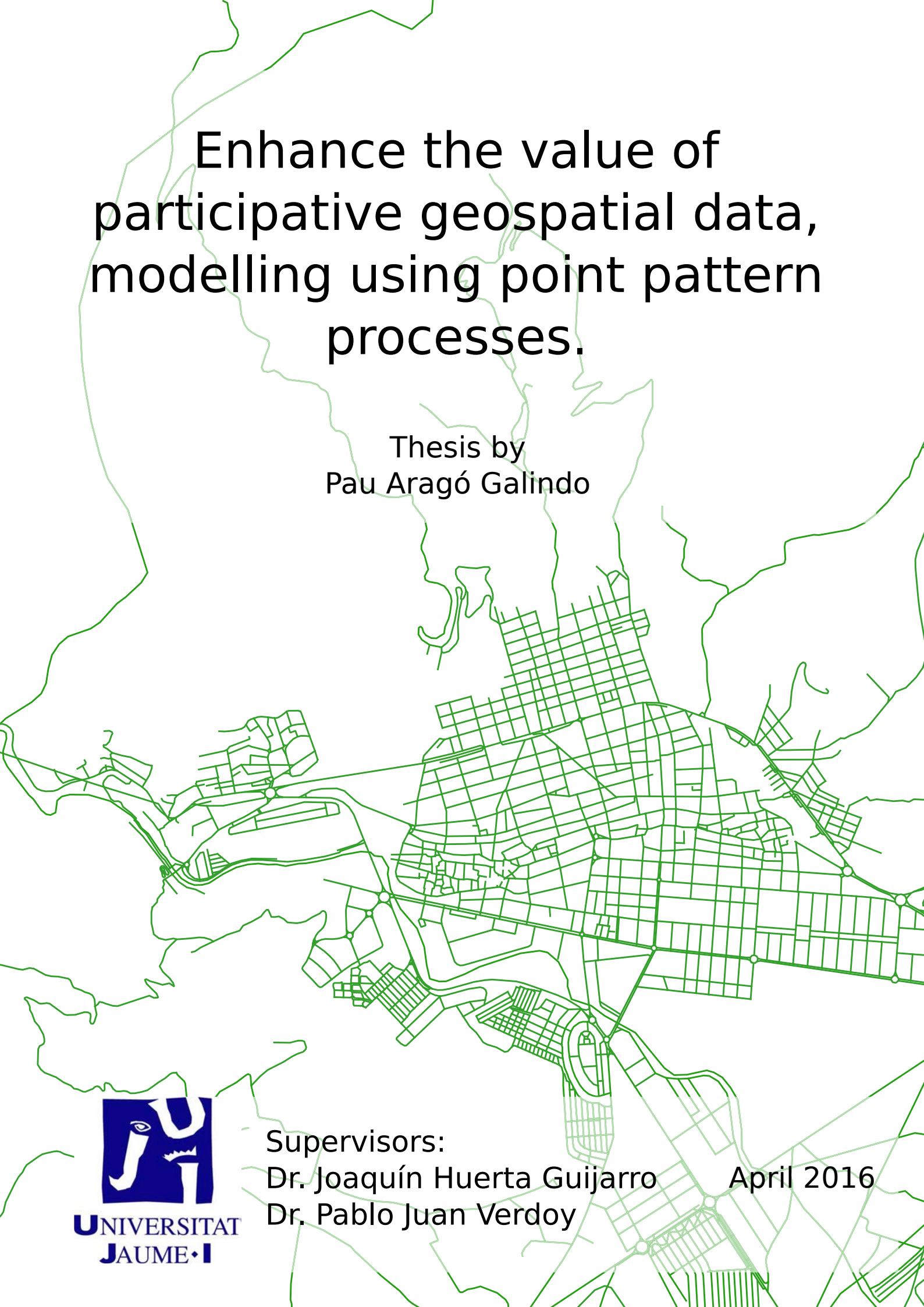
Enhance the value of participative geospatial data, modelling using point pattern processes.

Thesis by
Pau Aragó Galindo



Supervisors:
Dr. Joaquín Huerta Guijarro
Dr. Pablo Juan Verdoy

April 2016

UNIVERSITAT JAUME·I

Enhance the value of participative geospatial data, modelling using point pattern processes.

Dades espacials participatives ficades en valor, modelització mitjançant patrons de processos puntuals.

Datos geospaciales participativos puestos en valor, modelización mediante patrones de procesos puntuales

Thesis by

Pau Aragó Galindo

Dissertation submitted to the Institute of New Imaging Technologies (INIT) in partial fulfilment of the requirements for the Degree of Doctor by the University Jaume I

Supervisors:
Dr. **Joaquín Huerta Guijarro**
Dr. **Pablo Juan Verdoy**

(Castelló de la Plana, April 2016)

Vaig nàixer, pares;
vaig jugar i créixer, germaneta, tio;
vaig enamorar-me, Sandra;
vaig ser pare, Mateu , Laia i Lluc.
Per vosaltres sóc qui sóc,
per vosaltres he arribat on sóc,
per vosaltres seré,
Vos estime.

Agraïments

Voldria agraïr a Joaquín Huerta i Michel Gould per obrir-me la porta de la Investigació quan creia que ja estava tancada per sempre. A Carlos Granell i a Laura Díaz per donar-me els primers consells respecte al món de la investigació. A Rosana Peris i Antonio Grandio per rescatar-me quan tot pareixia perdut i ensenyar-me un altre món. A Jorge Mateu per recollir-me en el seu grup d'investigació. A Carlos Díaz per mostrar-me el camí des de la biologia a l'estadística. A Pablo Juan per recolzar-me en tot moment i creure en mi, en ell he trobat un amic.

També a tu que estas llegint aquestes pàgines ;)

Resum

Esta tesi tracta sobre la participació de les persones en la creació de dades espacials de forma més o menys voluntària i en l'anàlisi d'aquestes dades utilitzant patrons de processos puntuals i geoestadística. Les dades geogràfiques creades per voluntaris suposen un canvi radical en la forma com aquestes han segut creades tradicionalment. Fins fa uns anys, les dades geogràfiques han estat creades tan sols per experts i grans institucions, una aproximació des de dalt cap a baix. La revolució 2.0 a internet ha permés que la tecnologia estiga a l'abast de tothom. Per a les dades geogràfiques ha suposat que qualsevol puga contribuir a projectes voluntaris com OspenStreetMap aportant el seu coneixemnet de l'entorn. Una aportació des de baix cap a dalt. Aquest canvi deixa preguntes sobre si es poden emprar les dades, quina qualitat tenen, són creïbles els usuaris quan fan aportacions, hi ha errors... Aquesta tesi tracta de donar resposta a com fer mesures de la qualitat, la credibilitat i l'anàlisi de les dades així com facilitar la recol·lecció de dades geogràfiques provinents de les xarxes socials.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

I discovered the Geographic Information System(GIS) world with a MS-DOS Idrisi version. A user was able to do the basics GIS processes writing comands, it was an arduous task not error free whereas you were writing. Geographical data was introduced to the system using a digitalization tablet. Data introduction was an arduous task. Therefore, you had to be careful and know the limitations of digitalization tablets. Only experts well trained were able to create new and quality data. Later on comes the windows graphical interfaces such as Idrisi-32, ArcView, MapInfo. The task of creating manipulating and analysing spatial data become more easy and faster than previous MS-DOS versions but not human error free or without conceptual mistakes. The difference from paper era was not the concept but the tools and technologies. Overlay, color maps, thematic maps, etc... already existed before. The change was the computation facility, it was more easy to compute areas, to calculate distances, to make geoprocesses. Time needed for spatial analysis was decreasing. Moreover, it made possible to run geoprocesses just typing a command o clicking a button, the computer was in charge of the arduous and repetitive tasks. This procedure was a little beat dangerous because you can run a process and get results without knowing what the computer did, as a result you can have a misinterpreted layer or done mistakes. In the other hand there were quality issues with layers, precision, completeness, lineage... that become important. How good is this data? Can I overlay with other layer? In what cases?, which is the data source? Is valid for my research? Those are questions that a GIS professional or a researcher have to worry about before work with spatial data. That I can do it doesn't mean the result is right or have quality issues. Indeed, scientific has been working to ensure and measure spatial data quality (Devillers et al., 2010). Industry has worked in this field to create an ISO standard. ISO 19113 and ISO 19114 help spatial data producers to describe the quality of its products standard. Working with GIS was a qualified professional work where quality was easy to describe following predefined ISO

standards.

Web 2.0 has change the way data is produced (Oreilly, 2007),the know how is still in the hand of professional but you don't need to know programming languages to create web pages, upload content or share it. A web developer creates the tool, simplifying content creation, a company (Automattic [1] with Wordpress) or foundation (Wikimedia [2] with Wikipedia, Dropbox [3]), simplifies data storage and sharing content. Geospatial information has become as easy to create, just as it was a wikipedia content (OpenStreetMap [4]) or a cloud storage folder (CartoDB [5]), spatial data is being created by ordinary citizens in collaborative way (Goodchild, 2007). This new citizen's behaviour or movement has been called Neogeography (Turner, 2006), Cibercartography (Tulloch, 2007), user-generated content (Haklay and Weber, 2008; Goodchild, 2007) or Volunteered Geographic Information (VGI). VGI is the therm used in this thesis to refer to this kind of information in generally speaking (Goodchild, 2007).

Geographical data is not only produced by institutions big organizations, or professionals. Ordinary citizens can take advantage of web 2.0 and crowdsourcing tools to build up a database with geopositioned objects. VGI is a term used to define the personal contribution of people to collectively build a geographic information (GI) resource. VGI resource could be a street map such as OpenStreetMap[6] (OSM), a geotagged photo or photo-collection service such as Flickr[7] or Panoramio[8], a data validation service (e.g.,Geo-Wiki[9] (Fritz et al., 2009)), or information volunteered through a social network (Pultar et al., 2009) such as; Twitter[10], Facebook[11] or Instagram [12]

Geospatial world is changing as democratization has arrived, balance has changed, citizens main-streaming has taken his place and governmental data is not any more the only source (Sui et al., 2013). The rules are different. A unique type of information (up-down) can use well defined quality standards parameters and metadata. Citizens information more anarchic, so measure quality is becoming more complex. The Quality standards and metadata are not fitting within VGI. Therefore, a new quality definition may be defined to fit VGI variety of data. As an example Twitter is becoming a large source of VGI data which needs to be filtered to extract valuable information (Craglia

[1] https://automattic.com/
[2] https://wikimediafoundation.org
[3] https://www.dropbox.com
[4] http://www.openstreetmap.org
[5] https://cartodb.com/
[6] http://www.openstreetmap.org/
[7] http://www.flickr.com/
[8] http://www.panoramio.com/
[9] http://www.geo-wiki.org/
[10] https://twitter.com/
[11] https://www.facebook.com
[12] https://instagram.com/

et al., 2012).

Chapter 2 starts working with Forest fires where data is taken by professionals but not experts in spatial data information. This data is analysed using point pattern processes. A point pattern is a set locations within a study area, these locations has been recorded because there is an event (Gatrell et al., 1996), in this case a forest fire. Event could have additional data apart from its coordinates (covariates). Mathematically, point pattern, may be expressed in terms of a number of events (fires) in an arbitrary sub-regions areas, A, of the whole study region, R (Gatrell et al., 1996). Point Pattern Processes taking into account only forest fires data can explain about fires location or clustering, but covariates can tell more about its relationship with the landscape and human interaction. This point pattern analysis transform data to information. Spatial statistics is based on spatial data, that means a closer relationship between statistical models and GI. As far as data availability, amount and complexity is growing, the interdependence with geospatial technologies is needed to prepare data ready to be analysed by spatial statistic models. Data from chapter 2 has been prepared using geospatial technologies because point pattern processes models require spatial covariates not always coming from the same source.

When data is generated within an institution or company, the reward is clear, workers receive a payment for their job and the institution fulfil its goals. What happens when this relation is not clear and also there is difference level of knowledge or objectives?. Chapter 3 propose a feedback model to deal with spatial data creation among different participants working at different scale. The challenge is to keep a circular work-flow, where all the participants get information and contribute to the participatory project with their now-how , experience, expertise or simply uploading data because they can. This circular work-flow about geospatial information requires the implementation of a spatial data infrastructure (SDI) where data and information are working in circular way and not unidirectional. This concept behind is called participatory GIS (PGIS) (Sieber, 2006).

The thesis addresses some challenges to generate GI. We also introduce measures for a set of quality parameters for VGI in order to increase its utility. This thesis addresses how to measure and assure that VGI meets certain quality parameters affording reusability and provide additional value for others within various alternative scenarios such as spatial accuracy for road navigation or object annotation. This approach is useful for spatial data that is going to be used as a base data to build information over it, for instance OSM (Haklay, 2010). Projects such routing or Humanitarian OSM [13]need certain data quality to ensure its goals or a minimum quality. There are other projects within the citizen science sphere, for instance geo-

[13] https://hotosm.org/

wiki[14],where automatic classified remote sensing images are reviewed by volunteers. The quality of this data can be build base on a modified version of professional spatial data quality as you will see in chapter 4.

This thesis tries to look at the VGI, Crowsourcing GIS, Participatory GIS or citizen science. This way of spatial data creation has challenges regarding to user credibility because it is created by ordinary citizens or non GIS literates. Chapter 5 topic is credibility, where it is seen from the point of view of user descriptive credibility. In some cases descriptive credibility is not enough or only is an scratch on the surface. Therefore, an alternative should be taken, we address to this alternative in chapter 6 as a spatial credibility. Chapter 6 tries to model data to separate more credible data from data that should be reviewed. This approach is very useful to deal with Big Data where is impossible to review one by one. The methodology used is Integrated Nested Laplace Approximation (INLA) a deterministic algorithm proposed by Rue et al. (2009). INLA is a Bayesian method for modelling spatial and spatio-temporal data, it can merge point pattern events recorded on a given point and predict its values on unobserved regions, mapping continuous spatial (or spatio-temporal) variables (Blangiardo and Cameletti, 2015).Geostatistics use continuous space variables.

