considerations

While the research process has treated all data anonymously and only used publicly available data, the amount and detail of the information that can be collected is overwhelming, raising ethical considerations that should be addressed [396] regarding the privacy of users, especially in the wake of the recent Cambridge Analytica scandal.

This has been summarized with the phrase "if you're not paying for it, you become the product" [397], where services attract users with free services, only to sell the profile data to the highest bidder, granting unprecedented power to some corporations [398].

While we are still in the infancy of the digital revolution, the issues of the abuse of this technology by governments, corporations or malicious individuals must be seriously considered [399].

Table 10.1: Data collected from Flickr on the 20 highest ranked cities in the Mastercard Destination Cities Index.

Position	City	Pictures	Users
1	Bangkok	515,632	13,814
2	London	5,409,107	95,972
3	Paris	2,372,264	67,008
4	Dubai	147,617	7,990
5	Singapore	847,017	18,612
6	New York	4,582,429	87,649
7	Seoul	333,455	6,760
8	Kuala Lumpur	239,523	8,888
9	Tokyo	2,314,698	25,424
10	Istanbul	429,412	16,149
11	Hong Kong	1,075,140	22,331
12	Barcelona	1,166,704	34,283
13	Amsterdam	716,313	25,736
14	Milan	601,247	19,103
15	Taipei	1,890,940	20,425
16	Rome	1,126,483	39,547
17	Osaka	251,237	7,649
18	Vienna	590,330	13,892
19	Shanghai	3,137,522	35,361
20	Prague	476,657	17,787

Appendix: Open Source Tools

The majority of the development was conducted using open source tools. Beyond the obvious cost benefits, the objective was to promote open science and knowledge, as well as to participate in a helpful community and the possibility modifying and contributing code.

The text uses the Linux Libertine[7] digital typeface for the main body of text, and the microtype [400] package for micro-typographic enhancements. The dissertation was written using the LyX[8] 2.2.3 document processor, using the LaTeX 2_ε backend and packages from the MiKTeX[9] 2.9 distribution, and the bibliography was managed using Zotero[10]. The content of the dissertation was managed using the Git[11] distributed version control system and hosted on GitHub.

The development was conducted using the R programming language[12], from versions 3.3.1 to 3.4.3, using RStudio[13] as the coding environment, using a variety of packages for data retrieval, interoperability, wrangling, transformation and visualization (Table 10.2).

Most of the spatial data was processed using specialized R packages (Table 10.3), which were capable of manipulating geometries, perform complex spatial analysis and visualize the results. However, for exploratory analysis GIS software was more convenient, and in these cases the open source QGIS[14] and GRASS[15] were used, with SpatiaLite[16] as the database backend.

[7]The Linux Libertine typeface is available at *http://www.linuxlibertine.org/* at the time of writing.
[8]The LyX document processor is available at *http://www.lyx.org/* at the time of writing.
[9]The MiKTeX project page is available at *http://miktex.org/* at the time of writing.
[10]The Zotero bibliography manager is available at *http://www.zotero.org/* at the time of writing.
[11]The Git project is available at *http://git-scm.com/* at the time of writing.
[12]The R Project is available at *http://www.r-project.org/* at the time of writing.
[13]The RStudio environment is available at *http://www.rstudio.com/* at the time of writing.
[14]The QGIS project is available at *http://qgis.org/* at the time of writing.
[15]The GRASS GIS project is available at *http://grass.osgeo.org/* at the time of writing.
[16]The SpatiaLite library is available at *http://www.gaia-gis.it/gaia-sins/* at the time of writing.

Table 10.2: Main R non-spatial packages used in the research.

Package	Version	Description
broom	0.4.3	Convert Statistical Analysis Objects into Tidy Data Frames
classInt	0.1-24	Choose Univariate Class Intervals
curl	3.1	A Modern and Flexible Web Client for R
data.table	1.10.4-3	Extension of ‘data.frame‘
devtools	1.13.4	Tools to Make Developing R Packages Easier
dplyr	0.7.4	A Grammar of Data Manipulation
forcats	0.2.0	Tools for Working with Categorical Variables (Factors)
ggplot2	2.2.1	Create Elegant Data Visualisations Using the Grammar of Graphics
ggrepel	0.7.0	Repulsive Text and Label Geoms for ’ggplot2’
httr	1.3.1	Tools for Working with URLs and HTTP
jsonlite	1.5	A Robust, High Performance JSON Parser and Generator for R
lubridate	1.7.1	Make Dealing with Dates a Little Easier
magrittr	1.5	A Forward-Pipe Operator for R
pals	1.4	Color Palettes, Colormaps, and Tools to Evaluate Them
purrr	0.2.4	Functional Programming Tools
RColorBrewer	1.1-2	ColorBrewer Palettes
RCurl	1.95-4.10	General Network (HTTP/FTP/...) Client Interface for R
readr	1.1.1	Read Rectangular Text Data
reshape2	1.4.3	Flexibly Reshape Data: A Reboot of the Reshape Package
rtweet	0.6.0	Collecting Twitter Data
rvest	0.3.2	Easily Harvest (Scrape) Web Pages
scales	0.5.0	Scale Functions for Visualization
shiny	1.2.0	Web Application Framework for R
streamR	0.2.1	Access to Twitter Streaming API via R
stringr	1.2.0	Simple, Consistent Wrappers for Common String Operations
tibble	1.4.1	Simple Data Frames
tidyjson	0.2.2	A Grammar for Turning ’JSON’ into Tidy Tables
tidyr	0.7.2	Easily Tidy Data with ’spread()’ and ’gather()’ Functions
treemap	2.4-2	Treemap Visualization
twitteR	1.1.9	R Based Twitter Client
viridis	0.4.1	Default Color Maps from ’matplotlib’
xml2	1.1.1	Parse XML
xts	0.10-1	eXtensible Time Series
zoo	1.8-1	S3 Infrastructure for Regular and Irregular Time Series (Z’s Ordered Observations)

Table 10.3: Main R spatial packages used in the research.

Package	Version	Description
cartogram	0.0.2	Create Cartograms with R
ggmap	2.6.1	Spatial Visualization with ggplot2
GISTools	0.7-4	Some further GIS capabilities for R
gstat	1.1-5	Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation
leaflet	2.0.2	Create Interactive Web Maps with the JavaScript 'Leaflet' Library
mapproj	1.2-5	Map Projections
maptools	0.9-2	Tools for Reading and Handling Spatial Objects
mapview	2.2.0	Interactive Viewing of Spatial Data in R
micromap	1.9.2	Linked Micromap Plots
micromapST	1.1.1	Linked Micromap Plots for General U. S. and Other Geographic Areas
proj4	1.0-8	A simple interface to the PROJ.4 cartographic projections library
raster	2.6-7	Geographic Data Analysis and Modeling
rasterVis	0.43	Visualization Methods for Raster Data
rgdal	1.2-16	Bindings for the 'Geospatial' Data Abstraction Library
rgeos	0.3-26	Interface to Geometry Engine - Open Source ('GEOS')
rgrass7	0.1-10	Interface Between GRASS 7 Geographical Information System and R
rmapsharper	0.3.0	Client for 'mapshaper' for 'Geospatial' Operations
rnaturalearth	0.1.0	World Map Data from Natural Earth
sf	0.6-0	Simple Features for R
sp	1.2-7	Classes and Methods for Spatial Data
spacetime	1.2-1	Classes and Methods for Spatio-Temporal Data
spatstat	1.54-0	Spatial Point Pattern Analysis, Model-Fitting, Simulation, Tests
spdep	0.7-4	Spatial Dependence: Weighting Schemes, Statistics and Models
stpp	2.0-3	Space-Time Point Pattern Simulation, Visualisation and Analysis
tmap	1.11	Thematic Maps

References

[1] United Nations Human Settlements Programme (UN-HABITAT), *State of the World's Cities 2010/2011 - Cities for All: Bridging the Urban Divide*, ser. State of the World's Cities. London: Earthscan, 2010, no. 1249/09. [Online]. Available: http://mirror.unhabitat.org/pmss/getElectronicVersion.aspx?nr=2917&alt=1

[2] L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert, and G. B. West, "Growth, innovation, scaling, and the pace of life in cities," *Proceedings of the National Academy of Sciences*, vol. 104, no. 17, pp. 7301–7306, Apr. 2007. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0610172104

[3] D. Offenhuber and C. Ratti, Eds., *Decoding the City: Urbanism in the Age of Big Data*, 1st ed. Basel: Birkhauser Verlag AG, Aug. 2014.

[4] P. Queraltó and F. Valls, "Herramienta de cálculo de rutas óptimas según parámetros de accesibilidad física en itinerarios urbanos," *ACE: Architecture, City and Environment*, vol. 5, no. 13, pp. 161–184, Jun. 2010. [Online]. Available: http://hdl.handle.net/2099/9204

[5] F. Valls, "Metodología para la evaluación de rutas óptimas en entornos urbanos a partir de datos de LTS," Masters Research thesis, Universitat Politècnica de Catalunya, Jul. 2010. [Online]. Available: http://hdl.handle.net/2099.1/11696

[6] F. Valls, P. Garcia-Almirall, E. Redondo, and D. Fonseca, "From Raw Data to Meaningful Information: A Representational Approach to Cadastral Databases in Relation to Urban Planning," *Future Internet*, vol. 6, no. 4, pp. 612–639, Oct. 2014. [Online]. Available: http://www.mdpi.com/1999-5903/6/4/612

[7] F. Valls and J. Roca, "Herramienta de visualización de rutas accesibles en espacios urbanos utilizando tecnología HTML5," *ACE: Architecture, City and Environment*, vol. 11, no. 33, pp. 251–264, Feb. 2017. [Online]. Available: http://revistes.upc.edu/ojs/index.php/ACE/article/view/5138

[8] F. Valls, E. Redondo, and D. Fonseca, "E-Learning and Serious Games: New Trends in Architectural and Urban Design Education," in *Learning and Collaboration Technologies*, ser. Lecture Notes in Computer Science, P. Zaphiris and A. Ioannou, Eds., vol. 9192. Los Angeles, CA, USA: Springer International Publishing Switzerland, Aug. 2015, pp. 632–643. [Online]. Available: http://hdl.handle.net/2117/86762

[9] R. Erenhouse, "Mastercard Global Destination Cities Index," Mastercard, Tech. Rep., Sep. 2017. [Online]. Available: http://news.mstr.cd/gdci2017

[10] W. Geerts, "Top 100 City Destinations Ranking," Euromonitor International, Tech. Rep., Nov. 2017. [Online]. Available: http://go.euromonitor.com/Top_100_City_Destinations_WTM_Form_Download.html

[11] S. Paldino, I. Bojic, S. Sobolevsky, C. Ratti, and M. C. González, "Urban magnetism through the lens of geo-tagged photography," *EPJ Data Science*, vol. 4, no. 1, p. 5, May 2015. [Online]. Available: http://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-015-0043-3

[12] M. d. Solà-Morales i Rubió, *Ten lessons on Barcelona: urbanistic episodes that have made the modern city*. Barcelona: Col·legi d'Arquitectes de Catalunya, 2009, oCLC: 733868620.

[13] M. K. Saberi, A. Isfandyari-Moghaddam, and S. Mohamadesmaeil, "Web Citations Analysis of the JASSS: the First Ten Years," *Journal of Artificial Societies and Social Simulation*, vol. 14, no. 4, 2011. [Online]. Available: http://jasss.soc.surrey.ac.uk/14/4/22.html

[14] R. P. Dellavalle, E. J. Hester, L. F. Heilig, A. L. Drake, J. W. Kuntzman, M. Graber, and L. M. Schilling, "Going, Going, Gone: Lost Internet References," *Science*, vol. 302, no. 5646, pp. 787–788, Oct. 2003. [Online]. Available: http://science.sciencemag.org/content/302/5646/787

[15] C. P. Snow, *The Two Cultures and a Second Look: An Expanded Version of The Two Cultures and the Scientific Revolution*. Cambridge: Cambridge University Press, 1964.

[16] J. Portugali, *Complexity, cognition and the city*, ser. Understanding complex systems. Berlin: Springer, 2013, oCLC: 864536162.

[17] J. H. Von Thünen, *Der isolierte Staat (The Isolated State)*. Pergamon, 1826.

[18] C. Baudelaire, *Les Fleurs du Mal*, 1857. [Online]. Available: http://www.gutenberg.org/ebooks/6099

[19] C. Sitte, *City Planning According to Artistic Principles.* New York: Random House, 1889.

[20] W. Christaller, *Central Places in Southern Germany.* Englewood Cliffs, NJ: Prentice Hall, 1933.

[21] W. J. Reilly, *The Law of Retail Gravitation.* New York: Knickerbocker Press, 1931.

[22] L. Wirth, "Urbanism as a Way of Life," *American Journal of Sociology*, vol. 44, no. 1, pp. 1–24, 1938. [Online]. Available: http://www.jstor.org/stable/2768119

[23] L. Mumford, *The City in History: Its Origins, Its Transformations, and Its Prospects.* Harcourt, Brace & World, 1961.

[24] J. Jacobs, *The Death and Life of Great American Cities.* New York: Random House, 1961.

[25] C. Alexander, "A City is not a Tree," *Architectural Forum*, vol. 122, no. 1, pp. 58–62, 1965. [Online]. Available: http://www.rudi.net/pages/8755

[26] K. Lynch, *The Image of the City.* Cambridge (MA): The MIT Press, Jun. 1960.

[27] I. Burton, "The Quantitative Revolution and Theoretical Geography," *Canadian Geographer / Le Géographe canadien*, vol. 7, no. 4, pp. 151–162, Dec. 1963. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0064.1963.tb00796.x/abstract

[28] A. Lösch, *The Economics of Location.* New Haven: Yale Univ Press, 1954.

[29] W. Alonso, *Location and Land Use: Toward a General Theory of Land Rent.* Cambridge: HUP, 1963, oCLC: 935284812.

[30] H. Lefebvre, *La Révolution Urbaine.* Paris: Gallimard, 1970.

[31] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Mansfield Centre, CT: Martino Fine Books, Jun. 2012.

[32] G. B. Dantzig and T. L. Saaty, *Compact city; a plan for a liveable urban environment.* San Francisco: W. H. Freeman, 1973.

[33] M. Batty, "Competition in the Built Environment: Scaling Laws for Cities, Neighbourhoods and Buildings," *Nexus Network Journal*, pp. 1–20, Aug. 2015. [Online]. Available: http://link.springer.com/article/10.1007/s00004-015-0268-2

[34] M. Schlapfer, L. M. A. Bettencourt, S. Grauwin, M. Raschke, R. Claxton, Z. Smoreda, G. B. West, and C. Ratti, "The scaling of human interactions with city size," *Journal of The Royal Society Interface*, vol. 11, no. 98, pp. 20 130 789–20 130 789, Jul. 2014. [Online]. Available: http://rsif.royalsocietypublishing.org/cgi/doi/10.1098/rsif.2013.0789

[35] R. Cervero, *America's suburban centers: the land use-transportation link*. Boston: Unwin Hyman, 1989.

[36] M. Batty and P. Longley, *Fractal Cities: A Geometry of Form and Function*, 1st ed. Academic Press, Aug. 1994. [Online]. Available: http://www.fractalcities.org/

[37] S. Marshall, *Cities Design and Evolution*. Abingdon, Oxon ; New York, NY: Routledge, 2009.

[38] ——, *Streets and Patterns*, 1st ed. London ; New York: Routledge, Dec. 2004.

[39] J. Portugali, *Self-Organization and the City*, ser. Springer Series in Synergetics. Berlin Heidelberg: Springer-Verlag, 2000.

[40] M. Castells, *The rise of the network society*, ser. Information age. Oxford ; Malden, Mass: Blackwell Publishers, 1996.

[41] N. A. Salingaros, "Theory of the urban web," *Journal of Urban Design*, vol. 3, no. 1, pp. 53–71, 1998.

[42] G. MacLeod and M. Jones, "Renewing the Geography of Regions," *Environment and Planning D: Society and Space*, vol. 19, no. 6, pp. 669–695, Dec. 2001. [Online]. Available: https://doi.org/10.1068/d217t

[43] K. Dovey, F. Rao, and E. Pafka, "Agglomeration and assemblage: Deterritorialising urban theory," *Urban Studies*, vol. 55, no. 2, pp. 263–273, Feb. 2018. [Online]. Available: https://doi.org/10.1177/0042098017711650

[44] M. Batty, *Urban Modelling: Algorithms, Calibrations, Predictions*. Cambridge University Press, 1976. [Online]. Available: http://www.casa.ucl.ac.uk/urbanmodelling/UrbanModelling.pdf

[45] P. M. Torrens, "Agent-based Models and the Spatial Sciences," *Geography Compass*, vol. 4, no. 5, pp. 428–448, May 2010. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1749-8198.2009.00311.x/abstract

[46] T. C. Schelling, *Micromotives and Macrobehavior*, revised ed. W. W. Norton & Company, 1978.

[47] R. Koolhaas, *Delirious New York: a retroactive manifesto for Manhattan*, new ed ed. New York: Monacelli Press, 1994, oCLC: 243831153.

[48] J. Gehl and B. Svarre, *How to Study Public Life*, 2nd ed. Washington, DC: Island Press, Oct. 2013.

[49] J. Gehl, *Life Between Buildings: Using Public Space*. New York (NY), USA: Van Nostrand Reinhold, 1987.

[50] J. Gehl, L. J. Kaefer, and S. Reigstad, "Close encounters with buildings," *Urban Design International*, vol. 11, no. 1, pp. 29–47, 2006. [Online]. Available: http://www.palgrave-journals.com/udi/journal/v11/n1/abs/9000162a.html

[51] J. Gehl, *Cities for People*. Washington (DC): Island Press, 2010.

[52] J. Grant, *Planning the good community: new urbanism in theory and practice*, ser. The RTPI library series. London ; New York: Routledge, 2006, oCLC: ocm58455832.

[53] D. Gibson, *The Wayfinding Handbook: Information Design for Public Places*, 1st ed. Princeton Architectural Press, Feb. 2009.

[54] D. Sudjic, *The language of cities*. London: Allen Lane, an imprint of Penguin Books, 2016, oCLC: 965605665.

[55] A. M. Townsend, "Making Sense of the Science of Cities," New York University, New York, Tech. Rep., Jul. 2015. [Online]. Available: http://www.citiesofdata.org/wp-content/uploads/2015/04/Making-Sense-of-the-New-Science-of-Cities-FINAL-2015.7.7.pdf

[56] J. M. Epstein, "Why Model?" *Journal of Artificial Societies and Social Simulation*, vol. 11, no. 4, p. 12, 2008. [Online]. Available: http://jasss.soc.surrey.ac.uk/11/4/12.html

[57] C. Rohilla Shalizi, *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press, 2013. [Online]. Available: http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/

[58] D. Helbing and A. Johansson, "Pedestrian, Crowd and Evacuation Dynamics," in *Encyclopedia of Complexity and Systems Science*, 2009th ed., R. A. Meyers, Ed. Berlin/Heidelberg: Springer, Jun. 2009, pp. 6476–6495. [Online]. Available: http://www.springer.com/physics/complexity/book/978-0-387-75888-6

[59] E. D. Kuligowski, R. D. Peacock, and B. L. Hoskins, "A Review of Building Evacuation Models, 2nd Edition," National Institute of Standards and Technology, Gaithersburg, MD, Technical Note NIST TN - 1680, Nov. 2010. [Online]. Available: http://www.nist.gov/manuscript-publication-search.cfm?pub_id=906951

[60] M. Schreckenberg and S. D. Sharma, Eds., *Pedestrian and Evacuation Dynamics 2002*, 2002nd ed. Springer, 2002. [Online]. Available: http://www.springer.com/mathematics/applications/book/978-3-540-42690-5

[61] N. Waldau, *Pedestrian and evacuation dynamics 2005*. Berlin: Springer, 2007, oCLC: 185026865. [Online]. Available: http://public.eblib.com/choice/publicfullrecord.aspx?p=301846

[62] W. W. F. Klingsch, C. Rogsch, A. Schadschneider, and M. Schreckenberg, Eds., *Pedestrian and Evacuation Dynamics 2008*, 2010th ed. Springer, Feb. 2010. [Online]. Available: http://www.springer.com/mathematics/applications/book/978-3-642-04503-5

[63] R. D. Peacock, E. D. Kuligowski, and J. D. Averill, Eds., *Pedestrian and Evacuation Dynamics 2011*. Boston, MA: Springer US, 2011. [Online]. Available: http://link.springer.com/10.1007/978-1-4419-9725-8

[64] U. Weidmann, U. Kirsch, and M. Schreckenberg, *Pedestrian and evacuation dynamics 2012*, 2014, oCLC: 873820918. [Online]. Available: http://public.eblib.com/choice/publicfullrecord.aspx?p=1730951

[65] H. Timmermans, Ed., *Pedestrian Behavior: Data Collection and Applications*, 1st ed. Emerald Group Publishing Limited, Nov. 2009.

[66] C. Peters and C. Ennis, "Modeling Groups of Plausible Virtual Pedestrians," *IEEE Computer Graphics and Applications*, vol. 29, no. 4, pp. 54–63, Jul. 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1591926

[67] S. Ali, K. Nishino, D. Manocha, and M. Shah, Eds., *Modeling, simulation and visual analysis of crowds: a multidisciplinary perspective*, ser. The international series in video computing. New York, NY: Springer, 2013, no. 11, oCLC: 915381146.

[68] S. Zhou, D. Chen, W. Cai, L. Luo, M. Y. H. Low, F. Tian, V. S.-H. Tay, D. W. S. Ong, and B. D. Hamilton, "Crowd modeling and simulation technologies," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 20, no. 4, pp. 20:1–20:35, Nov. 2010. [Online]. Available: http://doi.acm.org/10.1145/1842722.1842725

[69] D. C. Duives, W. Daamen, and S. P. Hoogendoorn, "State-of-the-art crowd motion simulation models," *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 193–209, Dec. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X13000351

[70] M. Bezbradica and H. J. Ruskin, "Modelling Impact of Morphological Urban Structure and Cognitive Behaviour on Pedestrian Flows," in *Computational Science and Its Applications – ICCSA 2014*, ser. Lecture Notes in Computer Science, B. Murgante, S. Misra, A. M. A. C. Rocha, C. Torre, J. G. Rocha, M. I. Falcão, D. Taniar, B. O. Apduhan, and O. Gervasi, Eds. Springer International Publishing, Jan. 2014, no. 8582, pp. 268–283. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-09147-1_20

[71] B. Hillier, "The City as a Socio-technical System: A Spatial Reformulation in the Light of the Levels Problem and the Parallel Problem," in *Digital Urban Modeling and Simulation*, ser. Communications in Computer and Information Science, S. M. Arisona, G. Aschwanden, J. Halatsch, and P. Wonka, Eds. Springer Berlin Heidelberg, 2012, no. 242, pp. 24–48. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-29758-8_3

[72] I. Benenson and P. M. Torrens, *Geosimulation: automata-based modeling of urban phenomena*. Hoboken, NJ: John Wiley & Sons, 2004.

[73] M. Batty, *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*. MIT Press, 2005. [Online]. Available: http://mitpress.mit.edu/books/cities-and-complexity

[74] ——, *The New Science of Cities*. Cambridge, Massachusetts: The MIT Press, Nov. 2013.

[75] T. Lechner, B. Watson, U. Wilensky, and M. Felsen, "Procedural city modeling," in *In 1st Midwestern Graphics Conference*, 2003.