Social media networks such Facebook, Twitter or Instagram have an API to access some resources and develop third party applications. This is an obstacle that require some expert knowledge or programming skills, as a result there is a gap between source data and scientific community able to analyse this data. Chapter 7 is an effort to reduce this gap. An *R* package has been done to facilitate the task of get and store tweets to be analysed by the scientific community.

The motivation of this thesis is to reduce the gap between GIS experts and citizens, it is in the foundations of VGI and citizen science. The idea came as a feedback experience between experts and citizens. After this methodology is developed, came the next challenge, describe spatial data using quality parameters and credibility. This is based on quality standards to adapt it to VGI idiosyncrasy, but it is not enough. VGI contributions has a challenge regarding to the amount of data (Big Data) and contributors' background diversity. Therefore, it was necessary spatial statistics to describe and model it. A statistics approach was done using point pattern models and INLA, first approach to understand data (point pattern) and later on to apply it to VGI data (INLA). The goal is to develop a tool-set of methodologies able to get data from participatory projects, analyse, enhance it to be a valuable information extracted in a way that a user could decide which one is interesting for its purpose.

The goals of this thesis are:

- Analysis of geospatial data (created by non GIS experts) base on

[14] http://www.geo-wiki.org/

point pattern processes,

- create a methodology to measure VGI data quality,
- create a methodology to filter VGI base on data credibility,
- create a participatory methodology to collect spatial data by non GIS experts ,
- provide tools for scientific community to collect geo-tagged social network data.

Some chapter of this Thesis has been published in a Journal or are under review. This is the list of the publications:

- Chapter 2 published at European Journal of Forest Research, 2016, pp 1-14 http://dx.doi.org/10.1007/s10342-016-0945-z
- Chapter 3 published at Precision Agriculture October 2012, Volume 13, Issue 5, pp 594-610 http://dx.doi.org/10.1007/s11119-012-9267-4
- Chapter 4 partially published at 7th International Symposium on Spatial Data Quality (ISSDQ 2011). Raising awareness of Spatial Data Quality, pp 109-114. ISBN:978-989-95055-8-2
- Chapter 6 submitted to Ecological Modelling http://www.journals.elsevier.com/ecological-modelling
- Chapter 7 submitted to R-jorunal. https://journal.r-project.org/

Chapter 2

Spatial point process modelling applied to the assessment of risk factors associated with forest wildfires incidence

During the last decades, Spain as well as all the Mediterranean zone in Europe has experienced an increment in the incidence of forest fires and consequently an increase in the risk of forest fire ignition (Moreira et al., 2011; Wittenberg and Malkinson, 2009). It is worth noting here that we will mean by forest fire risk the probability that a fire ignites at a given location within a study area (Hardy, 2005), and by forest fire incidence the number of fires per unit area. Thus, fire incidence refers to the observed spatial pattern (number and location) of wildfires occurring in a given study area whilst fire risk refers to a probability measure obtained by means of mathematical model. In Spain, the increase in forest fires incidence is partly explained by climate change as well as socio-economic transformation in rural areas. Climate change has resulted in higher mean temperature and lower relative humidity, whilst socio-economic change has lead to the abandonment of farms, resulting in an increase and an unusual accumulation of forest fuels (Del Hoyo et al., 2011). The accumulation of forest fuels and the higher temperature can potentially lead to the outbreak of wildfires. Yet however, the risk of wildfires is not expected to be uniform since it is not only the quantity but also the vegetation type which is directly related to the proba-

Chapter published at European Journal of Forest Research, 2016, pp 1-14 http://dx.doi.org/10.1007/s10342-016-0945-z

bility of ignition of a wildfire. Other factors associated to the risk of wildfire ignitions are related to variables such as vegetation types, human activities, land use among others (Moreira et al., 2011; Carmo et al., 2011).

The spatial variation shown by climate and by vegetation composition and coverage and by socio-economic factors suggest strongly that wildfire risk is not evenly distributed over medium to large sized areas. It is known for example that quantity and quality of forest fuels are related to topography and geomorphology, which also affect relative humidity and rainfall (Caballero et al., 2007).

Wildfire incidence is high in the Mediterranean region of Spain due to the high temperatures and low humidity during summer months, a climatic trend with high seasonality that extends to the last months of spring and the beginning of autumn. In the province of Castellón (North-East of Spain), the process of afforestation in different agricultural areas and the increasing abandonment of rural activities have led to a situation of high vulnerability to fires, particularly in Mediterranean mountainous areas, where the aforementioned factors have led to forests being abandoned and their subsequent expansion, proliferation and fuel continuity (Bastarrika and Chuvieco, 2006). The four major natural causes of wildfire ignitions are lightning, volcanic eruption, sparks from rock falls, and spontaneous combustion (Scott, 2000; McRae, 1992). In Spain, a significant proportion of the wildfires is provoked by arsonists (Cubo María et al., 2012), making this kind of fires difficult to predict using only the weather risk index.

Predicting where a wildfire will occur is not possible due to the many factors involved in the process of wildfire incidence. However, one may construct maps showing the probability of wildfire ignition or risk maps, using data of wildfire incidence aggregated by non overlapping geographical regions with known area or by the use of the coordinates of the starting location of past wildfire events, assuming an adequate statistical model. Risk maps are needed in the process of risk management by government agencies managing natural resources, by fire fighting agencies and by the general public.

Methods to construct maps of wildfire ignition risk should be based on easily measurable variables (covariates) to be useful for forest and risk managers. An approach that has been followed to construct wildfire risk maps is based in logistic regression. For example, Del Hoyo et al. (2011) used logistic regression to model human-caused wildfire risk in central Spain. Díaz-Avalos et al. (2001) used a spatial autologistic model with covariates to construct forest fire risk maps in Oregon, using a bayesian approach. Other approaches have used spatial point process models, in which the so called intensity function is proportional to wildfire risk (Juan et al., 2012; Mateu et al., 1998; Møller and DÍaz-Avalos, 2010; Turner, 2009; Serra et al., 2014). The inclusion of covariates in the modeling helps to account for the spatial variation associated to the association between the expected mean

number of fires per unit area and the covariates, thus reducing the variance of the model parameter estimates (Martínez-Fernández et al., 2013).

Statistical models are usually preferred because they provide a measure of uncertainty for the inferences derived from the risk maps, leaving a quantitative error margin for managers and decision takers. Further, some model parameters often have a physical or a biological interpretation which can give ecologists and forest engineers answers about scientific questions of interest.

In this chapter we analyse the incidence of wildfires in the province of Castellón, in Spain in order to identify risk factors associated to wildfire incidences during the years 2001-2006 using the locations of the wildfires centroids. Our goals are to construct wildfire risk maps for the province, to identify which factors are relevant to explain the spatial variation of wildfire incidence and to analyse the interaction among wildfires. To attain our goals we need to find and fit statistical models to the observed spatial pattern of wildfire incidence. The resulting models and risk maps can provide aid in tasks such as planing wildfire fighting campaigns, to assess the hazard for the human populations and can also be helpful to plan fire prevention and pre-suppression activities. We use the discrete nature of wildfire events to build such models using point process theory and methods.

The rest of the chapter is organized as follows: In section 2 we describe the study area and the data sets. Section 3 gives all the details needed to clarify why and how we fit the models as well as some diagnostics to assess model fit. Section 4 describes and discusses the results and the interpretation and implication of such results. Also, in section 4 we discuss how the maps constructed can be used in planning fire fighting strategies, controlled burns and other related tasks. We finish with conclusions and description of future open research lines in section 5.

2.1 Study area and data.

The province of Castellón, is located in the North-Est of the Iberian peninsula (Figure 2.1). It is delimited by the mountain range called Sistema Iberico to the West, and the Mediterranean sea to the East. It also has borders with the provinces of Catalonia in the North and Valencia in the South. Castellón has a surface area of 6632 square kilometers, which represents 1.3% of the Spanish national territory. The topography and geology of Castellón allow the existence of a wide variety of vegetation types, which include coniferous forests, oak forests, mixed forests, shrubs and grasslands among others. According to the Coordination for the Information on the Environment (CORINE), there are forty four vegetation types in the province of Castellón. Some of the tree species that compose these vegetation types produce litter that is highly flammable, as is the case of the mediterranean pine (*Pinus halepensis*).

Figure 2.1: Geographic location of the province of Castellón, Spain. Source: Wikemedia Commons.

The database considered in this paper includes information about all the wildfires recorded in the study area during the period 2001-2006 . The information includes the geographic coordinates of the centroid of the fire at its final size, the year, elevation, slope, aspect, land use, distance to nearest road to the wildfire's centroid, isothermality and soil permeability. This last covariate may be taken as a proxy for fuel moisture, as less permeable soils tend to retain water and moisture. However, soil permeability may well relate with fuel moisture of the deeper horizon in the forest floor (humus) for some point after rain. Thus, moisture should not be expected to exert an influence over longer time periods, because the moisture of slow-drying fuels depends essentially of time since rain. Note also that slow-drying fuels contribute very weakly to fire ignition and fire spread. Although other authors such as Koutsias et al. (2013) have acknowledged the association between fire incidence and precipitation, we did not have those data available for our study area during the time scope of this study. Except for the geographic coordinates and the year, the rest of the information was obtained from the corresponding digital maps for the province of Castellón. Figure 2.2 shows the distribution of the wildfires occurred in Castellón per year for the time period considered in our analysis. Except for time, all the other variables were also used as spatial covariates. In particular, four continuous covariates: slope, aspect, distance to nearest road and elevation, and one categorical covariate (land use) were included in the modelling process.

Slope is the steepness or degree of incline of a surface. As slope cannot be directly computed from elevation points; one must first create either a

Figure 2.2: Spatial location of wildfires occurring in Castellón from 2001 to 2006. 2001*(top left)* and 2006 *(down right)*.

raster or TIN surface. In this paper, the slope for a particular location was computed as the maximum rate of change of elevation between that location and its surrounding pixels. Slope was expressed in degrees.

Regarding land use, we used the CORINE database (Coordination of Information on the Environment). In particular, we used the CORINE land cover map for the year 2006 (European Environment Agency, 2007 and

Table 2.1: Land use in Castellón and codes assigned. The categories represent an aggregation of a more diverse categorization for land use published by Heymann, 1994.

code	Land use
1	Coniferous forests
2	Dense forests
3	Fruit trees and berries
4	Artificial non-agricultural vegetated areas
5	Transitional woodland Scrub
6	Scrub
7	Natural grassland
8	Mixed forests
9	Urban, beaches, sand, bare rocks and water bodies

Heymann et al. 1994), on a 1:100.000 scale with a minimum mapping unit (MMU) of 25 hectares; the linear elements listed are those with a width of at least 100 meters.

In this paper we reclassified land use into nine categories (Table 2.1). We arbitrarily assigned a probability of zero to wildfire risk in category 9, so from now on we will refer exclusively to categories 1 through 8. All the covariates were defined on a grid of 360000 points covering Castellón. The digital images for some of the covariates are shown in Figure 2.3. The main causes of wildfires in Castellón during 2001-2006 were lightning strikes, followed by negligences and arsonism. It is expected that fires due to those causes be associated to different risk factors. For this reason, we made separate statistical models for wildfires caused by lighning strike (natural cause) and for negligence, arsonism and other human related causes.

2.2 Statistical analysis

The statistical analysis of the wildfire patterns for the years 2001-2006 in Castellón was done using the theory and methodology of spatial point processes (Cressie, 1993; Møller and Waagepetersen, 2007). Although we could have followed the most common logistic and auto logistic approaches, those methods cannot be extended easily to different spatial scales. On the other hand, as spatial point process modelling is based on the first and second order properties. The spatial continuity o such properties allows models fitted at a given scale to be adapted to a coarser or dinner resolution scale without much difficulty. A full description of such theory is beyond the scope of

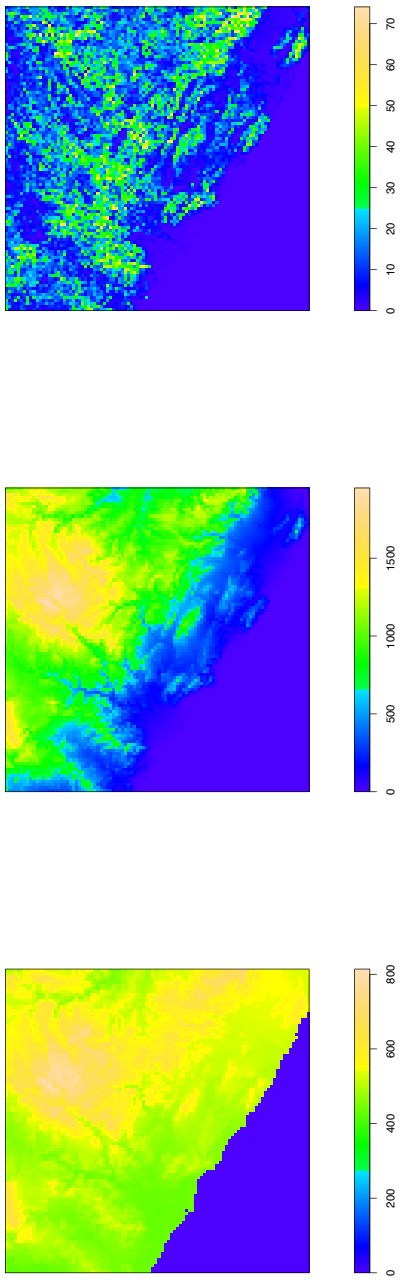


Figure 2.3: Spatial variation of the covariates available for this study. (a) Continuous: Slope (up), Elevation (middle), Permeability (down)).

this paper, and the interested reader is referred to the previously mentioned authors. Here we will only describe the theoretical parts that are relevant to explain our analytical methods.