[76] T. Lechner, P. Ren, B. Watson, C. Brozefski, and U. Wilenski, "Procedural Modeling of Urban Land Use," in *ACM SIGGRAPH 2006 Research Posters*, ser. SIGGRAPH '06. New York, NY, USA: ACM, 2006. [Online]. Available: http://doi.acm.org/10.1145/1179622.1179778

[77] D. C. Parker, S. M. Manson, M. A. Janssen, M. J. Hoffmann, and P. Deadman, "Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review," *Annals of the Association of American Geographers*, vol. 93, no. 2, pp. 314–337, Jun. 2003. [Online]. Available: http://dx.doi.org/10.1111/1467-8306.9302004

[78] J. Rocha, "Sistemas complexos, modelação e geosimulação da evolução de padrões de uso e ocupação do solo," Ph.D. dissertation, Universidade de Lisboa, Lisbon, 2012. [Online]. Available: http://hdl.handle.net/10451/6772

[79] P. M. Torrens, "Simulating Sprawl," *Annals of the Association of American Geographers*, vol. 96, no. 2, pp. 248–275, Jun. 2006. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8306.2006.00477.x/abstract

[80] D. G. Brown and D. T. Robinson, "Effects of heterogeneity in residential preferences on an agent-based model of urban sprawl," *Ecology and Society*, vol. 11, no. 1, p. 46, 2006. [Online]. Available: http://www.ecologyandsociety.org/vol11/iss1/art46/

[81] M. L. Zellner, R. L. Riolo, W. Rand, D. G. Brown, S. E. Page, and L. E. Fernandez, "The problem with zoning: nonlinear effects of interactions between location preferences and externalities on land use and utility," *Environment and Planning B: Planning and Design*, vol. 37, no. 3, pp. 408–428, 2010. [Online]. Available: http://www.envplan.com/abstract.cgi?id=b35053

[82] B. Hillier, *Space is the Machine: A Configurational Theory of Architecture*. London. UK: Space Syntax, 2007. [Online]. Available: http://www.spacesyntax.com/

[83] A. Bretagnolle, E. Daudé, and D. Pumain, "From theory to modelling: urban systems as complex systems," *Cybergeo*, Mar. 2006. [Online]. Available: http://cybergeo.revues.org/2420

[84] A. Bretagnolle and D. Pumain, "Simulating Urban Networks through Multiscalar Space-Time Dynamics: Europe and the United States, 17th-20th Centuries," *Urban Studies*, vol. 47, no. 13, pp. 2819–2839, Nov. 2010. [Online]. Available: http://usj.sagepub.com/content/47/13/2819.abstract

[85] R. L. Axtell, J. M. Epstein, J. S. Dean, G. J. Gumerman, A. C. Swedlund, J. Harburger, S. Chakravarty, R. Hammond, J. Parker, and M. Parker, "Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley," *Proceedings of the National Academy of Sciences of the United States of America*,

vol. 99, no. Suppl 3, pp. 7275–7279, 2002. [Online]. Available: http://www.pnas.org/content/99/suppl.3/7275.short

[86] K. M. Johnston, Ed., *Agent Analyst: Agent-Based Modeling in ArcGIS*. Redlands, CA: Esri Press, 2013. [Online]. Available: http://resources.arcgis.com/en/help/agent-analyst/pdf/AgentAnalyst.pdf

[87] U. Wilensky, "NetLogo," Evanston (IL), USA, 1999. [Online]. Available: http://ccl.northwestern.edu/netlogo/

[88] A. Alexiou, A. Singleton, and P. A. Longley, "A Classification of Multidimensional Open Data for Urban Morphology," *Built Environment*, vol. 42, no. 3, pp. 382–395, Oct. 2016. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article&issn=0263-7960&volume=42&issue=3&spage=382

[89] K. Dovey and E. Pafka, "What is functional mix? An assemblage approach," *Planning Theory & Practice*, pp. 1–19, Feb. 2017. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/14649357.2017.1281996

[90] H. J. Miller, "The data avalanche is here. Shouldn't we be digging?" *Journal of Regional Science*, vol. 50, no. 1, pp. 181–201, Feb. 2010. [Online]. Available: http://doi.wiley.com/10.1111/j.1467-9787.2009.00641.x

[91] A. Crooks, A. Croitoru, A. Jenkins, R. Mahabir, P. Agouris, and A. Stefanidis, "User-Generated Big Data and Urban Morphology," *Built Environment*, vol. 42, no. 3, pp. 396–414, Oct. 2016.

[92] A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, Mar. 2009. [Online]. Available: http://ieeexplore.ieee.org/document/4804817/

[93] E. P. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences," *Communications on Pure and Applied Mathematics*, vol. 13, no. 1, pp. 1–14, Feb. 1960. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/cpa.3160130102/abstract

[94] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Oct. 2009. [Online]. Available: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

[95] H. J. Miller and M. F. Goodchild, "Data-driven geography," *GeoJournal*, vol. 80, no. 4, pp. 449–461, Aug. 2015. [Online]. Available: http://link.springer.com/article/10.1007/s10708-014-9602-6

[96] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, 2nd ed. Cambridge, MA: Cambridge University Press, Dec. 2014. [Online]. Available: www.mmds.org

[97] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Los Angeles, California: SAGE Publications, 2014, oCLC: ocn871211376.

[98] G. Mcardle and R. Kitchin, "Improving the Veracity of Open and Real-Time Urban Data," *Built Environment*, vol. 42, no. 3, pp. 457–473, Oct. 2016. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article&issn=0263-7960&volume=42&issue=3&spage=457

[99] K. Schwab, *The fourth industrial revolution*, first u.s. edition ed. New York: Crown Business, 2017, oCLC: 961237143.

[100] E. Finn, *What algorithms want: imagination in the age of computing*. Cambridge, MA: MIT Press, 2017.

[101] H. Wickham and G. Grolemund, *R for data science: import, tidy, transform, visualize, and model data*, first edition ed. Sebastopol, CA: O'Reilly Media, 2016, oCLC: ocn968213225.

[102] D. E. Knuth, "Literate Programming," *The Computer Journal*, vol. 27, no. 2, pp. 97–111, Jan. 1984. [Online]. Available: https://academic.oup.com/comjnl/article/27/2/97/343244

[103] R. D. Peng, "Reproducible Research in Computational Science," *Science*, vol. 334, no. 6060, p. 1226, Dec. 2011. [Online]. Available: http://science.sciencemag.org/content/334/6060/1226

[104] G. Wilson, D. A. Aruliah, C. T. Brown, N. P. C. Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson, "Best Practices for Scientific Computing," *PLOS Biol*, vol. 12, no. 1, p. e1001745, Jan. 2014. [Online]. Available: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001745

[105] H. Wickham, "Tidy Data," *Journal of Statistical Software*, vol. 59, no. 10, pp. 1–23, 2014. [Online]. Available: http://www.jstatsoft.org/v59/i10

[106] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, 1st ed. Wiley, 1987.

[107] H. Wickham, D. Cook, H. Hofmann, and A. Buja, "Graphical inference for infovis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 973–979, Nov. 2010.

[108] A. E. Jinha, "Article 50 million: an estimate of the number of scholarly articles in existence," *Learned Publishing*, vol. 23, no. 3, pp. 258–263, Jul. 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1087/20100308

[109] H. A. Abt and E. Garfield, "Is the relationship between numbers of references and paper lengths the same for all sciences?" *Journal of the American Society for Information Science and Technology*, vol. 53, no. 13, pp. 1106–1112, Nov. 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/asi.10151/abstract

[110] L. Bornmann and H. Daniel, "What do citation counts measure? A review of studies on citing behavior," *Journal of Documentation*, vol. 64, no. 1, pp. 45–80, 2008.

[111] J. Bollen, H. Van de Sompel, A. Hagberg, L. Bettencourt, R. Chute, M. A. Rodriguez, and L. Balakireva, "Clickstream Data Yields High-Resolution Maps of Science," *PLoS ONE*, vol. 4, no. 3, p. e4803, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0004803

[112] G. Halevi, "Behind the Data: Citation characteristics in the Arts & Humanities," *Research Trends*, no. 32, pp. 23–25, Mar. 2013. [Online]. Available: http://www.researchtrends.com/issue-32-march-2013/citation-characteristics-in-the-arts-humanities-2/

[113] J. A. Banobi, T. A. Branch, and R. Hilborn, "Do rebuttals affect future science?" *Ecosphere*, vol. 2, no. 3, pp. 1–11, Mar. 2011. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1890/ES10-00142.1/abstract

[114] T. Kuhn, M. Perc, and D. Helbing, "Inheritance Patterns in Citation Networks Reveal Scientific Memes," *Physical Review X*, vol. 4, no. 4, p. 041036, Nov. 2014. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevX.4.041036

[115] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," *Science*, vol. 354, no. 6312, p. aaf5239, Nov. 2016. [Online]. Available: http://science.sciencemag.org/content/354/6312/aaf5239

[116] J. Carrión, "Literatura y 'big data'," *La Vanguardia*, pp. 20–23, Aug. 2016. [Online]. Available: http://www.lavanguardia.com/cultura/20160820/404063207130/literatura-big-data-franco-moretti-macroliteratura.html

[117] I. Mani, "How AI is revolutionising the role of the literary critic," Dec. 2016. [Online]. Available: https://aeon.co/essays/how-ai-is-revolutionising-the-role-of-the-literary-critic

[118] T. Lansdall-Welfare, S. Sudhahar, J. Thompson, J. Lewis, F. N. Team, N. Cristianini, A. Gregor, B. Low, T. Atkin-Wright, M. Dobson, and R. Callison, "Content analysis of 150 years of British periodicals," *Proceedings of the National Academy of Sciences*, vol. 114, no. 4, pp. E457–E465, Jan. 2017. [Online]. Available: http://www.pnas.org/content/114/4/E457

[119] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds, "The emotional arcs of stories are dominated by six basic shapes," *EPJ Data Science*, vol. 5, no. 1, Dec. 2016. [Online]. Available: http://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0093-1

[120] F. Karsdorp and A. v. d. Bosch, "The structure and evolution of story networks," *Open Science*, vol. 3, no. 6, p. 160071, Jun. 2016. [Online]. Available: http://rsos.royalsocietypublishing.org/content/3/6/160071

[121] M. D. Smith, A. Oglend, A. J. Kirkpatrick, F. Asche, L. S. Bennear, J. K. Craig, and J. M. Nance, "Seafood prices reveal impacts of a major ecological disturbance," *Proceedings of the National Academy of Sciences*, p. 201617948, Jan. 2017. [Online]. Available: http://www.pnas.org/content/early/2017/01/24/1617948114

[122] B. Green, T. Horel, and A. V. Papachristos, "Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence in Chicago, 2006 to 2014," *JAMA Internal Medicine*, Jan. 2017. [Online]. Available: http://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2594804

[123] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap, C. Weeg, E. E. Larson, L. H. Ungar, and M. E. P. Seligman, "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality," *Psychological Science*, p. 0956797614557867, Jan. 2015. [Online]. Available: http://pss.sagepub.com/content/early/2015/01/20/0956797614557867

[124] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205,

Mar. 2014. [Online]. Available: http://www.sciencemag.org/content/343/6176/1203

[125] D. Ruths and J. Pfeffer, "Social media for large studies of behavior," *Science*, vol. 346, no. 6213, pp. 1063–1064, Nov. 2014. [Online]. Available: http://www.sciencemag.org/content/346/6213/1063

[126] E. Pournaras, J. Nikolic, P. Velásquez, M. Trovati, N. Bessis, and D. Helbing, "Self-regulatory information sharing in participatory social sensing," *EPJ Data Science*, vol. 5, no. 1, p. 14, Apr. 2016. [Online]. Available: http://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0074-4

[127] V. Spaiser, T. Chadefaux, K. Donnay, F. Russman, and D. Helbing, "Communication power struggles on social media: A case study of the 2011–12 Russian protests," *Journal of Information Technology & Politics*, vol. 0, no. 0, pp. 1–22, Apr. 2017. [Online]. Available: http://dx.doi.org/10.1080/19331681.2017.1308288

[128] L. Shamir, "What makes a Pollock Pollock: a machine vision approach," *International Journal of Arts and Technology*, vol. 8, no. 1, pp. 1 – 10, 2015. [Online]. Available: http://www.inderscience.com/link.php?id=67389

[129] L. M. Bettencourt, "The Uses of Big Data in Cities," *Big Data*, vol. 2, no. 1, pp. 12–22, Feb. 2014. [Online]. Available: http://online.liebertpub.com/doi/abs/10.1089/big.2013.0042

[130] H. Karl and D. Eberl, "Origen etimológico-histórico del término "Catastro"," *Revista Geográfica*, no. 98, pp. 123–128, Jul. 1983. [Online]. Available: http://www.jstor.org/stable/40992450

[131] L. Bettencourt and G. West, "A unified theory of urban living," *Nature*, vol. 467, no. 7318, pp. 912–913, Oct. 2010. [Online]. Available: http://www.nature.com/nature/journal/v467/n7318/full/467912a.html

[132] L. M. A. Bettencourt, "The Origins of Scaling in Cities," *Science*, vol. 340, no. 6139, pp. 1438–1441, Jun. 2013. [Online]. Available: http://www.sciencemag.org/content/340/6139/1438

[133] D. G. Brown, R. Riolo, D. T. Robinson, M. North, and W. Rand, "Spatial process and data models: Toward integration of agent-based models and GIS," *Journal of Geographical Systems*, vol. 7, no. 1, pp. 25–47, May 2005. [Online]. Available: http://link.springer.com/article/10.1007/s10109-005-0148-5

[134] M. Behnisch and A. Ultsch, "Urban data-mining: spatiotemporal exploration of multidimensional data," *Building Research & Information*, vol. 37, no. 5-6, pp. 520–532, Nov. 2009. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/09613210903189343

[135] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, "Redrawing the Map of Great Britain from a Network of Human Interactions," *PLOS ONE*, vol. 5, no. 12, p. e14248, Dec. 2010. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0014248

[136] M. Claudel, T. Nagel, and C. Ratti, "From Origins to Destinations: The Past, Present and Future of Visualizing Flow Maps," *Built Environment*, vol. 42, no. 3, pp. 338–355, Oct. 2016. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article&issn=0263-7960&volume=42&issue=3&spage=338

[137] M. Batty, "Big Data and the City," *Built Environment*, vol. 42, no. 3, pp. 321–337, Oct. 2016. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article&issn=0263-7960&volume=42&issue=3&spage=321

[138] ——, "Editorial: Big Data, Cities and Herodotus," *Built Environment*, vol. 42, no. 3, pp. 317–320, Oct. 2016. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article&issn=0263-7960&volume=42&issue=3&spage=317

[139] R. Kitchin, T. P. Lauriault, and G. McArdle, "Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards," *Regional Studies, Regional Science*, vol. 2, no. 1, pp. 6–28, Jan. 2015. [Online]. Available: http://rsa.tandfonline.com/doi/full/10.1080/21681376.2014.983149

[140] M. Batty, "Does Big Data Lead to Smarter Cities? Problems, Pitfalls and Opportunities," *I/S: A Journal of Law and Policy for the Information Society*, vol. 11, pp. 127–525, 2015.

[141] S. Gray, O. O'Brien, and S. Hügel, "Collecting and Visualizing Real-Time Urban Data through City Dashboards," *Built Environment*, vol. 42, no. 3, pp. 498–509, Oct. 2016. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article&issn=0263-7960&volume=42&issue=3&spage=498

[142] F. Carrera, "Wise Cities: 'Old' Big Data and 'Slow' Real Time," *Built Environment*, vol. 42, no. 3, pp. 474–497, Oct. 2016. [Online].

Available: http://openurl.ingenta.com/content/xref?genre=article&issn=0263-7960&volume=42&issue=3&spage=474

[143] F. Valls, E. Redondo, D. Fonseca, R. Torres-Kompen, S. Villagrasa, and N. Martí, "Urban data and urban design: A data mining approach to architecture education," *Telematics and Informatics*, vol. 35, no. 4, pp. 1039–1052, Jul. 2018. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0736585317303416

[144] P. Salesses, K. Schechtner, and C. A. Hidalgo, "The Collaborative Image of The City: Mapping the Inequality of Urban Perception," *PLoS ONE*, vol. 8, no. 7, p. e68400, Jul. 2013. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0068400

[145] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo, "Computer vision uncovers predictors of physical urban change," *Proceedings of the National Academy of Sciences*, p. 201619003, Jul. 2017. [Online]. Available: http://www.pnas.org/content/early/2017/07/05/1619003114

[146] L. Jones, *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1949-1964*. London: Chapman & Hall, May 1986, vol. 3.

[147] N. Yau, *Data points: visualization that means something*. Indianapolis, IN: John Wiley & Sons, Inc, 2013, oCLC: ocn824725724.

[148] M. Friendly, "A Brief History of Data Visualization," in *Handbook of Data Visualization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 15–56. [Online]. Available: http://link.springer.com/10.1007/978-3-540-33037-0_2

[149] M. Friendly and D. J. Denis, "Milestones in the history of thematic cartography, statistical graphics, and data visualization," 2001. [Online]. Available: http://www.datavis.ca/milestones/

[150] D. A. Keim, "Information visualization and visual data mining," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, Jan. 2002.

[151] J. Bertin and M. Barbut, *Sémiologie Graphique. Les diagrammes, les réseaux, les cartes*. Paris: Mouton, 1967. [Online]. Available: http://books.google.es/books?id=aCItygAACAAJ

[152] J. Bertin, *La graphique et le traitement graphique de l'information.* Paris: Flammarion, 1977. [Online]. Available: http://books.google.es/books?id=kUeqQgAACAAJ

[153] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey, *Graphical Methods for Data Analysis.* Belmont, Calif.; Boston: Chapman and Hall/Cole Publishing Company, Feb. 1983.

[154] W. S. Cleveland, "Graphs in Scientific Publications," *The American Statistician*, vol. 38, no. 4, pp. 261–269, Nov. 1984. [Online]. Available: http://www.jstor.org/discover/10.2307/2683400?uid=3737952&uid=2&uid=4&sid=21101812306471

[155] W. S. Cleveland and R. McGill, "The Many Faces of a Scatterplot," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 807–822, Dec. 1984. [Online]. Available: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1984.10477098

[156] ——, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, Sep. 1984. [Online]. Available: http://www.jstor.org/discover/10.2307/2288400?uid=3737952&uid=2&uid=4&sid=21101803207581

[157] W. S. Cleveland, "Graphical Methods for Data Presentation: Full Scale Breaks, Dot Charts, and Multibased Logging," *The American Statistician*, vol. 38, no. 4, pp. 270–280, Nov. 1984. [Online]. Available: http://www.jstor.org/discover/10.2307/2683401?uid=3737952&uid=2&uid=4&sid=21101872504517

[158] W. S. Cleveland and R. McGill, "Graphical Perception and Graphical Methods for Analyzing Scientific Data," *Science*, vol. 229, no. 4716, pp. 828–833, Aug. 1985. [Online]. Available: http://www.sciencemag.org/content/229/4716/828

[159] ——, "Graphical Perception: The Visual Decoding of Quantitative Information on Graphical Displays of Data," *Journal of the Royal Statistical Society. Series A (General)*, vol. 150, no. 3, pp. 192–229, 1987. [Online]. Available: http://www.jstor.org/discover/10.2307/2981473?uid=3737952&uid=2&uid=4&sid=21101803207581

[160] W. S. Cleveland, *Visualizing Data*, 1st ed. Murray Hill, N.J. : Summit, N.J: Hobart Press, Mar. 1993.

[161] ——, *The Elements of Graphing Data*. Murray Hill, N.J: Hobart Press, Oct. 1994.

[162] E. R. Tufte, *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Feb. 1997.

[163] ——, *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, Conn: Graphics Press, Feb. 1997.

[164] ——, *Envisioning Information*. Cheshire, CT: Graphics Press, May 1990.

[165] ——, *The Visual Display of Quantitative Information*, 2nd ed. Graphics Pr, May 2001.

[166] ——, *Beautiful Evidence*, 1st ed. Cheshire, CT: Graphics Press, Jul. 2006.

[167] J. Steele and N. P. N. Iliinsky, Eds., *Beautiful Visualization: Looking at Data through the Eyes of Experts*, 1st ed. Sebastopol, CA: O'Reilly, 2010, oCLC: ocn471816105.

[168] D. Huff and I. Geis, *How to Lie With Statistics*, 1st ed. W. W. Norton, Jan. 1954.

[169] M. Friendly and SAS Institute, *Visualizing categorical data*. Cary, NC: SAS Institute, 2001.

[170] L. Wilkinson and G. Wills, *The Grammar of Graphics*. New York: Springer, 2005.

[171] H. Wickham, "A Layered Grammar of Graphics," *Journal of Computational and Graphical Statistics*, vol. 19, no. 1, pp. 3–28, 2010. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1198/jcgs.2009.07098

[172] ——, *ggplot2: elegant graphics for data analysis*, second edition ed., ser. Use R! Cham: Springer, 2016, oCLC: 958058958.

[173] D. J. Peuquet, "Representations of Geographic Space: Toward a Conceptual Synthesis," *Annals of the Association of American Geographers*, vol. 78, no. 3, pp. 375–394, Sep. 1988. [Online]. Available: http://dx.doi.org/10.1111/j.1467-8306.1988.tb00214.x

[174] J. L. Mennis, D. J. Peuquet, and L. Qian, "A conceptual framework for incorporating cognitive principles into geographical database representation," *International Journal of Geographical Information Science*, vol. 14, no. 6, pp. 501–520, Sep. 2000. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/.VA6v1_mSwbg#.VBNqJhbKRlY

[175] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, Nov. 2007. [Online]. Available: http://link.springer.com/article/10.1007/s10708-007-9111-y

[176] M. Friendly and G. Palsky, "Thematic Maps and Diagrams: Visualizing Nature and Society," in *Maps: Finding Our Place in the World*, J. R. Akerman and R. W. Karrow, Eds. Chicago, IL: University of Chicago Press, 2007, pp. 205–251.

[177] M. F. Goodchild, "Geographical data modeling," *Computers & Geosciences*, vol. 18, no. 4, pp. 401–408, May 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0098300492900694

[178] I. N. Gregory, "'A map is just a bad graph:' Why spatial statistics are important in historical GIS." in *Placing History: How maps, spatial data and GIS are changing historical scholarship*, A. K. Knowles, Ed. Redlands, CA: ESRI Press, 2008, pp. 123–149.

[179] K. Harmon, *You Are Here: Personal Geographies and Other Maps of the Imagination*, 1st ed. New York, NY: Princeton Architectural Press, Oct. 2003.

[180] M. Monmonier and H. J. d. Blij, *How to Lie with Maps*, 2nd ed. Chicago: University Of Chicago Press, May 1996.

[181] R. Kosara and J. Mackinlay, "Storytelling: The Next Step for Visualization," *Computer*, vol. 46, no. 5, pp. 44–50, May 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6412677/

[182] B. Tversky, "Visualizing Thought," *Topics in Cognitive Science*, vol. 3, no. 3, pp. 499–535, Jul. 2011. [Online]. Available: http://doi.wiley.com/10.1111/j.1756-8765.2010.01113.x

[183] J. R. Akerman, R. W. Karrow, F. M. of Natural History, and N. Library, Eds., *Maps: finding our place in the world*. Chicago: University of Chicago Press, 2007, oCLC: ocm80359919.

[184] W. Rankin, *After the map: cartography, navigation, and the transformation of territory in the twentieth century*. Chicago: University of Chicago Press, 2016.

[185] K. Harmon and G. Clemans, *The Map as Art: Contemporary Artists Explore Cartography*. New York, NY: Princeton Architectural Press, Sep. 2010.

[186] K. O'Rourke, *Walking and mapping: artists as cartographers*, ser. Leonardo. Cambridge, Massachusetts: The MIT Press, 2013. [Online]. Available: https://mitpress.mit.edu/books/walking-and-mapping

[187] Istvan, "Flowing City Maps," *Cartographic Perspectives*, no. 81, pp. 49–52, Nov. 2015.

[188] F. Jacobs, Ed., *Strange maps: an atlas of cartographic curiosities.* New York: Viking, 2009, oCLC: 837173737.

[189] M. Glaser, M. v. t. Hoff, H. Karssenberg, J. Laven, and J. V. Teeffelen, Eds., *The City at Eye Level: Lessons for Street Plinths.* Delft: Eburon Publishers, Delft, Jan. 2014.