2.2.1 First and second order properties.

We will denote the points of a spatial point process by $x = (\text{long}, \text{lat})$ and we will make a distinction of the theoretical spatial point process and the spatial point pattern observed whenever the context of the text does not make clear to which one of the two we are referring. Point process models can be characterised by their first and second order properties, which are analogous to first and second moments of ordinary random variables. We will denote by $\mathcal{A}$ the study area, also known as the observation window.

The (first-order) intensity function of a spatial point process is defined as

$$\lambda(x) = \lim_{|dx|\to 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\}, \tag{2.1}$$

where dx denotes a small region containing the point x. and $E[\cdot]$ denotes the expected value of the random variable inside the brackets.

$\lambda(x)$ may be interpreted as the expected number of events occurring in an infinitesimal set of area dx and is known as the intensity function of the point process. The intensety function is unknown and has to be estimated from the data $\{x_1, \ldots, x_n\}$ from the observed spatial point pattern. Estimation may be done by posting a parametric model $\lambda_\theta(x)$ for the intensity function or non parametrically, using kernel density estimators (Cressie, 1993; Diggle et al., 1983).

$$\widehat{\lambda}(x) = \frac{1}{n} \sum_{i=1}^{n} K_n \frac{1}{h^2} \left(\frac{x - x_i'}{h} \right) \tag{2.2}$$

where h is know as the bandwidth and controls the amount of smoothing in the estimation of λ. Non parametric estimates of $\lambda(s)$ are commonly used as exploratory tools to gain insight about the type of parametric model that should be fitted to the data and have been used in fire incidence mapping by Koutsias et al. (2004); De la Riva et al. (2004); Amatulli et al. (2007).

We fitted nonparametric kernel estimators to the observed wildfire patterns for each year considered in our study. The resulting estimates are shown in Figure 2.4. The maps for $\hat{\lambda}$ for each year show that wildfires in Castellón tend to occur in some fixed areas forming clusters of different sizes. Part of such clustering may be related to factors such as elevation, land use, or others. Nevertheless, the plots suggest that the class of point process models that should be fitted to the observed patterns need to consider clustering and the possible effect of covariates. Possible choices are the Non

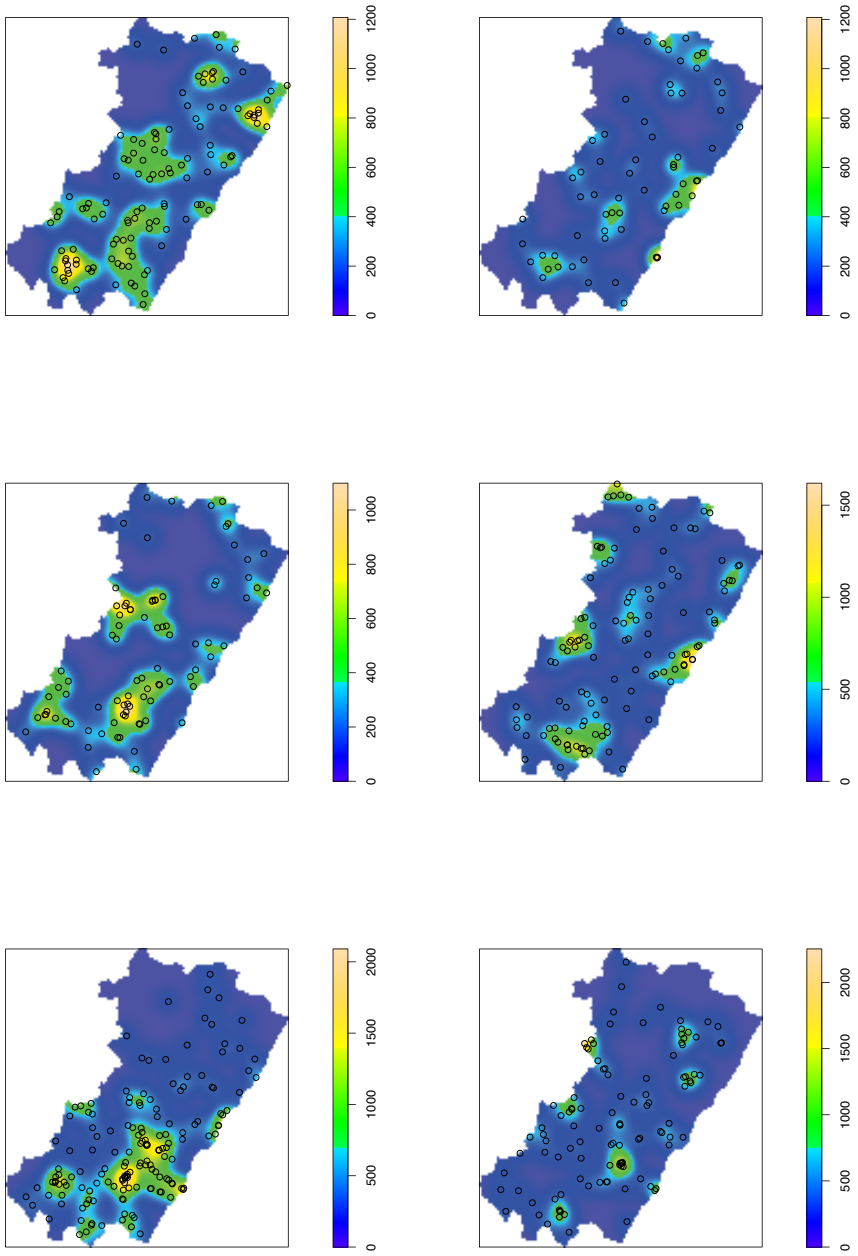


Figure 2.4: Nonparametric intensity function estimates for wildfires occurring in Castellón region in the different years considered in this study using kernel density estimator.

Homogeneous Poisson point process or an interaction model that can allow clustering, such as the Area Interaction.

The second-order intensity is a measure of the dependency structure of the events in $\mathcal{A}$ and is given by

$$\lambda_2(x_i, xj) = \lim_{|dx_i|,|dx_j|\to 0} \left\{ \frac{E[N(dx_i)N(dx_j)]}{|dx_i||dx_j|} \right\} \tag{2.3}$$

The second order intensity does not have a biological interpretation as it relates to data. An alternative very useful statistic is the reduced second moment function (Ripley 1976 and 1981)

$$K(h) = \lambda^{-1} E[N_o(h)] \tag{2.4}$$

where $N_o(h)$ is the number of extra events at a distance h from an arbitrary event $x \in \mathcal{A}$. When the points of the process distribute independently at random within $\mathcal{A}$ we speak of Complete Spatial Randomness (CSR), which can be modelled with an homogeneous Poisson process (Cressie, 1993; Diggle et al., 1983). In this case, the intensity function (2.1) is a constant λ, and a non parametric estimator of $K(h)$is given by

$$\hat{K}(h) = \frac{|\mathcal{A}|}{n(n-1)} \sum_{x \neq y} I(|x - y| \leq h) w_{\mathcal{A}}(x, y) \tag{2.5}$$

where $|\mathcal{A}|$ is the area or volume of the study area, x and y are arbitrary data points and $w_{\mathcal{A}}(x, y)$ is a border effect correction.

The reduced second moment function is also known as Ripley's K-function or simply as the K-function.

The intensity function provides information related to the homogeneity of the process within $\mathcal{A}$, whilst second-order properties provide information related to the interaction between points in a spatial point pattern, and can be used to test the null hypothesis of CSR Illian et al. (2008). For a point pattern with n points, non parametric tests for CSR are constructed by simulating s spatial patterns of size n from an homogeneous spatial Poisson process, under CSR and using such simulated patterns, compute confidence bands for the K-function

$$U(h) = max_{i=2,\ldots,s} \left\{ \hat{K}_i(h) \right\}$$

$$L(h) = min_{i=2,\ldots,s} \left\{ \hat{K}_i(h) \right\}$$

The hypothesis of CSR is rejected if the empirical K-function obtained from the observed point pattern fall outside the confidence bands for some set of distances h. Non parametric tests based on second order moments allow to conclude whether or not an observed spatial point pattern follows a random, clustered or dispersed distribution pattern. Non homogeneity

of $\lambda(s)$ may often be explained through spatial varying covariates such as temperature, litter depth and elevation above sea level among others. When information on spatial covariates is available, a common choice for the form of the intensity function is

$$log[\lambda(x)] = \mathbf{z}^T\beta \tag{2.6}$$

where $\mathbf{z}$ is a vector whose entries are the values of the covariates at the point x, and β is a vector of coefficients. For model identifiability we used sum contrasts for the categorical variables, i.e., the coefficients associated to the satisfy the restriction $\sum_j \beta_j = 0$.

Tests of CSR which are constructed from functional summary statistics of an observed pattern such as the $K - function$ are useful for two reasons: when CSR is conclusively rejected, the behavior of such summary statistics provide clues about the kind of model which might provide a reasonable fit to the data. They also suggest preliminary estimates of model parameters. In our case, we tested first the null hypothesis of Complete Spatial randomness (CSR) using the K-function in order to gain further insight about weather clustering or repulsion was evident in the observed wildfire patterns.

For those years for which the CSR hypothesis was rejected and the evidence was for clustering, we fitted first non homogeneous Poisson point process (NHPP) models, with intensity function of the form 2.6. NHPP are a natural alternative to CSR, since they preserve the assumption of independence between events of Poisson point processes. Although the NHPP models may not be good to capture strong clustering in the observed patterns, they are useful as a first step to check if the spatial variability in the observed spatial pattern may be explained by the spatial variation of the covariates. Also, in case the NHPP gives a good fit to the data, the significance of the covariate effect may be done using the fact that the coefficient estimators are asymptotically normal if the number of data is large enough (Waagepetersen, 2007), as is our case. Because the intensity function is not constant on $\mathcal{A}$, we used the corrected version for the $K - function$ (Møller and Waagepetersen, 2007) to test goodness of fit, namely

$$\hat{K}_{inhom}(r) = |A|^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{I(0 < \|x_i - x_j\| \leq r)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j)} w_{ij}^{-1}. \tag{2.7}$$

For an inhomogeneous random Poisson process with intensity function $\lambda(u)$, the inhomogeneous K-function is $K_{inhom}(r) = \pi r^2$, exactly as for the homogeneous case. We can then define the L_{inhom}-function as

$$L_{inhom}(r) = \sqrt{K_{inhom}(r)/\pi}$$

Note that the inhomogeneous K-function depends on the first-order properties of the point process.

Wildfires burn areas that occupy a wide surface. Although the shape of the burned area is highly irregular, it is expected that the incidence of a wildfire in a given zone affects the probability of another wildfire starting within the burned area. This results in negative interaction between wildfires at least for some short term. Therefore, we also explored the goodness of fit of area interaction models (Baddeley and Van Lieshout, 1995). For a bounded number of events $s_1, \ldots, s_n$, the area interaction model has a density of the form

$$f(s_1, \ldots, s_n) = \alpha \beta^{n(s)} \gamma^{-A(s)} \tag{2.8}$$

where α is a proportionality constant, β is the intensity function of the process, γ is the interaction parameter, $n(s)$ is the number of points of the process and $A(s)$ is the area of the region formed by the intersection of balls of radius r centered at the points of the process. For the intensity β we choose a log linear form as in (2.6). The interaction parameter γ can be any positive number. If $\gamma = 1$ then the model reduces to a Poisson process with intensity κ. If $\gamma < 1$ then the process is regular, while if $\gamma > 1$ the process is clustered. Thus, an Area-Interaction process can be used to model either clustered or regular point patterns. Two points interact if the distance between them is less than $2r$. If $\gamma = 0$, then a hardcore process is encountered in which there are no points within the hardcore distance (r) of any point in the point process; that is, there is total inhibition.

Goodness of fit for the NHPP and for the area-interaction models was done by simulating patterns under the fitted models with the same number of wildfires for each year. Each of the simulated patterns was the used to compute the empirical non homogeneous K-function to obtain the envelopes. We used the rank test (Grabarnik et al., 2011) to compute the 95% envelopes to get a better approximation to the true p-value for the test. To fit the models we used the *R* free software Team (2014), and the contributed package *spatstat* (Baddeley and Turner, 2000; Baddeley et al., 2004), which can be obtained from the *R* project website www.r-project.org.

In all cases, the intensity function was standardized to integrate to 1.0 in the study area, thus becoming a probability density function. Therefore, the intensity function $\lambda(x)$ is proportional to the probability of ignition of a wildfire in x, and thus to the risk of wildfire at x.

2.3 Results

The number of wildfires occurring on each year considered in this study is shown in Table 2.2. The number of fires per year showed high variability during the time span of this study, with an average of 110.5 fires per year. Years 2002, 2003 and 2005 were those that deviated more from the average,

Table 2.2: Annual distribution of wildfires occurring in Castellón by year.

Year	Number of wildfires
2001	120
2002	65
2003	88
2004	120
2005	160
2006	110
All years	663

due to high precipitation in both years during the summer months, resulting in an increase in fuel moisture.