[190] A. Semmo, M. Trapp, M. Jobst, and J. Döllner, "Cartography-Oriented Design of 3d Geospatial Information Visualization – Overview and Techniques," *The Cartographic Journal*, vol. 52, no. 2, pp. 95–106, Apr. 2015. [Online]. Available: http://dx.doi.org/10.1080/00087041.2015.1119462

[191] F. Biljecki, J. Stoter, H. Ledoux, S. Zlatanova, and A. Çöltekin, "Applications of 3d City Models: State of the Art Review," *ISPRS International Journal of Geo-Information*, vol. 4, no. 4, pp. 2842–2889, 2015. [Online]. Available: http://www.mdpi.com/2220-9964/4/4/2842

[192] F. Döner, R. Thompson, J. Stoter, C. Lemmen, H. Ploeger, P. van Oosterom, and S. Zlatanova, "4d cadastres: First analysis of legal, organizational, and technical impact—With a case study on utility networks," *Land Use Policy*, vol. 27, no. 4, pp. 1068–1081, Oct. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0264837710000207

[193] P. van Oosterom, "Research and development in 3d cadastres," *Computers, Environment and Urban Systems*, vol. 40, pp. 1–6, Jul. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0198971513000033

[194] P. Garcia-Almirall, F. Valls Dalmau, and M. Moix Bergada, "Planned vs. Real City: 3d GIS for Analyzing the Transformation of Urban Morphology," in *17th AGILE Conference on Geographic Information Science*, J. Huerta, S. Schade, and C. Granell, Eds., Castellón de la Plana (Spain), Jun. 2014. [Online]. Available: http://hdl.handle.net/10234/98968

[195] J. Snow, *On the Mode of Communication of Cholera.* John Churchill, 1855. [Online]. Available: https://books.google.es/books?id=-N0_AAAAcAAJ

[196] K. Dovey, E. Pafka, and M. Ristic, Eds., *Mapping urbanities: morphologies, flows, possibilities*. New York, NY: Routledge, 2018.

[197] J. Cheshire and O. Uberti, *London: the information capital : 100 maps and graphics that will change how you view the city*. London: Particular Books, 2014, oCLC: 922605067.

[198] R. Salerno, "Rethinking Kevin Lynch's Lesson in Mapping Today's City," in *Innovative Technologies in Urban Mapping*, A. Contin, P. Paolini, and R. Salerno, Eds. Cham: Springer International Publishing, 2014, vol. 10, pp. 25–31. [Online]. Available: http://link.springer.com/10.1007/978-3-319-03798-1_3

[199] L. Anselin and S. Williams, "Digital neighborhoods," *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, vol. 9, no. 4, pp. 305–328, Oct. 2016. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/17549175.2015.1080752

[200] C. Mattioli, "Crowd Sourced Maps: Cognitive Instruments for Urban Planning and Tools to Enhance Citizens' Participation," in *Innovative Technologies in Urban Mapping*, A. Contin, P. Paolini, and R. Salerno, Eds. Cham: Springer International Publishing, 2014, vol. 10, pp. 145–156. [Online]. Available: http://link.springer.com/10.1007/978-3-319-03798-1_13

[201] R. Kitchin and M. Dodge, *Atlas of Cyberspace*, 1st ed. Pearson Education, Jan. 2002. [Online]. Available: http://www.kitchin.org/atlas/

[202] J. Van Schaick, "Tracking research – An agenda for urban design and planning," *Research in Urbanism Series*, vol. 1, pp. 181–194, Sep. 2008. [Online]. Available: https://journals.open.tudelft.nl/index.php/rius/article/view/RiUS.1.181-194

[203] B. S. Pushkarev, *Urban Space for Pedestrians: A Quantitative Approach*, 1st ed. Cambridge, MA: The MIT Press, Jan. 1976.

[204] W. H. Whyte, *The Social Life of Small Urban Spaces*, unknown edition ed. New York: Project for Public Spaces, 1980.

[205] J. Corry, "About New York," *The New York Times*, Mar. 1974. [Online]. Available: https://www.nytimes.com/1974/03/01/archives/about-new-york-on-schmoozing-and-smooching.html

[206] J. Aggarwal and M. Ryoo, "Human Activity Analysis: A Review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011. [Online]. Available: http://doi.acm.org/10.1145/1922649.1922653

[207] M. Batty, "Predicting where we walk," *Nature*, vol. 388, no. 6637, pp. 19–20, Jul. 1997. [Online]. Available: http://www.nature.com/nature/journal/v388/n6637/full/388019a0.html

[208] J. Procházka and K. Olševičová, "Data-Driven Pedestrian Model: From OpenCV to NetLogo," in *Computational Collective Intelligence. Technologies and Applications*, ser. Lecture Notes in Computer Science, D. Hwang, J. J. Jung, and N.-T. Nguyen, Eds. Springer International Publishing, Sep. 2014, no. 8733, pp. 322–331. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-11289-3_33

[209] B. Morris and M. Trivedi, "A Survey of Vision-Based Trajectory Learning and Analysis for Surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1114–1127, Aug. 2008.

[210] N. Ihaddadene and C. Djeraba, "Real-time crowd motion analysis," in *19th International Conference on Pattern Recognition, 2008. ICPR 2008*, Dec. 2008, pp. 1–4.

[211] M. Boltes, A. Seyfried, B. Steffen, and A. Schadschneider, "Automatic Extraction of Pedestrian Trajectories from Video Recordings," in *Pedestrian and Evacuation Dynamics 2008*, W. W. F. Klingsch, C. Rogsch, A. Schadschneider, and M. Schreckenberg, Eds. Springer Berlin Heidelberg, Jan. 2010, pp. 43–54. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-04504-2_3

[212] D. Makris and T. Ellis, "Path detection in video surveillance," *Image and Vision Computing*, vol. 20, no. 12, pp. 895–903, Oct. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885602000987

[213] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera People Tracking with a Probabilistic Occupancy Map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, Feb. 2008.

[214] U. Jaenen, U. Feuerhake, T. Klinger, D. Muhle, J. Haehner, M. Sester, and C. Heipke, "QTrajectories: Improving the Quality of Object Tracking Using Self-Organizing Camera Networks," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. I-4, pp. 269–274, Jul. 2012. [Online]. Available: http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/I-4/269/2012/isprsannals-I-4-269-2012.html

[215] M. Boltes and A. Seyfried, "Collecting pedestrian trajectories," *Neurocomputing*, vol. 100, pp. 127–133, Jan. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231212003189

[216] Y. Benabbas, N. Ihaddadene, and C. Djeraba, "Motion Pattern Extraction and Event Detection for Automatic Visual Surveillance," *EURASIP Journal on Image and Video Processing*, vol. 2011, no. 1, p. 163682, Jan. 2010. [Online]. Available: http://jivp.eurasipjournals.com/content/2011/1/163682/abstract

[217] S. Seer, N. Brändle, and C. Ratti, "Kinects and human kinetics: A new approach for studying pedestrian behavior," *Transportation Research Part C: Emerging Technologies*, vol. 48, pp. 212–228, Nov. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X14002289

[218] S. Z. Nielsen, R. Gade, T. B. Moeslund, and H. Skov-Petersen, "Taking the Temperature of Pedestrian Movement in Public Spaces," *Transportation Research Procedia*, vol. 2, pp. 660–668, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352146514001070

[219] A. Willis, N. Gjersoe, C. Havard, J. Kerridge, and R. Kukla, "Human movement behaviour in urban spaces: implications for the design and modelling of effective pedestrian environments," *Environment and Planning B: Planning and Design*, vol. 31, no. 6, pp. 805–828, 2004. [Online]. Available: http://www.envplan.com/abstract.cgi?id=b3060

[220] A. Johansson, D. Helbing, and P. K. Shukla, "Specification of the social force pedestrian model by evolutionary adjustment to video tracking data," *Advances in Complex Systems*, vol. 10, no. supp02, pp. 271–288, Dec. 2007. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0219525907001355

[221] A. Crooks, A. Croitoru, X. Lu, S. Wise, J. M. Irvine, and A. Stefanidis, "Walk This Way: Improving Pedestrian Agent-Based Models through Scene Activity Analysis," *ISPRS International Journal of Geo-Information*, vol. 4, no. 3, pp. 1627–1656, Sep. 2015. [Online]. Available: http://www.mdpi.com/2220-9964/4/3/1627

[222] B. One, "Bang Goes the Theory Episode 47," Apr. 2012. [Online]. Available: http://www.bbc.co.uk/programmes/b01fnjd6

[223] "Who enjoys shopping in IKEA? (18 Jan 2011)," Jan. 2011. [Online]. Available: http://www.youtube.com/watch?v=NkePRXxH9D4&feature=youtube_gdata_player

[224] F. Cedó, "D-Lab, tecnología digital para el progreso social," *La Vanguardia*, p. 8, Feb. 2017. [Online]. Available: http://hemeroteca-paginas.lavanguardia.com/LVE05/PUB/2017/02/26/LVG20170226008SU1.pdf

[225] J. Cheshire and O. Uberti, *Where the animals go: tracking wildlife with technology in 50 maps and graphics*, 2016, oCLC: 964699381.

[226] N. Shoval, G. Auslander, K. Cohen-Shalom, M. Isaacson, R. Landau, and J. Heinik, "What can we learn about the mobility of the elderly in the GPS era?" *Journal of Transport Geography*, vol. 18, no. 5, pp. 603–612, Sep. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S096669231000044X

[227] S. Van Der Spek, "Spatial Metro – Tracking pedestrians in historic city centres," *Research in Urbanism Series*, vol. 1, pp. 77–97, Sep. 2008. [Online]. Available: https://journals.open.tudelft.nl/index.php/rius/article/view/RiUS.1.77-97

[228] S. Van der Spek, J. Van Schaick, P. De Bois, and R. De Haan, "Sensing Human Activity: GPS Tracking," *Sensors*, vol. 9, no. 4, pp. 3033–3055, Apr. 2009. [Online]. Available: http://www.mdpi.com/1424-8220/9/4/3033

[229] J. R. B. Palmer, T. J. Espenshade, F. Bartumeus, C. Y. Chung, N. E. Ozgencil, and K. Li, "New Approaches to Human Mobility: Using Mobile Phones for Demographic Research," *Demography*, vol. 50, no. 3, pp. 1105–1128, Nov. 2012. [Online]. Available: http://link.springer.com/article/10.1007/s13524-012-0175-z

[230] Ö. Balaban and B. Tunçer, "Visualizing Urban Sports Movement," in *Complexity & Simplicity - Proceedings of the 34th eCAADe Conference*, vol. 2. Oulu, Finland: University of Oulu, Aug. 2016, pp. 89–94.

[231] ——, "Visualizing and Analising Urban Leisure Runs by Using Sports Tracking Data," in *ShoCK! - Sharing Computational Knowledge! - Proceedings of the 35th eCAADe Conference*, vol. 1. Rome, Italy: Sapienza University of Rome, Sep. 2017, pp. 533–540.

[232] H. Koller, P. Widhalm, M. Dragaschnig, and A. Graser, "Fast Hidden Markov Model Map-Matching for Sparse and Noisy Trajectories," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. Gran Canaria, Spain: IEEE, Sep. 2015, pp. 2557–2561. [Online]. Available: http://ieeexplore.ieee.org/document/7313503/

[233] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008. [Online]. Available: https://www.nature.com/articles/nature06958

[234] R. Ahas and Ü. Mark, "Location based services—new challenges for planning and public administration?" *Futures*, vol. 37, no. 6, pp. 547–561, Aug. 2005. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0016328704001521

[235] C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams, "Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis," *Environment and Planning B: Planning and Design*, vol. 33, no. 5, pp. 727–748, Oct. 2006. [Online]. Available: https://doi.org/10.1068/b32047

[236] P. Pucci, F. Manfredini, and P. Tagliolato, "A New Map of the Milan Urban Region Through Mobile Phone Data," in *Innovative Technologies in Urban Mapping*, A. Contin, P. Paolini, and R. Salerno, Eds. Cham: Springer International Publishing, 2014, vol. 10, pp. 83–92. [Online]. Available: http://link.springer.com/10.1007/978-3-319-03798-1_8

[237] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular Census: Explorations in Urban Data Collection," *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 30–38, Jul. 2007.

[238] V. Soto and E. Frías-Martínez, "Automated Land Use Identification Using Cell-phone Records," in *Proceedings of the 3rd ACM International Workshop on MobiArch*, ser. HotPlanet '11. New York, NY, USA: ACM, 2011, pp. 17–22. [Online]. Available: http://doi.acm.org/10.1145/2000172.2000179

[239] K. Kloeckl, O. Senn, and C. Ratti, "Enabling the Real-Time City: LIVE Singapore!" *Journal of Urban Technology*, vol. 19, no. 2, pp. 89–112, Apr. 2012. [Online]. Available: https://doi.org/10.1080/10630732.2012.698068

[240] C. Ratti, "Phone-Call Cartography," *The New York Times*, Jul. 2011. [Online]. Available: https://www.nytimes.com/2011/07/03/sunday-review/03phone-map.html

[241] M. Lenormand and J. J. Ramasco, "Towards a Better Understanding of Cities Using Mobility Data," *Built Environment*, vol. 42, no. 3, pp. 356–364, Oct. 2016. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article&issn=0263-7960&volume=42&issue=3&spage=356

[242] C. Zhong, M. Schläpfer, S. M. Arisona, M. Batty, C. Ratti, and G. Schmitt, "Revealing centrality in the spatial structure of cities from human activity patterns," *Urban Studies*, p. 0042098015601599, Oct. 2015. [Online]. Available: http://usj.sagepub.com/content/early/2015/09/30/0042098015601599

[243] J. Reades, C. Zhong, E. Manley, R. Milton, and M. Batty, "Finding Pearls in London's Oysters," *Built Environment*, vol. 42, no. 3, pp. 365–381, Oct. 2016.

[244] A. C. Gallup, J. J. Hale, D. J. T. Sumpter, S. Garnier, A. Kacelnik, J. R. Krebs, and I. D. Couzin, "Visual attention and the acquisition of information in human crowds," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7245–7250, May 2012. [Online]. Available: http://www.pnas.org/content/109/19/7245

[245] A. Sussman, "Recording human responses with eye tracking," *Planning*, vol. 82, no. 6, pp. 31–34, 2016.

[246] R. Passini, *Wayfinding in architecture*. Van Nostrand Reinhold, Mar. 1984.

[247] M. Moussaïd, M. Kapadia, T. Thrash, R. W. Sumner, M. Gross, D. Helbing, and C. Hölscher, "Crowd behaviour during high-stress evacuations in an immersive virtual environment," *Journal of The Royal Society Interface*, vol. 13, no. 122, p. 20160414, Sep. 2016. [Online]. Available: http://rsif.royalsocietypublishing.org/content/13/122/20160414

[248] D. Quercia, "Playful Cities: Crowdsourcing Urban Happiness with Web Games," *Built Environment*, vol. 42, no. 3, pp. 430–440, Oct. 2016. [Online]. Available: http://openurl.ingenta.com/content/xref?genre=article&issn=0263-7960&volume=42&issue=3&spage=430

[249] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.

[250] B. Marcus, F. Machilek, and A. Schütz, "Personality in cyberspace: Personal web sites as media for personality expressions and impressions." *Journal of Personality and Social Psychology*, vol. 90, no. 6, pp. 1014–1031, 2006. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.90.6.1014

[251] T. Herrera, "Is Your Digital Life Ready for Your Death?" *The New York Times*, Jan. 2017. [Online]. Available: https://www.nytimes.com/2017/01/18/technology/is-your-digital-life-ready-for-your-death.html

[252] F. Roesner, B. T. Gill, and T. Kohno, "Sex, Lies, or Kittens? Investigating the Use of Snapchat's Self-Destructing Messages," in *Financial Cryptography and Data Security*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Mar. 2014, pp. 64–76. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-662-45472-5_5

[253] R. Lemov, "Anthropology's Most Documented Man, Ca. 1947: A Prefiguration of Big Data from the Big Social Science Era," *Osiris*, vol. 32, no. 1, pp. 21–42, Sep. 2017. [Online]. Available: http://www.journals.uchicago.edu/doi/abs/10.1086/694171

[254] S. Brand, *How Buildings Learn: What Happens After They're Built.* Penguin Books, Oct. 1995.

[255] E. S. Raymond, *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary.* O'Reilly Media, Oct. 1999.

[256] S. O'Grady, *The Software Paradox: The Rise and Fall of the Commercial Software Market*, 1st ed. Sebastopol (CA), USA: O'Reilly Media, 2015. [Online]. Available: http://www.oreilly.com/programming/free/software-paradox.csp

[257] C. Myhill, "Commercial Success by Looking for Desire Lines," in *Computer Human Interaction*, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, M. Masoodian, S. Jones, and B. Rogers, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, vol. 3101, pp. 293–304. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-27795-8_30

[258] K. Stock, "Mining location from social media: A systematic review," *Computers, Environment and Urban Systems*, vol. 71, pp. 209–240, Sep. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0198971518301133

[259] R. Lemov, *Database of dreams: the lost quest to catalog humanity.* New Haven London: Yale University Press, 2015.

[260] J. R. French, "A formal theory of social power," *Psychological Review*, vol. 63, no. 3, pp. 181–194, 1956.

[261] E. Bakshy, S. Messing, and L. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, p. aaa1160, May 2015. [Online]. Available: http://www.sciencemag.org/content/early/2015/05/08/science.aaa1160

[262] A. L. Schmidt, F. Zollo, M. D. Vicario, A. Bessi, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "Anatomy of news consumption on Facebook," *Proceedings of the National Academy of Sciences*, p. 201617052,

Mar. 2017. [Online]. Available: http://www.pnas.org/content/early/2017/02/28/1617052114

[263] A. Friggeri, L. Adamic, D. Eckles, and J. Cheng, "Rumor Cascades," in *Eighth International AAAI Conference on Weblogs and Social Media*, May 2014. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8122

[264] R. Meyer, "When You Fall in Love, This Is What Facebook Sees," *The Atlantic*, Feb. 2014. [Online]. Available: https://www.theatlantic.com/technology/archive/2014/02/when-you-fall-in-love-this-is-what-facebook-sees/283865/

[265] C. Ingraham, "The path of the solar eclipse is already altering real-world behavior," *Washington Post*, Aug. 2017. [Online]. Available: https://wapo.st/2vkgIBv

[266] Y. Huang, D. Guo, A. Kasakoff, and J. Grieve, "Understanding U.S. regional linguistic variation with Twitter data analysis," *Computers, Environment and Urban Systems*, vol. 59, pp. 244–255, Sep. 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0198971515300399

[267] R. A. Hill, "You've Come a Long Way, Dude: A History," *American Speech*, vol. 69, no. 3, pp. 321–327, 1994. [Online]. Available: http://www.jstor.org/stable/455525

[268] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web." Stanford InfoLab, Technical Report 1999-66, Nov. 1999, previous number = SIDL-WP-1999-0120. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[269] R. T. Fielding, "Architectural Styles and the Design of Network-based Software Architectures," Doctoral dissertation, University of California, Irvine, 2000.

[270] P. Barbera, *streamR: Access to Twitter Streaming API via R*, 2014. [Online]. Available: https://CRAN.R-project.org/package=streamR

[271] J. Gentry, *twitteR: R Based Twitter Client*, 2015. [Online]. Available: https://CRAN.R-project.org/package=twitteR

[272] M. W. Kearney, *rtweet: Collecting Twitter Data*, 2016. [Online]. Available: https://cran.r-project.org/package=rtweet

[273] H. Wickham, *httr: Tools for Working with URLs and HTTP*, 2016. [Online]. Available: https://CRAN.R-project.org/package=httr

[274] J. Gentry and D. T. Lang, *ROAuth: R Interface For OAuth*, 2015. [Online]. Available: https://CRAN.R-project.org/package=ROAuth

[275] H. Wickham, *rvest: Easily Harvest (Scrape) Web Pages*, 2016. [Online]. Available: https://CRAN.R-project.org/package=rvest

[276] A. South, *rnaturalearth: World Map Data from Natural Earth*, 2017. [Online]. Available: https://CRAN.R-project.org/package=rnaturalearth

[277] R. Bivand and C. Rundel, *rgeos: Interface to Geometry Engine - Open Source (GEOS)*, 2017. [Online]. Available: https://CRAN.R-project.org/package=rgeos

[278] E. Val, B. Navarro, J. M. Palau Riberaygua, R. Suñé, and B. Corzo, "SOS: ciudades acorraladas por el turismo," *La Vanguardia Magazine Digital*, Jul. 2017. [Online]. Available: http://www.magazinedigital.com/historias/reportajes/sos-ciudades-acorraladas-por-turismo

[279] C. Milano, "Overtourism and Tourismphobia: Global trends and local contexts," Ostelea School of Tourism and Hospitality, Tech. Rep., 2017. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.13463.88481

[280] H. Q. Vu, G. Li, R. Law, and Y. Zhang, "Tourist Activity Analysis by Leveraging Mobile Social Media Data," *Journal of Travel Research*, vol. 57, no. 7, pp. 883–898, Sep. 2018. [Online]. Available: https://doi.org/10.1177/0047287517722232

[281] D. Leung, R. Law, H. v. Hoof, and D. Buhalis, "Social Media in Tourism and Hospitality: A Literature Review," *Journal of Travel & Tourism Marketing*, vol. 30, no. 1-2, pp. 3–22, Jan. 2013. [Online]. Available: https://doi.org/10.1080/10548408.2013.750919

[282] M. C. Burns, J. R. Cladera, and M. M. Bergadà, "The spatial implications of the functional proximity deriving from air passenger flows between European metropolitan urban regions," *GeoJournal*, vol. 71, no. 1, pp. 37–52, Apr. 2008. [Online]. Available: http://link.springer.com.recursos.biblioteca.upc.edu/article/10.1007/s10708-008-9144-x

[283] B. McKercher, "The impact of distance on tourism: a tourism geography law," *Tourism Geographies*, vol. 0, no. 0, pp. 1–5, Feb. 2018. [Online]. Available: https://doi.org/10.1080/14616688.2018.1434813

[284] J. Bertin, "General Theory, from Semiology of Graphics," in *The Map Reader*, M. Dodge, R. Kitchin, and C. Perkins, Eds. Chichester, UK: John Wiley & Sons, Ltd, Apr. 2011, pp. 8–16. [Online]. Available: http://doi.wiley.com/10.1002/9780470979587.ch2

[285] R. E. Roth, A. W. Woodruff, and Z. F. Johnson, "Value-by-alpha Maps: An Alternative Technique to the Cartogram," *The Cartographic Journal*, vol. 47, no. 2, pp. 130–140, May 2010. [Online]. Available: http://dx.doi.org/10.1179/000870409X12488753453372

[286] S. Openshaw, *The modifiable areal unit problem*, ser. Concepts and Techniques in Modern Geography (CATMOG). Norwich: Geo Books, 1984, no. 38. [Online]. Available: http://qmrg.org.uk/files/2008/11/38-maup-openshaw.pdf

[287] C. A. Brewer, "A Transition in Improving Maps: The ColorBrewer Example," *Cartography and Geographic Information Science*, vol. 30, no. 2, pp. 159–162, Jan. 2003. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1559/152304003100011126

[288] R. G. Cromley and E. K. Cromley, "Choropleth map legend design for visualizing community health disparities," *International Journal of Health Geographics*, vol. 8, no. 1, p. 52, 2009. [Online]. Available: http://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-8-52

[289] M. Harrower and C. A. Brewer, "ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps," *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, Jun. 2003. [Online]. Available: http://www.maneyonline.com/doi/abs/10.1179/000870403235002042

[290] L. W. Pickle, M. Mungiole, G. K. Jones, and A. A. White, *Atlas of United States mortality*. Hyattsville, Maryland: National Center for Health Statistics, Dec. 1996, no. (PHS) 97-1015. [Online]. Available: http://www.cdc.gov/nchs/products/other/atlas/atlas.htm

[291] J. P. Snyder, *Flattening the earth: two thousand years of map projections*. Chicago London: The University of Chicago Press, 1993, oCLC: 26764604.

[292] ——, "Map Projections: A Working Manual," U.S. Geological Survey, Washington, DC, USGS Numbered Series 1395, 1987. [Online]. Available: http://pubs.er.usgs.gov/publication/pp1395

[293] J. P. Snyder and P. M. Voxland, "An album of map projections," U.S. Geological Survey, Washington, DC, USGS Numbered Series 1453, 1989. [Online]. Available: http://pubs.er.usgs.gov/publication/pp1453

[294] H. Wickham and H. Hofmann, "Product Plots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2223–2230, Dec. 2011.