The spatial distribution of the wildfires (Figure 2.2) and the kernel estimators of the intensity function (Figure 2.4) indicated the presence of multiple hot spots, where wildfires tend to occur at a significant higher rate than in the rest of the province. The majority of the wildfires observed in Castellón between 2001 and 2006 occurred at distances closer than 400 meters from the nearest road. This last covariate is a proxy of human caused wildfires, as most of human activities tend to occur near roads. The empirical odds of wildfire for the different vegetation classes is shown in Figure 2.7, where one can see that the risk of wildfires standardised by the area covered by each land use category was higher for classes one, six and seven, which correspond to coniferous forests, scrub and natural grasslands. Areas covered by that kind of vegetation have been recognised as fire prone (Gonzalez et al., 2006).

The multiplicative structure of the area-interaction model permits to analyze each component separately. The intensity parameter was given a log linear structure dependent on the covariates, as in (2.6), where Z is a matrix whose entries are the values covariates. The parameter estimates of the significant covariates for each year are shown in table 2.3. The absolute value of the coefficient estimates related to elevation, slope and distance to the nearest road are small due to the variation range of those continuous covariates. When multiplied by the values of the corresponding covariate and exponentiating the result gives either the increase or the decrease in the intensity function associated to the covariate. Comparing the values of $\exp \beta_j(Z_{jk} - Z_{jl})$ provides a measure of the relative change in the probability of wildfire caused by a change $Z_{jk} - Z_{jl}$ in the covariate. Thus, for example, in 2001, a change of $100m$ in elevation resulted in a decrease of 10% in the probability of wildfire, keeping the rest of the covariates at fixed values. Analogous results are obtained for slope and distance to nearest road.

Table 2.3 shows the coefficient estimates for the logarithm of the intensity

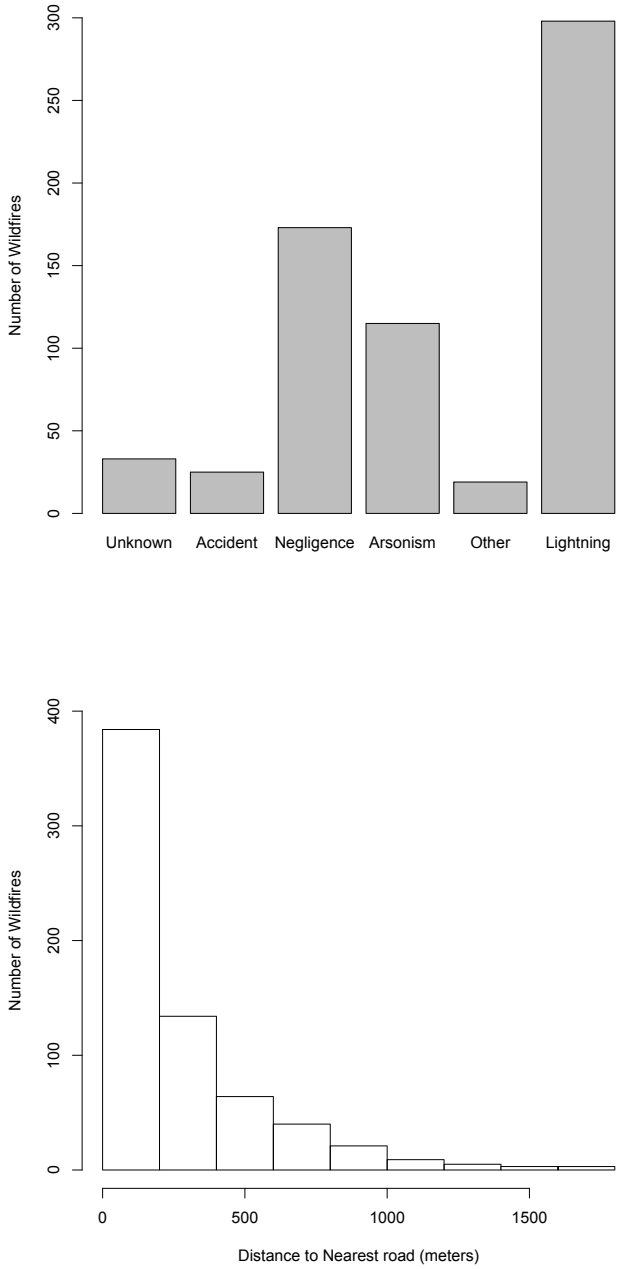


Figure 2.5: Histograms of wildfire incidences by covariate (cause and distance) in Castellón, for the years 2001-2006.

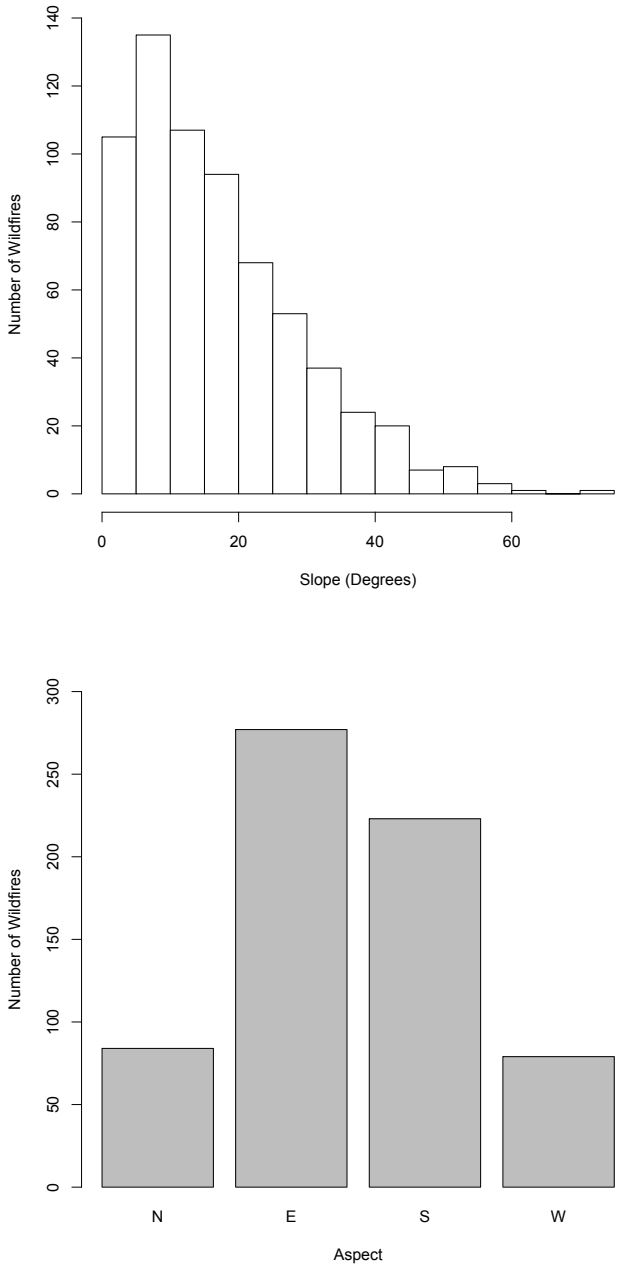


Figure 2.6: Histograms of wildfire incidences by covariate (slope and aspect) in Castellón, for the years 2001-2006.

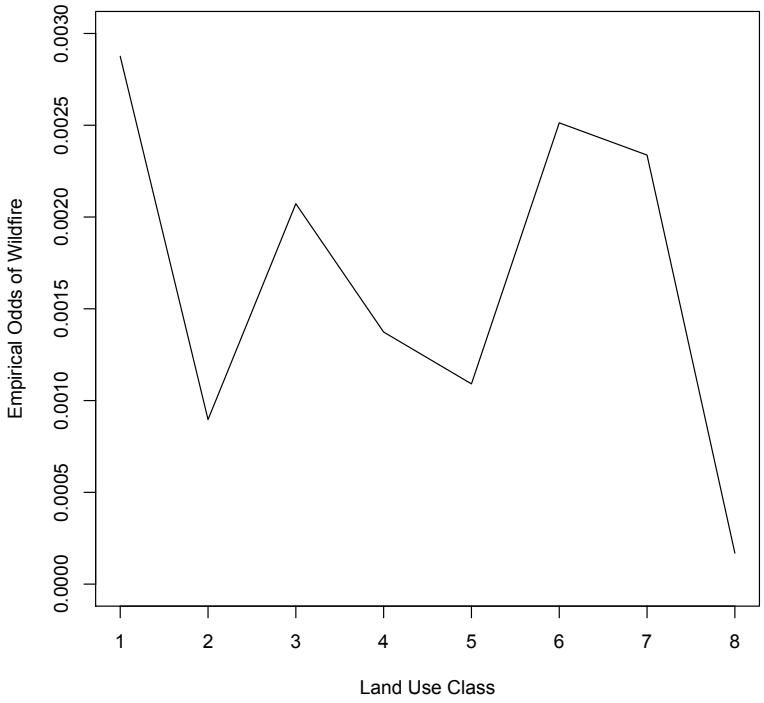


Figure 2.7: The odds of wildfire incidence for different land use classes.

Table 2.3: Parameter estimates for the inhomogeneous Area-Interaction models fitted to the wildfire incidences in Castellón, Spain, for the years 2001 to 2006 for natural causes. Only significant coefficients are shown.

Coefficient	Year					
Estimate	2001	2002	2003	2004	2005	2006
β_0	4.481	5.329	6.726	-1.639	6.726	-9.749
Elevation	4.49E-4	-8.65E-4	0.002	6.99E-4	0.001	-0.001
Slope	0.022	0.002	0.030	0.023	0.028	0.048
Distance to Nearest Road	-5.90E-4	8.87E-5	-0.002	5.97E-4	3.28E-4	-0.001
Isothermality	0.045	-0.180	-0.048	0.149	0.095	0.370
Permeability	-0.003	0.011	-0.005	1.04E-5	-0.013	0.001
Interaction ($log[\gamma]$)	0.621	-12.645	4.108	1.058	-27.613	0.732

function (equation 2.6) for naturally-caused wildfires. The sign and value for the different coefficients for the years considered is not constant and indicates that the importance and association of the covariates to wildfire incidence

is not constant over time. The coefficients for the log intensity function for human caused wildfires are shown in Table 2.4. Unlike the coefficients for the naturally-caused wildfires, the sign of the coefficients for most of the covariates remains constant for the years considered in this study.

Distance to nearest road has a negative or neutral effect on the probability of wildfire. This covariate has been considered a proxy of human activity, which is another important cause of wildfires in the study area as well as in other parts of the Mediterranean basin (Wittenberg and Malkinson, 2009; Del Hoyo et al., 2011). The negative coefficients for this covariate indicate that human caused wildfires are more likely to occur near the roads, where arsonists and other human activities are often observed.

Table 2.4: Parameter estimates for the inhomogeneous Area-Interaction models fitted to the wildfire incidences in Castellón, Spain, for the years 2001 to 2006 for human related causes. Only significant coefficients are shown.

Coefficient	Year					
Estimate	2001	2002	2003	2004	2005	2006
β_0	-1.035	4.375	-9.717	-5.453	7.426	5.739
Elevation	-0.002	-7.85E-4	-5.47E-4	-0.001	-5.59E-4	-0.001
Slope	0.027	0.0024	0.07	0.013	0.004	-0.017
Distance to Nearest Road	-0.001	-0.002	6.02E-4	-0.001	-7.27E-4	4.54E-5
Isothermality	0.202	0.163	0.457	0.333	0.185	0.090
Permeability	-1.80E-4	-0.011	-0.004	-1.003	-0.017	-0.007
Interaction ($log[\gamma]$)	-1.256	-70.960	3.025	1.479	2.492	4.193

The parameter β_0 can be thought as the basal risk for each year if the other β coefficients are kept equal to zero and the interaction parameter γ is equal to unity. Under this consideration, we see that the variation of β_0 along the years followed the same trend as the yearly number of wildfires shown in table2.2. Thus, we may think of β_0 as a parameter related to the average annual risk in the whole province and in consequence to the number of wildfires per year $\{n_t, t = 2001, \ldots, 2006\}$ and that the remaining β coefficients increase or decrease the wildfire risk according to the spatial variation shown by the linear combination of the covariates. The table also shows that for areas covered with coniferous forests the risk of wildfire was high, in particular during 2001 and 2002. Except for areas with land use 2, 4 and 7, the fire risk associated to land use followed a similar trend along the years considered in our study, with higher wildfire risk during 2001 and 2002 and moderate risk the rest of the years.

The estimates of the interaction parameter γ for naturally-caused wildfires took values below 1.0 for the years 2002 and 2005 (Table 2.4), and for the years 2001 and 2002 for human caused wildfires (2.4) indicating the presence

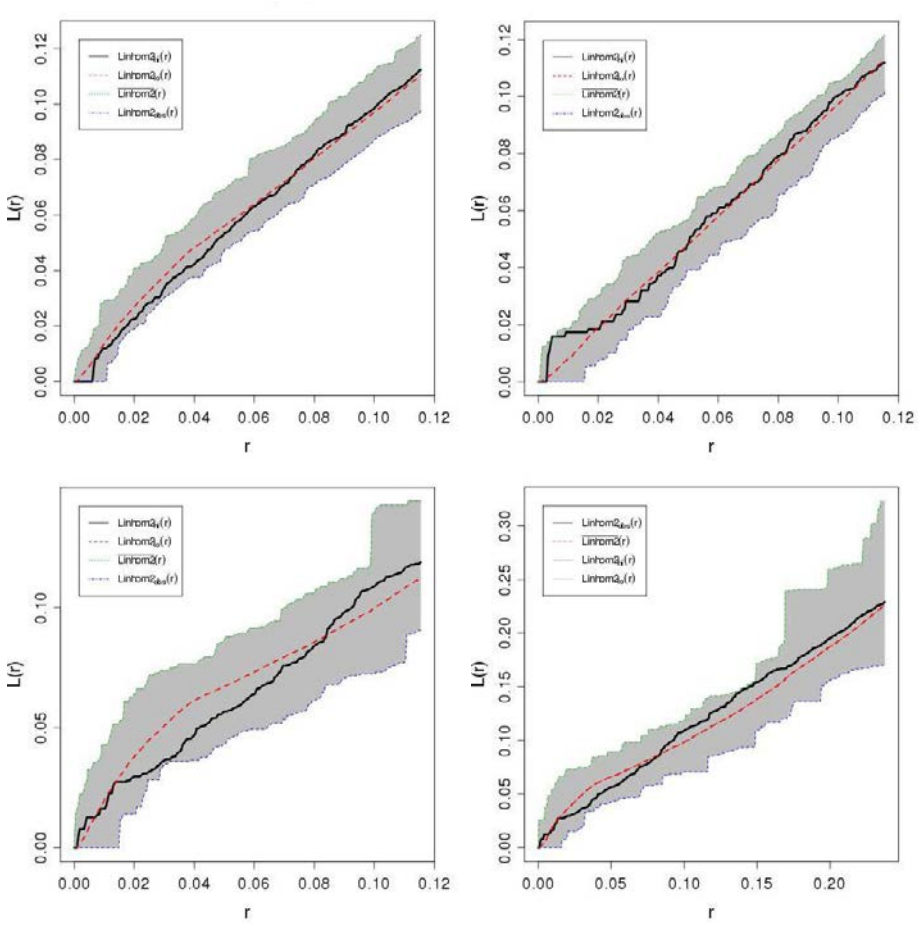


Figure 2.8: The inhomogeneous L-function for Area Interaction models fitted to the wildfire incidence data in Castellón including all the covariates, for the years 2001-2004.

of repulsion among the wildfires for those years and causes. This results in a more regular pattern of wildfire spatial distribution than under complete spatial randomness. For the rest of the years in this study, the positive value of γ indicates the presence of clustering of wildfire incidences. The fact that the parameter γ in not closer to 1.0 provides evidence that the incidences of wildfires in Castellón are not independent events, but the result of a complex process where the incidence of a wildfire in a given location somehow affects the probability of ignition of other wildfires.

Figure 2.9 shows the intensity function estimates for the naturally-caused wildfires. The areas of high intensity show different location for the different years. This is a result of the variability of the covariate effects described in

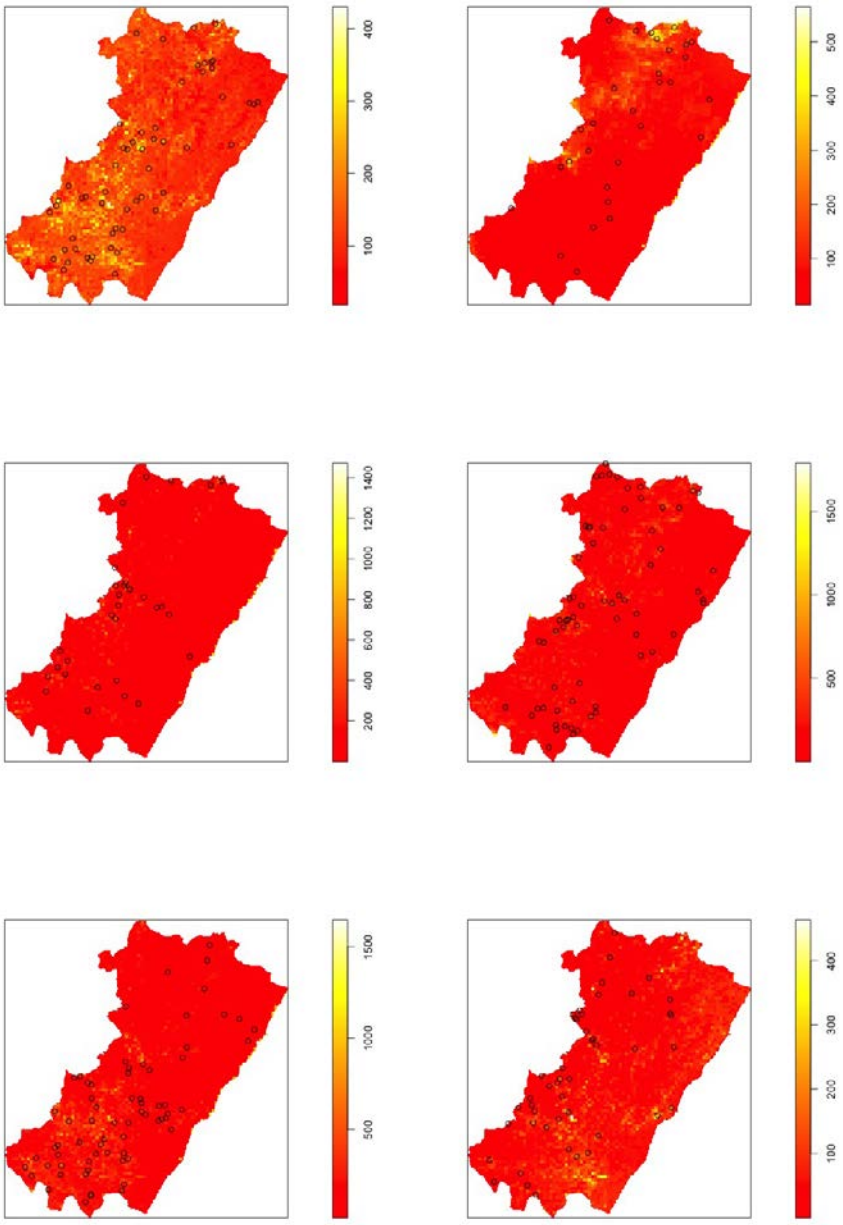


Figure 2.9: Esimated intensity function obtained from the area-Interation models for the years 2001 to 2006 in Castellón, natural causes.

previous paragraphs where we described the results of the models fitted to the intensity function for the observed wildfire observed patterns. The map for the intensity function of human caused wildfires (Figure 2.10) shows that the higher risk values occurred in a centered band going from south west to North-East, where the elevation ranges between 0 and $500m$, where the highest risk was located in the south east part of the province, not in high elevation areas where the Iberian Range is located in Castellón. Note that the maps do not resemble the spatial pattern of the covariates (Figure 2.3), as what we show in the maps in Figure 2.10 is the combination of all the components in model 2.8.

The risk maps also provide insight about the spatial variation of wildfire risk, and for non overlapping areas A and B within the study area, the ratio (q).

$$q = \frac{\int_A \lambda(u)du}{\int_B \lambda(u)du}$$

may be evaluated to get a measure of the relative risk of wildfire in those areas or to compare the estimated annual changes in wildfire risk in the same area if we take $A = A_t$ and $B = A_{t+k}$ with $k = 1, 2, \ldots$. This information is useful for risk managers in the process of risk evaluation and as an input for risk management Caballero et al. (2007); Hanewinkel et al. (2011). Government agencies and forest managers can also use relative risks to decide where to put the resources for fire control and fighting and provide a more efficient response in case of wildfire occurrence.

2.4 Discussion

The leading cause of wildfires in Castellón during 2001-2006 were lightning strikes, followed by negligences and arsonism (Figure 2.5). These three causes accounted for over 75% of the wildfires during the time span of our study. This supports the assertion that climate change and changes in land use are among the main factors associated to wildfire incidence (Pausas, 2004). For the particular case of Castellón, land use changes are the consequence of rural abandonment. Terrains with slopes lower than 30 degrees and closer than 400 m to the nearest road are relatively easy to access for humans. Most of the wildfires occurred in hill sides facing South or East, in slopes with less than 30 degrees. Hills sides facing East and South receive a higher amount of solar energy along the year, and thus the moisture content in the fuel in those hillsides is expected to be lower than in hillsides facing North or West, resulting in different risk of wildfire ignition. This was confirmed through formal statistical analysis. Higher temperatures are perhaps associated to the highest number of wildfires per year during 2005.

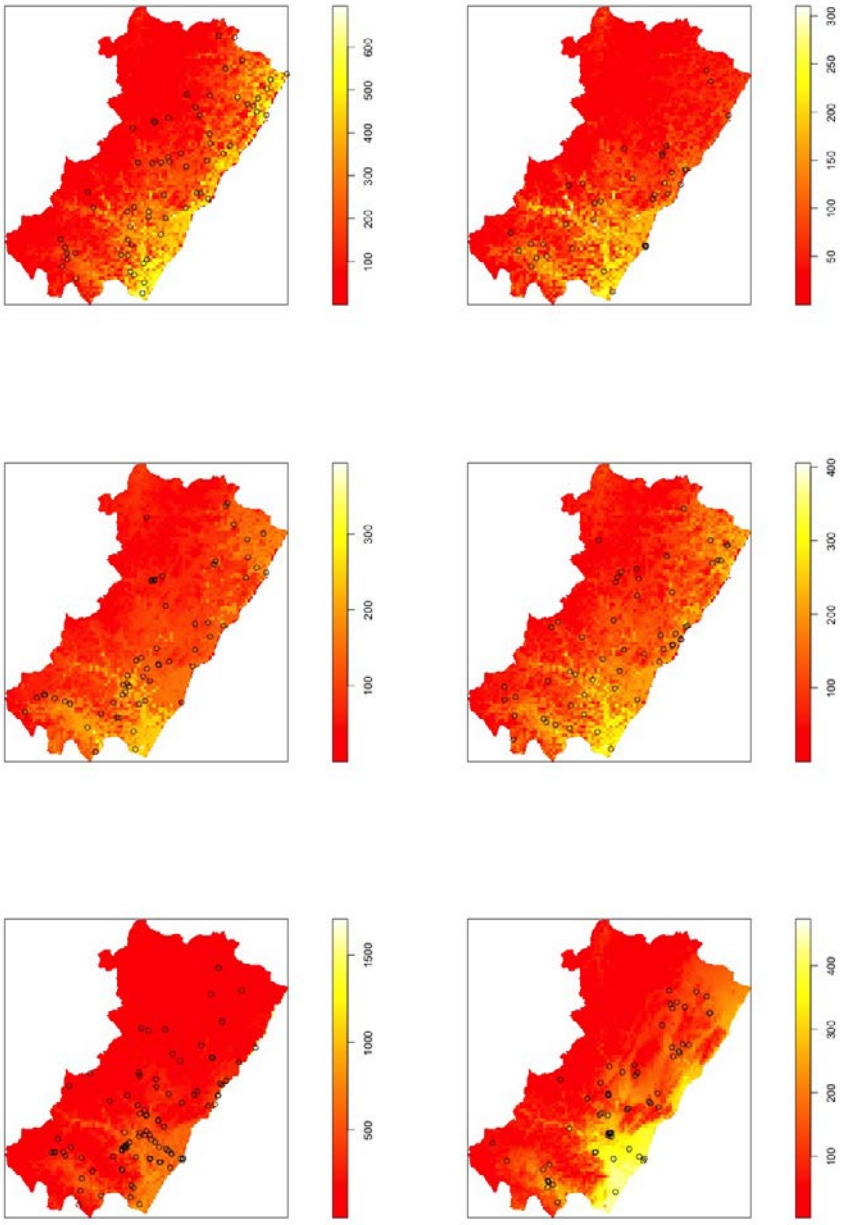


Figure 2.10: Estimated intensity function obtained from the area-Interation models for the years 2001 to 2006 in Castellón, human causes.

Wildfires tend to be less selective as their size increase, so factors such as vegetation type are less significant for bigger wildfires (Barros and Pereira, 2014). This explains in part why land use was not a significant covariate in the models.

For 2004 all the covariates showed a positive association to wildfire incidence in Castellón. For year 2001 on the other hand the elevation, slope and isothermality were associated to increase in the wildfire incidence whilst distance to the nearest road and soil permeability were associated to lower incidence. For the rest of the years there was also variation of the signs of the coefficients, indicating that the effect of most of the covariates on the wildfire incidence is not constant a long time. Slope was the only covariate that showed positive association for the whole study period for naturally-caused wildfires. This variability in the effects of the covariates is indicative that naturally-caused wildfires may occur in the whole province of Castellón but show a slight association to physical and climatic factors, which show variation along time.

Elevation and permeability showed negative coefficients, something that could be expected because the higher the elevation the lesser human presence in the province. Because soil permeability can be considered a proxy of fuel moisture, the negative coefficients for this covariate suggest that human caused wildfires tend to occur in areas where fuel moisture is lower. On the other hand, slope and isothermality showed positive coefficients for the six years. The positive coefficients for isothermality indicate that human caused wildfires are more likely to occur in areas where temporal variation of temperature is low, perhaps due to the higher population density in those areas. In Castellón such areas are found mostly in coastal areas, where dry and hot weather conditions prevail most of the year. Regarding slope, the coefficients have the same sign as for naturally-caused wildfires, which suggests that the effect of slope is the same regardless of the cause, and the positive values may be due to the fact that fire propagates easier in steep terrains, facilitating a small fire to become a wildfire.

The relative risks between the different land use categories were not constant across time. This may be because in a given year wildfires may occur at a slightly higher rate in areas with a particular land use or have larger wildfires. This brings as a result that for the next year the fuel loads are lower and therefore the expected number of wildfires will be smaller. This is in fact the idea behind the use of controlled fires to reduce wildfire incidence (Hutchinson et al., 2005). The presence of interaction between the wildfires is probably a consequence of the burned area in previous years, which left a patchy arrangement of unburned areas where wildfires can occur. It is not clear the reason for the change from repulsion to clustering in the observed wildfire pattern, although a possible explanation is that the repulsion during years 2001 and 2002 for human related causes and 2002 and 2005 for natural causes left wide unburned areas that became fire prone

the last two years. Another possible explanation is that the number of wildfires and the burned area are not too high and therefore the change from a regular distribution pattern to a clustered one is due to the variation in the number of wildfires along the years under study. The area interaction model covers a wide range of possible scenarios, going from regular patterns with strong repulsion to point processes with moderate clustering. The fit to the wildfire data of Castellón was acceptable, as no empirical $L - function$ lied outside the confidence bands obtained by simulating point patterns using the parameter estimates for each year.