[295] M. Friendly, "Extending Mosaic Displays: Marginal, Conditional, and Partial Views of Categorical Data," *Journal of Computational and Graphical Statistics*, vol. 8, no. 3, pp. 373–395, Sep. 1999. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10618600.1999.10474820

[296] M. Tennekes, *treemap: Treemap Visualization*, 2017. [Online]. Available: https://CRAN.R-project.org/package=treemap

[297] D. Wilkins, *treemapify: Draw Treemaps in 'ggplot2'*, 2017. [Online]. Available: https://CRAN.R-project.org/package=treemapify

[298] M. Bruls, K. Huizing, and J. J. v. Wijk, "Squarified Treemaps," in *Data Visualization 2000*, ser. Eurographics. Springer, Vienna, 2000, pp. 33–42. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-7091-6783-0_4

[299] L. Wilkinson, A. Anand, and R. Grossman, "Graph-theoretic scagnostics," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, Oct. 2005, pp. 157–164.

[300] L. Wilkinson, "Dot Plots," *The American Statistician*, vol. 53, no. 3, pp. 276–281, 1999. [Online]. Available: https://www.jstor.org/stable/2686111

[301] P. Gould, "Letting the Data Speak for Themselves," *Annals of the Association of American Geographers*, vol. 71, no. 2, pp. 166–176, Jun. 1981. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8306.1981.tb01346.x/abstract

[302] J. Bednar, "Big Data Visualization With Datashader," Anaconda, Inc., Tech. Rep., 2017. [Online]. Available: http://know.continuum.io/big-data-visualization-with-datashader.html

[303] R. J. Hijmans, *raster: Geographic Data Analysis and Modeling*, 2017. [Online]. Available: https://CRAN.R-project.org/package=raster

[304] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2009. [Online]. Available: http://ggplot2.org

[305] C. E. Gehlke and K. Biehl, "Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material," *Journal of the American Statistical Association*, vol. 29, no. 185, pp. 169–170, 1934. [Online]. Available: http://www.jstor.org/stable/2277827

[306] T. Hengl, "Finding the right pixel size," *Computers & Geosciences*, vol. 32, no. 9, pp. 1283–1298, Nov. 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0098300405002657

[307] K. Wright, *pals: Color Palettes, Colormaps, and Tools to Evaluate Them*, 2017. [Online]. Available: https://CRAN.R-project.org/package=pals

[308] S. Garnier, *viridis: Default Color Maps from 'matplotlib'*, 2017. [Online]. Available: https://CRAN.R-project.org/package=viridis

[309] D. A. Green, "A colour scheme for the display of astronomical intensity images," *arXiv:1108.5083 [astro-ph]*, Aug. 2011, arXiv: 1108.5083. [Online]. Available: http://arxiv.org/abs/1108.5083

[310] K. Thyng, C. Greene, R. Hetland, H. Zimmerle, and S. DiMarco, "True Colors of Oceanography: Guidelines for Effective and Accurate Colormap Selection," *Oceanography*, vol. 29, no. 3, pp. 9–13, Sep. 2016. [Online]. Available: https://tos.org/oceanography/article/true-colors-of-oceanography-guidelines-for-effective-and-accurate-colormap

[311] P. Kovesi, "Good Colour Maps: How to Design Them," *arXiv:1509.03700 [cs]*, Sep. 2015, arXiv: 1509.03700. [Online]. Available: http://arxiv.org/abs/1509.03700

[312] P. Tol, "Colour Schemes," Technical Note SRON/EPS/TN/09-002, Dec. 2012. [Online]. Available: https://personal.sron.nl/~pault/

[313] W. R. Tobler, "A Computer Movie Simulating Urban Growth in the Detroit Region," *Economic Geography*, vol. 46, pp. 234–240, 1970. [Online]. Available: http://www.jstor.org/stable/143141

[314] R. C. Geary, "The Contiguity Ratio and Statistical Mapping," *The Incorporated Statistician*, vol. 5, no. 3, p. 115, Nov. 1954. [Online]. Available: http://www.jstor.org/discover/10.2307/2986645?uid=3737952&uid=2&uid=4&sid=21101782451343

[315] L. Anselin, "A Local Indicator of Multivariate Spatial Association: Extending Geary's c," *Geographical Analysis*, Aug. 2018. [Online]. Available: http://doi.wiley.com/10.1111/gean.12164

[316] S. J. Rey and L. Anselin, "PySAL: A Python Library of Spatial Analytical Methods," in *Handbook of Applied Spatial Analysis*, M. M. Fischer and A. Getis, Eds. Springer, Berlin, Heidelberg, 2010, pp. 175–193. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-03647-7_11

[317] L. Anselin and S. J. Rey, *Modern spatial econometrics in practice: a guide to GeoDa, GeoDaSpace and PySAL*. Chicago, IL: GeoDa Press, 2014, oCLC: 923797910.

[318] S. Rey, L. Anselin, X. Li, R. Pahle, J. Laura, W. Li, and J. Koschinsky, "Open Geospatial Analytics with PySAL," *ISPRS International Journal of Geo-Information*, vol. 4, no. 2, pp. 815–836, May 2015. [Online]. Available: http://www.mdpi.com/2220-9964/4/2/815/

[319] R. Bivand, J. Hauke, and T. Kossowski, "Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods," *Geographical Analysis*, vol. 45, no. 2, pp. 150–179, 2013.

[320] R. Bivand and G. Piras, "Comparing Implementations of Estimation Methods for Spatial Econometrics," *Journal of Statistical Software*, vol. 63, no. 18, pp. 1–36, 2015.

[321] J. J. Allaire, K. Ushey, and Y. Tang, *reticulate: Interface to 'Python'*, 2018. [Online]. Available: https://CRAN.R-project.org/package=reticulate

[322] E. Neuwirth, *RColorBrewer: ColorBrewer Palettes*, 2014. [Online]. Available: https://CRAN.R-project.org/package=RColorBrewer

[323] N. Fry, "Random point distributions and strain measurement in rocks," *Tectonophysics*, vol. 60, no. 1, pp. 89–105, Nov. 1979. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0040195179901355

[324] A. L. Patterson, "A Fourier Series Method for the Determination of the Components of Interatomic Distances in Crystals," *Physical Review*, vol. 46, no. 5, pp. 372–376, Sep. 1934. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.46.372

[325] M. Morisita, "Measuring of the dispersion of individuals and analysis of the distributional patterns," *Memoirs of the Faculty of Science, Kyushu University. Ser. E. Biol*, vol. 3, pp. 65–80, 1959.

[326] A. Baddeley, E. Rubak, and R. Turner, *Spatial point patterns: methodology and applications with R*, ser. Champan & Hall/CRC Interdisciplinary Statistics Series. Boca Raton ; London ; New York: CRC Press, Taylor & Francis Group, 2016, oCLC: ocn933300812.

[327] A. Baddeley and R. Turner, "spatstat: An R Package for Analyzing Spatial Point Patterns," *Journal of Statistical Software*, vol. 12, no. 6, 2005. [Online]. Available: http://www.jstatsoft.org/v12/i06/

[328] L. Anselin, "SpaceStat, a Software Program for Analysis of Spatial Data," 1992.

[329] R. Bivand, "Implementing Spatial Data Analysis Software Tools in R," *Geographical Analysis*, vol. 38, no. 1, pp. 23–40, Jan. 2006. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.0016-7363.2005.00672.x/abstract

[330] P. A. P. Moran, "Notes on Continuous Stochastic Phenomena," *Biometrika*, vol. 37, no. 1/2, p. 17, Jun. 1950. [Online]. Available: http://www.jstor.org/discover/10.2307/2332142?uid=3737952&uid=2&uid=4&sid=21101781280923

[331] L. Anselin, "Local Indicators of Spatial Association—LISA," *Geographical Analysis*, vol. 27, no. 2, pp. 93–115, Apr. 1995. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1995.tb00338.x/abstract

[332] A. Getis and J. K. Ord, "The Analysis of Spatial Association by Use of Distance Statistics," *Geographical Analysis*, vol. 24, no. 3, pp. 189–206, Jul. 1992. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1992.tb00261.x/abstract

[333] J. K. Ord and A. Getis, "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application," *Geographical Analysis*, vol. 27, no. 4, pp. 286–306, Oct. 1995. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1995.tb00912.x/abstract

[334] B. W. Silverman, *Density estimation for statistics and data analysis*, ser. Monographs on statistics and applied probability. London ; New York: Chapman & Hall/CRC, 1986, no. 26.

[335] V. Gómez-Rubio, P. Zheng, P. Diggle, D. C. Sterratt, R. D. Peng, D. Murdoch, and B. Rowlingson, *spatialkernel: Non-Parametric Estimation of Spatial Segregation in a Multivariate Point Process*, 2017. [Online]. Available: https://CRAN.R-project.org/package=spatialkernel

[336] T. M. Davies, J. C. Marshall, and M. L. Hazelton, "Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk," *Statistics in Medicine*, p. [in press], 2017.

[337] B. S. Rowlingson and P. J. Diggle, "Splancs: Spatial point pattern analysis code in S-plus," *Computers & Geosciences*, vol. 19, no. 5, pp. 627–655, May 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/009830049390099Q

[338] A. Perrot, R. Bourqui, N. Hanusse, F. Lalanne, and D. Auber, "Large interactive visualization of density functions on big data infrastructure." IEEE, Oct. 2015, pp. 99–106. [Online]. Available: http://ieeexplore.ieee.org/document/7348077/

[339] P. Diggle, "A Kernel Method for Smoothing Point Process Data," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 34, no. 2, pp. 138–147, 1985. [Online]. Available: http://www.jstor.org/stable/2347366

[340] P. Hall, "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density," *The Annals of Statistics*, vol. 20, no. 2, pp. 675–694, Jun. 1992. [Online]. Available: https://projecteuclid.org/euclid.aos/1176348651

[341] D. Stoyan and H. Stoyan, *Fractals, random shapes, and point fields: methods of geometrical statistics*, ser. Wiley series in probability and mathematical statistics. Chichester ; New York: Wiley, 1995.

[342] C. Loader, *Local regression and likelihood*, ser. Statistics and computing. New York: Springer, 1999.

[343] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*, ser. Wiley series in probability and mathematical statistics. New York: Wiley, 1992.

[344] D. Carr, A. Olsen, J.-Y. P. Courbois, S. M. Pierson, and A. Carr, "Linked Micromap Plots: Named and Described," *Statistical Computing and Graphics Newsletter*, vol. 9, pp. 24–32, 1998. [Online]. Available: http://stat-computing.org/newsletter/issues/scgn-09-1.pdf

[345] D. B. Carr and L. W. Pickle, *Visualizing data patterns with micromaps*, ser. Chapman & Hall/CRC interdisciplinary statistics series. Boca Raton, FL: Chapman & Hall/CRC, 2010, oCLC: ocn463674429.