For human caused wildfires the high intensity areas are located near coastal areas, in particular in the southern part of the province, where the capital city Castellón is located (Figure 2.10). This Figure shows a well defined pattern, associated to the areas where most of the human population (about 80%) lives, making those areas more likely to have a human caused ignition, which combined to the dry weather prevailing in the coastal areas makes the presence of a wildfire to have a higher probability of ignition. The expansion of the low risk zone is in part because of the increased amount of rain, which has a direct association to elevation and to the reduced human presence due to farm abandonment. Overall, we see from Figures 2.9, 2.10 and that in terms of wildfire risk the province of Castellón may be roughly divided in three zones: A high risk zone comprising the coastal plains where most of the human activities take place, a medium risk zone is distributed in a central strip in the NE-SW direction, which contains areas where farms and agricultural activities used to take place and where field abandonment has increased fuel loads in the recent years, and a low risk area, located in the NW part of the province, where most of the forest of the province are located.

Other authors have approached the problem of producing forest fire risk maps using methods such as GIS (Vázquez and Moreno, 1998), logistic regression (Preisler and Westerling, 2007), autologistic regression (Koutsias, 2003) and empirical generalized linear models (Barbero et al., 2014) among others. Most of such approaches are either based on pure descriptive methods or ignore the presence of spatial dependence observed in the data. The use of inferences based on pure descriptive methods may be misleading because the variability of empirical data sets, and therefore posing a model for the observed data as we have done, has the advantage of providing external validation to the inferences obtained from such models. Our approach also acknowledges the presence of spatial dependence through the incorporation of spatially varying covariates, allowing the assessment of the effect of changes in the covariate values on the risk of wildfires for different locations within the study area. Although posing and fitting models may be difficult in some cases, it is a clear improvement as compared to pure descriptive or non spatial methods of analysis of fire risk.

2.5 Conclusions

Decision making about forest ecosystems requires the integration of risk management in process. An important part of risk management is the assessment of risk of ecological disturbances. In forest ecosystems perhaps the most important perturbation is wildfire, as it may potentially influence the post fire successional process and induce landscape changes. The first step in risk assessment is the estimation of the probability of wildfire. The standard tool to model the probability of fires in forest ecosystems has been logistic regression as in Wittenberg and Malkinson (2009) or in Del Hoyo et al. (2011). Most authors however have used standard logistic regression, which assumes independence between wildfire events, ignoring the spatial association inherent in spatial processes. Here we have proposed the use of spatial point process models with interactions, a well known methodology that acknowledges the non independence between wildfire events. The general form of the model we used allows its application in wider areas as long as information about wildfire incidence and covariate information is available. It is worth pointing that the significance of the covariates used for modelling wildfire risk in a given area is likely to change when the model is applied in a different area, or if the study area becomes broader than the original. This however, is not a disadvantage of the model used here, as the general form allows to adapt the model in case drastic spatial changes in vegetation composition or in topography may occur. Our results show that the spatial variation of forest wildfire risk is associated to the environmental covariates and to human factors considered in our analysis. Distance to nearest road was used as a proxy for human activity in the study area, but the model fit may improve if information about population density, socioeconomic activities and other related information becomes available. Castellón's province is a mountainous area and forest are restricted to the interior part of the province. Unlike what happens in other provinces of Spain, such as Catalonia, where forest wildfire risk is associated to abandonment of agricultural land (Serra et al., 2014), wildfire presence in Castellón is higher in forest covered areas. However, for the last two years analysed the higher risk zone occurred in the flat area close to the sea. This change in the spatial distribution of wildfire risk is likely associated to non observed human related covariates, as the vegetation cover and geographical features of the province did not change during the time span of this study.

We have compared the incidence of wildfires separated in two wide classes, which has showed that both classes have a different association with the climatic and physical factors included as covariates. Naturally-caused wildfires do not show a specific spatial pattern for the different years but are nevertheless slightly associated to the covariates we have included in our study. On the other hand, human caused wildfires show a well defined pattern with higher incidence associated to the covariate values observed in

coastal areas.

Overall, we see from Figures 2.9 and 2.10 that in terms of wildfire risk the province of Castellón may be roughly divided in three zones: A high risk zone comprising the coastal plains where most of the human activities take place, a medium risk zone is distributed in a central strip in the NE-SW direction, which contains areas where farms and agricultural activities used to take place and where field abandonment has increased fuel loads in the recent years, and a low risk area, located in the NW part of the province, where most of the forest of the province are located.

Although kernel estimates have been used previously to study the spatial distribution of forest fires incidence (Avalos and Alvarado, 1996; Koutsias et al., 2004; De la Riva et al., 2004; Amatulli et al., 2007), those are only non-parametric estimators are not useful for spatial prediction nor to asses the existence of association between wildfire incidence to suspected risk factors. Parametric models such as the ones fitted in our study besides providing predictive estimators for the association between the intensity function of wildfire incidence and are also useful to asses the effect of changes of those factors on such intensity function.

The statistical modeling approach to forest wildfire risk mapping provides to forest fire managers a valuable tool for planning forest wildfire prevention task and surveillance. These maps show not only areas with high wildfire risk but they also may be used to distinguish between those forest areas with high risk of wildfire by natural causes and those with high risk of wildfire because of human activities. Point process methods are a sensible approach to model the probability of wildfire ignition. However, this approach works better if reliable data bases including historical records of forest fire locations as well as digitalized maps of spatial varying variables are available to the modeler. Such data bases are becoming available due to the increasing quantity of data acquired with remote sensing technologies and the increasing accessibility to such information. Thus, it is reasonable to expect that the number of applications of spatial point process models to forest fires data will increase in the near future and that improved models and fire risk maps will become available to risk managers and to general public. Predictive models for the final size of wildfires and the factors influencing such size are yet to be developed. Better management of wildfire risk is therefore a goal that is being attained as research continues.

Acknowledgements

We would like to thank the Department of Infrastructure Territory and Environment of the Generalitat Valenciana, Section of Forest Fire Prevention for the assignment of geographic data.

Chapter 3

Participative site-specific agriculture analysis with low technology approach

Precision agriculture or site-specific farming can be defined as a method for managing soil and crop production in a spatial and precise manner, taking into account the conditions of various parcels that, when combined, define the farming land (Schueller, 1992). Sometimes this concept is mixed with high technological equipment such as computational devices, Global Positioning Systems (GPS), Geographical Information Systems (GIS) and others. Technology providers and developers are pushing users to adopt the newest technologies (Lamb et al., 2008). Nevertheless, precision agriculture is not a high-tech discipline by definition (Molin, 1997), but is based on an "observe-interpret-evaluate-implementation" methodology regardless of the means used (Cook and Bramley, 1998). Furthermore, a low technology approach is also suitable for precision agriculture developments (Bouma et al., 1999).

Precision and site-specific farming became an attractive idea for most farmers and agriculture experts in developed countries as a method for optimizing agricultural production (Roberts et al., 2004; Sassenrath et al., 2008; Cook et al., 2003) (Mann et al., 2011). Precision agriculture has been based mostly on information technology, high levels of machinery and computation knowledge. This refers to an increase in economic resources as "input". For example, the application of high positional accuracy involves implementation costs (Booltink et al., 2001) and training time. Nevertheless, even in developed countries, using the latest technology in precision agriculture is not as widespread as believed (Lamb et al., 2008).Moreover, in Southern Europe the use of site-specific agriculture "has been delayed because of small

Chapter published at Precision Agriculture October 2012, Volume 13, Issue 5, pp 594-610 http://dx.doi.org/10.1007/s11119-012-9267-4

farm size" (Fountas et al., 2010) among other reasons.

Precision agriculture is more feasible when the farmland is larger or based on the educational level of the owner (Roberts et al., 2004). What are the effects on precision agriculture based on these two factors? Site-specific concepts remain the same, regardless of the farm size and the farmer's educational level. Nevertheless, technological equipment and the implementation of sophisticated procedures and their application pose a different question (Aggelopoulou et al., 2009). Computation, machinery and even education are scarce in rural environments (Diagne, 2009). Even without the potential of being able to use high technology, small farmers are still able to apply site-specific concepts and ideas by just referring to paper maps. This is possible because small farmland owners are more familiar with their own land (Altieri, 2004). Since most smallholders are traditional families that have lived on the land for quite some time, they can utilize their "mental maps" to manage their land (Cook et al., 2003). However, it is important to provide farmers with environmental and agricultural education by using a methodology that will allow them to make appropriate decisions (Ma et al., 2009).

Farmers can acquire information in a short time frame by observing the environmental resources and production, consequently learning how to improve the management of their land. For example, farmers tend to know which part of the land might be better than another part by simply observing the crop's progression. These observations are in fact low technology site-specific information that can and should be applied.

Many smallholders already have the idea of site-specific management in their minds (Cook et al., 2003), even if it is in their subconscious. An example of this is the fact that limited quantity of fertilizer has to be applied to only a specific location where and when it is needed and not evenly spread across all the farmland (Stoorvogel, 2006). Other research has found that farmers know their farm's features and variability (Booltink et al., 2001). Hence, is possible to follow a site specific precision agricultural methodology to manage the small farm.

In the geospatial information community, data collection is moving, from a top-down approach to a bottom-up approach. A top-down approach is a traditional way of collecting data by official institutions and experts. A bottom-up approach means that ordinary people are working as voluntary 'sensors' (Goodchild, 2007). People can be like sensors providing spatial information directly from a source. In this scenario experts can collect spatial information, but can also take advantage of an individual's (such as farmers/samllholders) contribution as a voluntary 'sensor'. The farmer's individual contribution is part of a wider contribution collected by the expert. This way of collecting spatial information has been called participatory GIS (PGIS) by Sieber (2006). PGIS facilitates knowledge and learning interchange between participants (Hall et al., 2010)

3.2

3.1 Problem

Smallholders know their land. They know which areas are the best, and they can estimate their crop yield according to their observations. Nevertheless, this appreciation and knowledge is not recorded and shared. In contrast, experts have academic and technical knowledge. Experts can provide advice to smallholders based on their knowledge and on information provided by smallholders. This information exchange among experts and smallholders is done mostly orally without any documentation.

This concept in developing countries is similar to developed countries. Moreover, developing countries are usually faced with the fact that a good portion of the population has a lack of knowledge, expertise and access to the 'digital world' that surrounds many others; it has been called the digital division between developed and developing countries (ITU, 2010). This fact is even worse in rural areas where the penetration of digital technology is lower than urban areas (James, 2008). This gap is filled mostly but not always by "leading farmers" who are often more highly educated, or fill a local/regional "leadership role" (Lamb et al., 2008), who are early adopters of technology for precision agriculture.

The problem raised here is how to communicate the concepts of precision agriculture or site-specific management strategies to smallholders in those cases where it is potentially feasible without the introduction of high technology. To implement the methodology a cooperation among smallholders and experts is needed to interchange information and advice.

The methodology developed might be applied to smallholders in developed and developing countries.

3.2 Objective

The aim was to find out if it is possible to perform monitoring and analysis of farm production and management, implementing site-specific agricultural principles and low technology dependency within a participatory context. This research focused on the use of precision agricultural concepts (Srinivasan, 2006) and techniques for smallholders. The strategy will use as much of the available concepts as possible without introducing new technologies to the smallholders. For example, paper maps and spreadsheets will be used for collecting the data rather than GIS and GPS.

The goal is to ensure information exchange in a participatory context between the smallholders and the experts. This participatory approach has had higher adoption rates (Booltink et al., 2001) than a non participatory approach. Data collection will be carried on by the farmer, using paper maps to later on share the data with the experts. Experts will be able to process the data and analyze it, to get spatial information, and finally

provide feedback to the farmer in a personalized way.

The result of this methodology will be a continuous collaboration among the different participants. The participants will interchange information: for example the expert does not know the field as well as the smallholders; on the other hand, the farmer requires advice regarding advanced agricultural topics. The use of paper maps will help to improve spatial communication among different participants and integrate the collected data with other data sources (Van Wart et al., 2010). The farmer should be able to collect spatial information as easily as possible, by taking advantage of his/her knowledge so as to locate the crop variables on a paper map. Experts require information in GIS format in order to work with it. Hence, paper maps need to be introduced in GIS to process this data.

3.3 Methodology

The proposed methodology is based on a bottom-up (Goodchild, 2007) approach to get spatial information. The farmer will collect spatial data using paper maps and spreadsheets. This data will be shared with an expert, who then uploads it into a GIS application. Then data analysis can be performed to provide feedback to the farmer.

The principles of site-specific farming (Srinivasan, 2006), are: reducing costs, optimization of yield and quality in relation to the productivity capacity of each site, better management of the resource base and protection of the environment. A farmer has to be able to gather information about his field in a way that spatial and temporal variation of the field conditions can be recorded and archived. The collected information should be quantitative, in order to be able to do a critical analysis. Nevertheless, qualitative information may also be recorded as the farmer deems useful for crop management. With the input and output records and expert feedback, the farmer can perform site specific management of the field, according to predefined objective parameters.

As shown, these principles (Srinivasan, 2006) mention spatial and temporal management strategies, and do not focus only on the technology that makes it possible. Therefore, it is possible to say that the use of Global Positioning Systems (GPS), Remote Sensing, GIS and other electronic and expensive equipment is not essential (Cook and Bramley, 1998). It is possible for small holders to take advantage of his/her field knowledge to locate and represent different variables spatially. There have been efforts to gather information from the orange crop yield using a differential GPS(DGPS) (Whitney et al., 2001) *"to have enough accuracy to record the geographical position of trees. Nevertheless, the spatial information will be represented using paper maps instead of using a DGPS o GPS with not enough accuracy"*.

For this research, the principles of precision agriculture or site-specific agriculture (Srinivasan, 2006) will be applied to a small orange tree farmland in Spain. There are two participants involved with different technological experience. The first will use GIS technologies (expert user) that are not always accessible to smallholders. The second will use only paper and pencils (farm user) while still implementing the principles behind precision and site-specific agriculture.

3.3.1 Study Area

The Study area is located in La Vall d'Uixó, Valencian Community, Spain. The area of the study is a single field with an area around 2.71 *ha*, Fig 3.1, with only one owner. The owner has another job not related with agriculture. The field is cultivated with orange trees with an irrigation system based on drip irrigation. The orange tree varieties are mostly Clemenules and one parcel of Hernandina. The field was transformed from an olive and carob tree cultivation to the present orange trees between 1951 and 1968. With this change, a new layer of soil was laid in the area to improve soil conditions for the new cultivation.

This farm is representative of the Valencian Community's most common orange orchard farm. In the Valencian Community 3 *ha* is the average size of an orange farm (M.A.P.A, 2003). 78% of farmers have another job not related with agriculture (M.A.P.A, 2003)

Expert User

The steps followed by the technology user to upload inputs and spatial information into a GIS and perform the analysis using GIS tools are described below.

- Digitization of the field boundaries and parcels

  The most important issue in site specific or precision agriculture is location. The location is needed to assign inputs and outputs, in order to perform a posterior analysis, focused on the results per parcel. In other words, site-specific management can not be performed without the data being associated to a concrete location.

  The map of the field must be drawn for this task; the tool used for this task is going to be gvSIG[1] software along with the use of PNOA[2] (Plan Nacional de Orotofotografia Aerea) aerial image and the layer of the Spanish land registration office[3] WSM (Web Map Service).

[1] http://www.gvsig.org
[2] http://pnoa.ign.es/
[3] http://www.sedecatastro.gob.es/

Figure 3.1: Aerial image of the Study Area. Study area parcels marked with red lines.

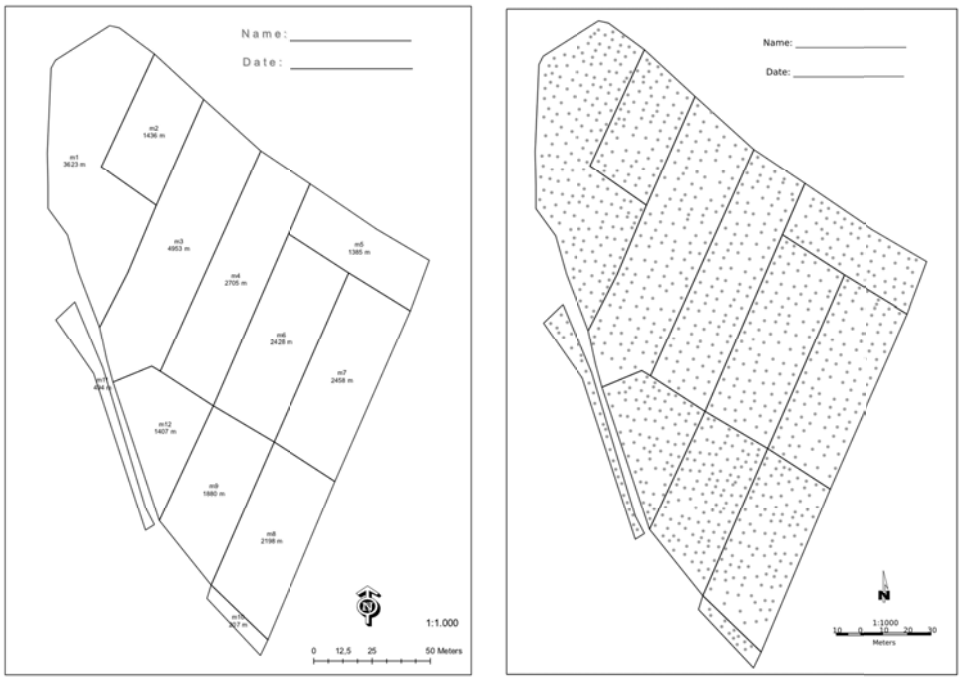


Figure 3.2: Map provided to the owner by the expert. Left map with parcel name and area, right map shows trees' position.

The orange cultivation in the Valencian Community is traditionally performed on a terrace; each terrace has similar characteristics of soil, and tree variety. Moreover, the tasks performed in the field are planned on the basis of the natural division in terraces with homogeneous conditions. In this case, a regular grid would not apply, this technique is more adequate for cereal species cultivation. Hence, the subdivision of the field is going to be digitized according to the terrace distribution, (Figure3.1). It is also the same division used for the farmer's handmade map. The use of the PNOA image allows for the digitization of tree position. The expert user provides a paper map with the parcels (Figure3.2), based on the farmer's handmade map and tree position.

- Data collection

The data collection is going to be done exclusively by the farmer. In the case of smallholders, it is sometimes it difficult to discriminate outputs from each parcel. Therefore, an effort has been made to measure the crop yield for each terrace. Yield was measured by harvested oranges per terrace. The extra amount of fertilizer dispensed by the farmer is measured in kg per parcel and tree. On the other hand the

orange data quality is measured by the Orange Packing Cooperative, where the fruit is processed. The Co-operative provides feedback to the farmer with a report of the orange yield quality of the farm, not per terrace. The farmer estimates the orange quality per terrace according to his experience. Hence, parcel orange quality is not considered in this work because it is an estimation.

- Data translation to a GIS

  The tool for data translation is gvSIG. Data is stored in a postGIS [4] database using gvSIG as an interface. GvSIG and PostTGIS are available in several languages, such as open source which have on-line documentation and tutorials. GvSIG has a mailing list to help users. Data analysis in this case is based only on the computation of some crop yield production parameters, such as harvested oranges in kg/ha, kg/tree and difference between years.

- Information feedback

  Feedback information could be provided to the farmer using printed maps or an on-line map server to which the farmer could connect using a simple client visualizer. Geoserver [5] was chosen to provide on-line spatial feedback to the farmer. Figure 3.3 shows the spatial information work-flow. Paper maps produced by the farmer can be available as historical records in GEO-TIFF[6] format.

3.3.2 Farmer user

In a non-technology approach, the farmer must use the site-specific management tools and principles (Srinivasan, 2006) without the implementation of high levels of GI technology.

- Field boundaries and parcels

  The farmer has a sketched map of his land divided into parcels for managing reasons, although the farmer is going to use the output map from gvSIG, provided by the expert user. All parcels have been measured and labeled according to their area and parcel number. The farmer is also provided with a map, Figure 3.2 of the land describing

[4] http://postgis.refractions.net/
[5] http://geoserver.org
[6] http://trac.osgeo.org/geotiff/

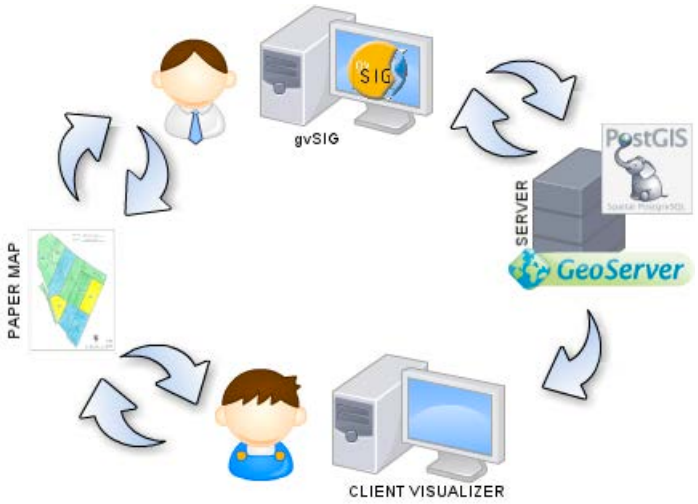


Figure 3.3: Spatial information work-flow.

the trees' positions.

- Data collection

  The orange tree production is provided by the owner of the field as is explained in section 3.3.1. Spatial data is collected using paper maps and spreadsheets. There is no difference in the data acquisition for the expert user or the farmer user.

  Production will be shown in $kg/1000\ m^2$ to be more easily interpreted by the farmer. A colored classification labeling system will be created by the farmer for the orange production where a total of 4 classes should be applied. The farmer, according to the field outputs, will color each parcel. The result is a map with production information and easy to view colored classification. Trees' data is annotated freely by the farmer with just one condition, the annotation must be clear for the expert user.

  Each production year map will be stored as a hard-copy document to be used in the following years as a guide to fill in the information in the same way. The owner will be able to modify the inputs or perform special care according to the analysis of the paper maps and the expert's feedback. For example, parcels with good productivity in the last years will receive less or no input, whereas less productive areas should receive more input or special care.

3.3.3 Participatory GIS

The farmer contributes to the process providing data from his farmland. The expert user gets data from the farmer and uploads it to the GIS making the data available for other users. Expert users can provide spatial feedback to the farmer with processed information or with other spatial information that is considered important for the farmer. This methodology will provide a dialogue between the farmer and the expert with a never ending work-flow of information, look at Figure 3.3.

3.4 Results

According to the steps described in the Methodology the results for the farmer user and for the expert user are described in the following sections.

3.4.1 Farmer user

The farmer user used paper maps to graphically describe the production of the parcel, look at Figure 3.4. The farmer drew the production to create the map using a quantitative scale, but also performed fast mental calculations to estimate the production maps regarding the $kg/1000\ m^2$. The farmer did this because he thought it would be better to compare the production between parcel, and also has noticed that it will be better to have the production in $kg/trees$, because there are parcels that have some young, old or ill trees that are not producing oranges. This observation by the farmer has been taken into account by the expert to produce a map with the location of the trees and to compute the $tree/kg$ ratio. Paper maps with tree locations were used by the farmer in successive seasons to annotate qualitative information about the trees, such as old trees, new plant trees, non productive trees, and special care (fertilizer addition) trees. The first orange production paper maps motivated the farmer to draw a map with his own observations about the soil quality.

The time used by the farmer to record the information on a map, Figure 3.2, and spreadsheets was around 4 hours, having all the data previously collected and distributed per parcel. Finally, the data collected by the farmer is the following:

- Production for seasons 2007-2008 to 2010-2011.
- New plant trees (2007-2011).
- Old trees 2011.
- Yellow trees with extra addition of special fertilizer (2009-2010).
- Qualitative observation of soil quality.

Figure 3.4: Colored map of the production distribution in the parcels. Seasons 2007-2008 to 2010-2011.

3.4.2 Expert user

It is necessary to point out that input data is provided only by the farmer. The use of GIS software is an advantage, because with this tool it is possible to implement faster computation processes and advanced analysis using spatial data. Spatial information can be displayed in a more interactive way by using GIS applications. GIS allows for changes of colors as well interactive information display. Furthermore, computations between table's fields can be made. The $tree/kg$ ratio has been calculated with gvSIG by the expert user. The computations in this case are simple. The work done is:

- Calculate Surface in m^2, this surface is the same used for the farmer for his maps.
- Ratio production per tree and per parcel surface. The ratio allows for direct comparative between parcels.
- Tree information is uploaded into the system according to farmer paper maps.
- Spatial data uploading to Geoserver.
- Paper maps uploading to Geoserver.

The Figure 3.5 shows output maps from gvSIG. The use of a GIS tool allows creating maps interactively, changing the scale and the interval thereby enhancing the way the data is displayed. The farmer receives the gvSIG output maps from the expert. The farmer can also retrieve spatial information from the Geoserver using a light client visualizer, Figure 3.3. These maps contain more information than the paper maps that the farmer has provided to the expert.