[346] J. Symanzik and D. B. Carr, "Interactive Linked Micromap Plots for the Display of Geographically Referenced Statistical Data," in *Handbook of Data Visualization*, C.-h. Chen, W. Härdle, and A. Unwin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 267–294. [Online]. Available: http://link.springer.com/10.1007/978-3-540-33037-0_12

[347] Q. C. Payton, M. G. McManus, M. H. Weber, A. R. Olsen, and T. M. Kincaid, "micromap: A Package for Linked Micromaps," *Journal of Statistical Software*, vol. 63, no. 2, Feb. 2015. [Online]. Available: http://www.jstatsoft.org/v63/i02/

[348] L. W. Pickle, J. B. Pearson, and D. B. Carr, "micromapST: Exploring and Communicating Geospatial Patterns in US State Data," *Journal of Statistical Software*, vol. 63, no. 3, Feb. 2015. [Online]. Available: http://www.jstatsoft.org/v63/i03/

[349] U. Ramer, "An iterative procedure for the polygonal approximation of plane curves," *Computer Graphics and Image Processing*, vol. 1, no. 3, pp. 244–256, Nov. 1972. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0146664X72800170

[350] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 10, no. 2, pp. 112–122, Dec. 1973. [Online]. Available: http://utpjournals.press/doi/10.3138/FM57-6770-U75U-7727

[351] M. Visvalingam and J. D. Whyatt, "Line generalisation by repeated elimination of points," *The Cartographic Journal*, vol. 30, no. 1, pp. 46–51, Jun. 1993. [Online]. Available: http://www.tandfonline.com/doi/full/10.1179/000870493786962263

[352] A. Teucher and K. Russell, *rmapshaper: Client for 'mapshaper' for 'Geospatial' Operations*, 2017. [Online]. Available: https://CRAN.R-project.org/package=rmapshaper

[353] Z. Wang and J.-C. Müller, "Line Generalization Based on Analysis of Shape Characteristics," *Cartography and Geographic Information Systems*, vol. 25, no. 1, pp. 3–15, Jan. 1998. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1559/152304098782441750

[354] S. Zhou and C. B. Jones, "Shape-Aware Line Generalisation With Weighted Effective Area," in *Developments in Spatial Data Handling: 11th International Symposium on Spatial Data Handling*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 369–380. [Online]. Available: https://doi.org/10.1007/3-540-26772-7_28

[355] GRASS Development Team, *Geographic Resources Analysis Support System (GRASS GIS) Software*. USA: Open Source Geospatial Foundation, 2015. [Online]. Available: http://grass.osgeo.org

[356] J. A. Dougenik, N. R. Chrisman, and D. R. Niemeyer, "An Algorithm to Construct Continuous Area Cartograms," *The Professional Geographer*, vol. 37, no. 1, pp. 75–81, Feb. 1985. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.0033-0124.1985.00075.x/abstract

[357] S. Jeworutzki, *cartogram: Create Cartograms with R*, Sep. 2016. [Online]. Available: https://CRAN.R-project.org/package=cartogram

[358] M. T. Gastner and M. E. J. Newman, "From The Cover: Diffusion-based method for producing density-equalizing maps," *Proceedings of the National Academy of Sciences*, vol. 101, no. 20, pp. 7499–7504, May 2004. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0400280101

[359] C. A. Kennedy, I. Stewart, A. Facchini, I. Cersosimo, R. Mele, B. Chen, M. Uda, A. Kansal, A. Chiu, K.-g. Kim, C. Dubeux, E. L. L. Rovere, B. Cunha, S. Pincetl, J. Keirstead, S. Barles, S. Pusaka, J. Gunawan, M. Adegbile, M. Nazariha, S. Hoque, P. J. Marcotullio, F. G. Otharán, T. Genena, N. Ibrahim, R. Farooqui, G. Cervantes, and A. D. Sahin, "Energy and material flows of megacities," *Proceedings of the National Academy of Sciences*, p. 201504315, Apr. 2015. [Online]. Available: http://www.pnas.org/content/early/2015/04/22/1504315112

[360] M. Batty, "A Theory of City Size," *Science*, vol. 340, no. 6139, pp. 1418–1419, Jun. 2013. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1239870

[361] G. Caldarelli, *Scale-free networks: complex webs in nature and technology*. Oxford: Oxford University Press, 2007.

[362] M. Newman, *Networks: An Introduction*, 1st ed. New York: Oxford University Press, May 2010.

[363] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999. [Online]. Available: http://www.sciencemag.org/content/286/5439/509

[364] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00107510500052444

[365] M. T. Gastner and M. E. Newman, "The spatial structure of networks," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 49, no. 2, pp. 247–252, 2006. [Online]. Available: http://dx.doi.org/10.1140/epjb/e2006-00046-8

[366] T. Courtat, C. Gloaguen, and S. Douady, "Mathematics and morphogenesis of cities: A geometrical approach," *Physical Review E*, vol. 83, no. 3, Mar. 2011. [Online]. Available: http://pre.aps.org/abstract/PRE/v83/i3/e036106

[367] B. Hillier, A. Penn, J. Hanson, T. Grajewski, and J. Xu, "Natural movement: or, configuration and attraction in urban pedestrian movement," *Environment and Planning B: Planning and Design*, vol. 20, no. 1, pp. 29–66, 1993. [Online]. Available: http://www.envplan.com/abstract.cgi?id=b200029

[368] B. Hillier and S. Iida, "Network and Psychological Effects in Urban Movement," in *Spatial Information Theory*, ser. Lecture Notes in Computer Science, A. G. Cohn and D. M. Mark, Eds. Springer Berlin Heidelberg, 2005, no. 3693, pp. 475–490. [Online]. Available: http://link.springer.com/chapter/10.1007/11556114_30

[369] C. Ratti, "Space syntax: some inconsistencies," *Environment and Planning B: Planning and Design*, vol. 31, no. 4, pp. 487–499, 2004. [Online]. Available: http://www.envplan.com/abstract.cgi?id=b3019

[370] H. Timmermans, X. van der Hagen, and A. Borgers, "Transportation systems, retail environments and pedestrian trip chaining behaviour: Modelling issues and applications," *Transportation Research Part B: Methodological*, vol. 26, no. 1, pp. 45–59, Feb. 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/019126159290019S

[371] M. Duckham and L. Kulik, ""Simplest" Paths: Automated Route Selection for Navigation," in *Spatial Information Theory. Foundations of Geographic Information Science*, ser. Lecture Notes in Computer Science, W. Kuhn, M. F. Worboys, and S. Timpf, Eds. Springer Berlin Heidelberg, 2003, no. 2825, pp. 169–185. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-39923-0_12

[372] M. Barthelemy, P. Bordin, H. Berestycki, and M. Gribaudi, "Self-organization versus top-down planning in the evolution of a city," *Scientific Reports*, vol. 3, p. 2153, Jul. 2013. [Online]. Available: http://www.nature.com/srep/2013/130708/srep02153/full/srep02153.html

[373] H. Makse, J. Andrade, M. Batty, S. Havlin, and H. Stanley, "Modeling urban growth patterns with correlated percolation," *Physical Review E*, vol. 58, no. 6, pp. 7054–7062, Dec. 1998. [Online]. Available: http://pre.aps.org/abstract/PRE/v58/i6/p7054_1

[374] S. Lämmer, B. Gehlsen, and D. Helbing, "Scaling laws in the spatial structure of urban road networks," *Physica A: Statistical Mechanics and its Applications*, vol. 363, no. 1, pp. 89–95, Apr. 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378437106000938

[375] P. Garcia-Almirall, "Estudi de caracterització i avaluació de terrasses en espai públic: resultats de l'estudi: anàlisi i proposta d'apreuament," Universitat Politècnica de Catalunya, Barcelona, Tech. Rep., Sep. 2016. [Online]. Available: http://hdl.handle.net/2117/97548

[376] C. Chen, M. Batty, and T. v. Vuren, "Editorial," *Transportation*, vol. 42, no. 4, pp. 537–540, Jul. 2015. [Online]. Available: https://link.springer.com/article/10.1007/s11116-015-9614-1

[377] T. Nagel, C. Pietsch, and M. Dörk, "Staged analysis: From evocative to comparative visualizations of urban mobility," in *Proceedings of the IEEE VIS 2016 Arts Program*, Baltimore, Maryland, Oct. 2016, pp. 23–30. [Online]. Available: http://visap.uic.edu/2016/materials/03_visap-0143-paper_Nagel.pdf

[378] Land Transport NZ, *Pedestrian planning and design guide*. Wellington, N.Z.: Land Transport New Zealand, 2007. [Online]. Available: http://www.nzta.govt.nz/resources/pedestrian-planning-guide/docs/pedestrian-planning-guide.pdf

[379] Great Britain Department for Transport, *Manual for streets*. London: Thomas Telford Pub., 2007. [Online]. Available: https://www.gov.uk/government/publications/manual-for-streets

[380] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, Nov. 2009. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/070710111

[381] A. Okabe and K. Sugihara, *Spatial Analysis Along Networks: Statistical and Computational Methods*, 1st ed. Wiley, Aug. 2012.

[382] A. Okabe, T. Satoh, and K. Sugihara, "A kernel density estimation method for networks, its computational method and a GIS-based tool," *International Journal of Geographical Information Science*, vol. 23, no. 1, pp. 7–32, Jan. 2009. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/13658810802475491

[383] G. McSwiggan, A. Baddeley, and G. Nair, "Kernel Density Estimation on a Linear Network," *Scandinavian Journal of Statistics*, vol. 44, no. 2, pp. 324–345, 2017. [Online]. Available: http://dx.doi.org/10.1111/sjos.12255

[384] S. Scheider, B. Gräler, E. Pebesma, and C. Stasch, "Modeling spatiotemporal information generation," *International Journal of Geographical Information Science*, pp. 1–29, Mar. 2016. [Online]. Available: http://www.tandfonline.com/doi/full/10.1080/13658816.2016.1151520

[385] E. Pebesma, "spacetime: Spatio-Temporal Data in R," *Journal of Statistical Software*, vol. 51, no. 7, 2012. [Online]. Available: http://www.jstatsoft.org/v51/i07/

[386] B. Gräler, E. Pebesma, and G. Heuvelink, "Spatio-Temporal Interpolation using gstat," *The R Journal*, vol. 8, no. 1, pp. 204–218, 2016. [Online]. Available: https://journal.r-project.org/archive/2016/RJ-2016-014/index.html

[387] G. Grolemund and H. Wickham, "Dates and Times Made Easy with lubridate," *Journal of Statistical Software*, vol. 40, no. 3, 2011. [Online]. Available: http://www.jstatsoft.org/v40/i03/

[388] R. Bivand and N. Lewin-Koh, *maptools: Tools for Reading and Handling Spatial Objects*, 2017. [Online]. Available: https://CRAN.R-project.org/package=maptools

[389] A. Zeileis and G. Grothendieck, "zoo: S3 Infrastructure for Regular and Irregular Time Series," *Journal of Statistical Software*, vol. 14, no. 6, 2005. [Online]. Available: http://www.jstatsoft.org/v14/i06/

[390] J. A. Ryan and J. M. Ulrich, *xts: eXtensible Time Series*, 2017. [Online]. Available: https://CRAN.R-project.org/package=xts

[391] R. B. Cleveland, W. S. Cleveland, and I. Terpenning, "STL: A seasonal-trend decomposition procedure based on loess," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3–73, 1990.

[392] J. J. V. Wijk and E. R. V. Selow, "Cluster and calendar based visualization of time series data," in *1999 IEEE Symposium on Information Visualization, 1999. (Info Vis '99) Proceedings*, 1999, pp. 4–9, 140.

[393] S. Pileggi and R. Amor, "Addressing Semantic Geographic Information Systems," *Future Internet*, vol. 5, no. 4, pp. 585–590, Nov. 2013. [Online]. Available: http://www.mdpi.com/1999-5903/5/4/585/

[394] E. F. Codd, "A relational model of data for large shared data banks," *Commun. ACM*, vol. 13, no. 6, pp. 377–387, Jun. 1970. [Online]. Available: http://doi.acm.org/10.1145/362384.362685

[395] A. Etz, "Introduction to the Concept of Likelihood and Its Applications," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 1, pp. 60–69, Mar. 2018. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2515245917744314

[396] A. Bechmann and P. B. Vahlstrup, "Studying Facebook and Instagram data: The Digital Footprints software," *First Monday*, vol. 20, no. 12, Dec. 2015. [Online]. Available: http://firstmonday.org/ojs/index.php/fm/article/view/5968

[397] S. Goodson, "If You're Not Paying For It, You Become The Product," Mar. 2012. [Online]. Available: http://www.forbes.com/sites/marketshare/2012/03/05/if-youre-not-paying-for-it-you-become-the-product/

[398] P. Baugh, "'Techplomacy': Denmark's ambassador to Silicon Valley," *Politico*, Jul. 2017. [Online]. Available: https://www.politico.eu/article/denmark-silicon-valley-tech-ambassador-casper-klynge/

[399] D. Helbing, "From remote-controlled to self-controlled citizens," *The European Physical Journal Special Topics*, 2017. [Online]. Available: http://dx.doi.org/10.1140/epjst/e2016-60372-1

[400] H. T. Thành, "Micro-typographic extensions to the TEX typesetting system," Ph.D. dissertation, Masaryk University Brno, Brno, Oct. 2000. [Online]. Available: http://www.pragma-ade.com/pdftex/thesis.pdf

Index