3.5 Discussion

The farmer likes using handmade maps, as they are easy for him to create. He can use these maps to follow the increase or decrease in production. The difference in the production maps clearly shows the parcels that have increased in production, and those that have maintained the same level of production. Consequently, action plans will be defined according to the results of each parcel. The farmer suggests an improvement for future maps by adding orange quality, which has a relationship with the final price. Nevertheless, the orange's quality refers to the total amount of the farm's production, as was explained in section 3.3.1. He says that in the future he is going to follow this system. The farmer also wants to perform a prediction of the yield using tree flowering (Aggelopoulou et al., 2009), to later on compare

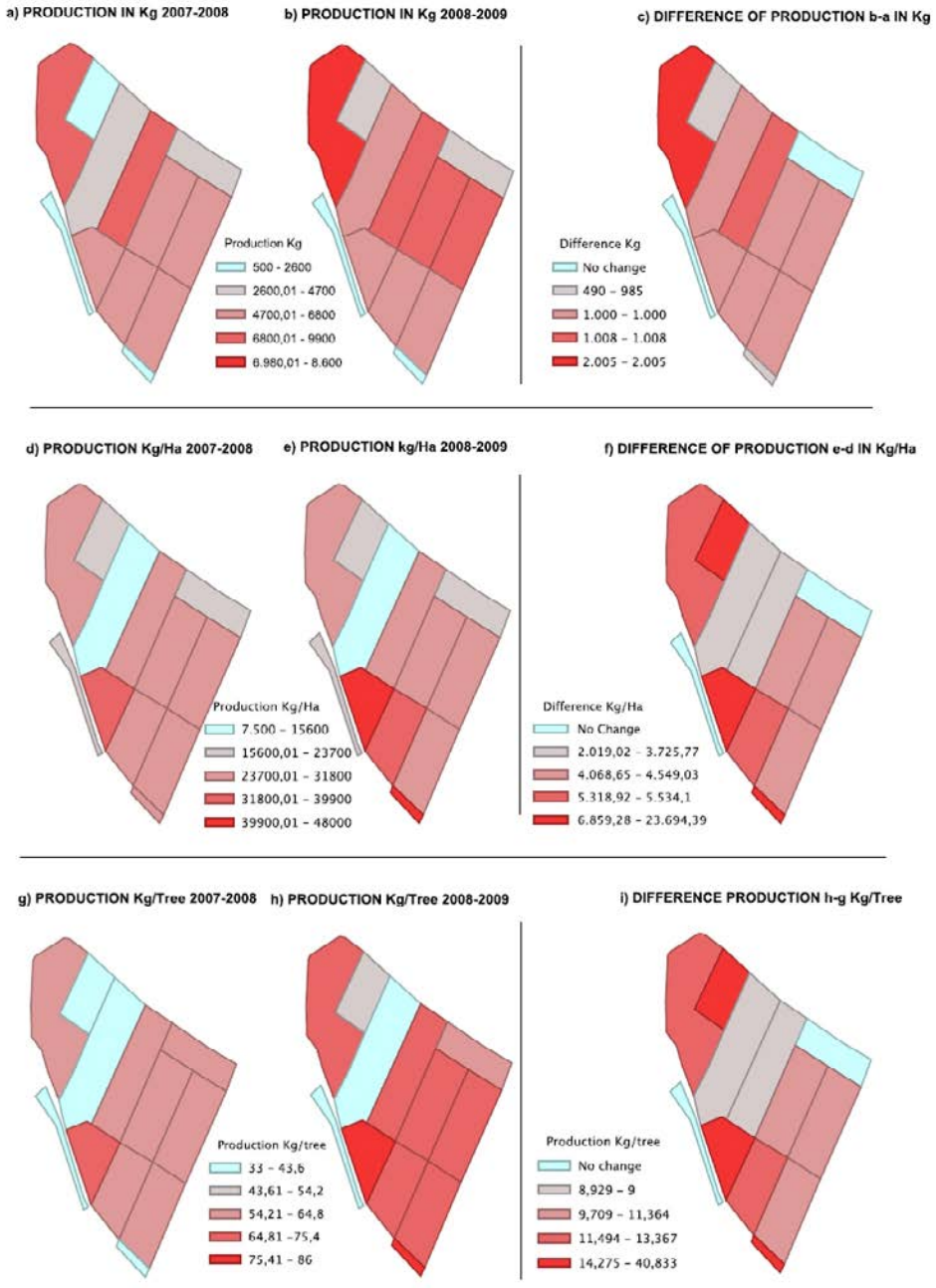


Figure 3.5: GVSIG output maps of the production distribution in the parcels. Seasons 2007-2008 and 2008-2009.

it with the real production. All these new estimations and annotations will be done by the farmer using the handmade maps.

The GIS tools provide a more efficient way to visualize the results, changing the intervals or setting visualization in a variety of ways. Using GIS tools at the farmer's knowledge level is not significantly different from using handmade maps, the farmer also can perform hand calculations of the ratios and draw them on the maps. Nevertheless, if the farmer provides the amount of production to the expert he can receive the production result in kg/per tree. The difference for the farmer is the GIS output visualization.

The farmer has said, he will not attend a basic course of GIS tools for producing the maps, because he admits that: "I'm too old (64)", but he will continue with the methodology of using handmade maps, recording more parameters of the field inputs and outputs, such as quality. Using this data he will produce some maps, plan his tasks and in some cases show the data and paper maps to a consultant expert.

The farmer prefers the maps created with GIS, as he can see more things with these maps than with his handmade ones. He can easily visualize more information in different ways. He has noticed the evolution of parcel m12 where he has added fertilizer, because the trees had symptoms of a low level of Iron. The production has increased in this parcel. GIS maps provide feedback by the expert, and provide a visual description of the situation.

This is collaborative approach to data collection directly from the source, the farmer. With this data the expert can complete his/her spatial information with a wider overview of the situation and the farmer's concerns.

3.6 Conclusion

Is it possible to perform site-specific management of the farm with a low technology dependency using a participatory GIS approach? The next points are the conclusions of this research:

- The methodology for farmer user provides useful and easy instructions to follow.
- The handmade maps provide enough information to allow the farmer to understand his crop situation.
- GIS outputs provide extra information to allow the farmer to analyze the current situation.
- Precision agricultural principles (Srinivasan, 2006) can be followed for smallholders without the help of high level technology dependency.
- A consultant expert is always needed and can guide the farmer in the recollection of the data, and in the management decisions.

- The work-flow provides a dynamic dialogue between the farmer and the expert. Both actors can benefit from this collaborative approach

The farmer realized that he had all the data in his mind and in some of his notes, but only after making the maps, he was able to point out what was occurring in each parcel and why. This exercise has provided the farmer with a new tool to collect and improve the management of his field according to the obtained results. The expert user can also use the information provided by papers maps. If this methodology is adopted by the farmers of a region the expert user will have an overview of the past and current situation. The participatory methodology provides the expert an overview about the farmer's concerns. The farmer receives feedback by an expert giving added value to the data provided by the farmer. As the spatial information is centralized in a server, different experts can have access to the data to analyze it and give feedback to other users or to the farmer. There is a server where all the data is stored and can be retrieved if needed.

This paper explores the possibilities of involving smallholders in the process of decision making together with experts in a participatory approach using paper maps and geospatial technologies. The use of a participatory methodology with feedback by the expert to the farmers will provide a significant change in the adoption of low-tech site-specific agriculture. The farmer will continue providing more data to the expert as far as the expert will provide spatial information and useful advice to the farmer.

Future work is need to test this methodology on a larger scale. This testing will require the participation of farmer communities, associations or cooperatives initiatives.

Chapter 4

Quality

Since 1980 has been done and effort to describe GI quality (Devillers et al., 2010; Goodchild and Li, 2012). The quality of geospatial data and information refers to how well a real-world object (such as a road or tree) is represented digitally in a geospatial database Devillers and Jeansoulin (2006) according to its fitness for final use (Veregin, 1999). With regards to GI quality may focus on a core strategy: spatial data quality associated with standards or uncertainty (Devillers et al., 2010).

Official GI quality parameters are divided into the following seven parameters (Van Oort 2005; Devillers et al. 2010; Goodchild & Li 2012):

1. **Lineage**. This aspect concerns data history, the origin or source of the data, and the processes implemented in order to acquire the data.

2. **Positional accuracy**. How accurate is the object's position in relation to the earth.

3. **Attribute accuracy**. This measures how accurately an object has been tagged with its attributes. The attribute accuracy is measured with a misclassification matrix or confusion matrix.

4. **Logical consistency**. Concerns the fidelity of the representation and its consistence in terms of topological correctness and the relationships encoded in the database.

5. **Completeness**. Completeness checks for gaps in the data for the coverage region. Results from this parameter include missing objects and over-represented area. Completeness gives the user an idea about data coverage representation for a given area.

Chapter partially published at 7th International Symposium on Spatial Data Quality (ISSDQ 2011). Raising awareness of Spatial Data Quality, pp 109-114. ISBN:978-989-95055-8-2

6. **Semantic accuracy**. How close is the description of the spatial object to the meaning it represents (Haklay et al., 2010).

7. **Temporal quality**. This describes the temporal resolution, e.g., how fast a spatial object is captured and represented.

A second block of parameters (Van Oort, 2005; Devillers et al., 2010)

1. **Variation in quality.** Describe the differences in the quality elements along the geospatial database.

2. **Metaquality.** Information describing the quality parameters measured for the GI. For example if a quality element (positional accuracy, semantic,..) of the data set is estimated from a smaller sample size then the metaquality is low.

3. **Usage purpose**. This aspect represents the fitness and constrains of the geospatial information for the user goals (accuracy, coverage region,...)

Metaquality as well as variation in quality are elements that cannot be extracted directly from GI. Instead they are derived from quality parameters in the first block. The usage purpose is strongly related with other quality elements.

Apply this quality parameters directly to VGI data will fail, because were designed for a complete different data source such as professional GI or official topographic maps. This chapter addresses some of the challenges in generating GI quality for VGI. We explore some measurements of quality parameters for VGI in order to increase the utility of VGI. VGI has to afford re-usability and additional value for others within various alternative scenarios such as spatial accuracy for road navigation or object annotation.

4.1 Fitting quality within VGI context

VGI has some advantages, such us every citizen is a potential contributor (Goodchild, 2007), facility to register changes (Coleman, 2010) or local knowledge (Wiersma, 2010) among others. Despite the advantages of VGI, such as massive amount of information and local knowledge (e.g., billions of citizens around the world that are familiar with their local environment) there are quality issues associated to this kind of information (Aragó Galindo et al., 2011).

A VGI project is built by users for users. It is based on the collaborative work of citizens and is more focused on the act of uploading geographic objects and attributes rather than being critical of the contribution's spatial quality (Goodchild, 2008). A user contributes to a VGI project according

to their skill-sets and capabilities. Hence, a VGI project is populated by data with varying levels of quality. For instance we could find users with professional skills whereas other users don't have experience at all.

On the other hand, an "official" project starts with a work-plan and criteria such as, defining the project's minimum quality level, scope of the work and the usage of a data source. Therefore, GI is built and checked to fit all the features within a previously defined quality level (e.g., ISO19113, ISO19114 and ISO19138 (GIS., 2005)). Applying "official" geospatial data quality criteria to a VGI project might provide poor results since the quality criteria could have not been established, be simply recommendations or depend on user skills. Thus, VGI sources may have poor spatial data quality control. Nevertheless, this quality element can be adjusted according to the idiosyncrasy of VGI, for instance projects in which contributions are only made by a selected group of individuals, or specialists, for which the error rate in positioning and attributes is low. Last, we need to take into account the user's freedom and flexibility to contribute to the VGI project. This freedom should not be much restricted since it can reduce also data provision rate.

The quality issue could be assessed by comparing the accuracy and quality of VGI data against similar "official" data (Haklay, 2010), in other words comparing with the spatial data produced by official institutions and governments. However, it is not always possible to compare VGI data with official data because sometimes a comparable "official" source does not exist. In other cases VGI is less structured and lacks description, to assess whether it is valuable as a complement to "official" data. Using the industry standard to measure quality.

4.2 Open Street Map data description and test parameters

In order to assess our methodology and parameter, we have chosen the OSM project. OSM data has been used in order to assess the quality parameters. OSM is a crowd-sourcing project to map principally roads, streets and other features of the world. OSM is base in the same philosophy that Wikipedia, everybody can contribute to the project freely. OSM data are readily available for download in .osm1 format, which is XML-base format. Its data model is basically composed of three different entities: nodes, ways, and relations. As an example, a road may be represented by several nodes or vertices, road is a way that contains references to all of its nodes. A relation links together the ways and nodes which are related geographically.

For evaluation purposes we have chosen study area which is located in the region of Valencia in Spain. The data covering this area was extracted from the OSM database using the bounding box defined as fol-

lows: left=-0.42778 right=-0.30967 top=39.51145 bottom=39.42346, EPSG 900013. The Spain.osm file was retrieved 14 June 2011 and has been imported into a spatial database (PostGIS) using the Osmosis and Osm2pgsql software tools. Note that OSM data is updated every day and this work has been done with a static database (available at section A.1) from the specified date. Valencia city was chosen because is a city with a large number of contributors, is well known by the authors and OSM data is quite updated (corresponding to fall of 2010).

OSM data has been use to test and define a VGI quality data approach for lineage, positional accuracy, attributes accuracy, logical consistency, completeness, semantic accuracy, and temporal quality.

4.3 Lineage

Lineage is defined as data history. VGI's basic data history could be the user name or id and the date of creation(time stamp). Nevertheless, depending on the VGI' project the basic data could be extended. An important issue for data history is automatic annotation without user intervention. There are some parameters such time stamp, editor, number of edition (version), that could be annotated automatically.

Furthermore, this information may be completed by defining the data source, such as a GPS. For example GPS tracks are one of the primary VGI data sources in OSM. In order for VGI data to have scientific value is necessary to provide data's lineage, however, in VGI projects such as OSM and Wikimapia data sources is recommended but not mandatory. Here is another point to be addressed to evaluate the data quality requirements for a VGI project.

From the Valencia's OSM data we can extract the following information: contributing user, version number, and updatting time stamp. This information is automatically generated at the same time a user creates a spatial object, meaning that the user does not have input it manually. On the other hand Valencia OSM data retrieved contains a source tag and note tag, which is information that is not mandatory but will increase the OSM data quality if it is available. Source and note tags are not required but are recommended. Therefore, annotating these tags requires an additional effort by the user when publishing geospatial information.

Table 4.1 shows the number of ways with lineage information for the city of Valencia, Spain. The unrequired lineage information is provided for less than 2% of the contributions. This means that users are not able to know the data source from 98% of the data.

Every OSM object has an attribute with it's version number. This version history could tell us the number of revisions done to each object. We examined the ways and number of revisions of our data which is shown in

Table 4.1: Valencia OSM ways features with lineage information tag.

	Total	Percentatge
user	8306	100,00%
version	8306	100,00%
time stamp	8306	100,00%
note	3	0,04%
source	101	1,22%
Source survey	58	0,70%
Source other	43	0,52%
Source URI	0	0,00%

Table 4.2. The results are that more than 23% of the data has only one version and has never been updated. The numbers of versions indicates the number of reviews done; which affects and improves the quality of the feature. When a way has more revisions the possibilities to have a higher and logical consistency theoretically are greater (See section Atributte acuracy).

4.4 Positional accuracy

Depending on the positional accuracy geospatial data has limitations in its usability. For instance, a car navigation device based on the Global Positioning System (GPS) needs cartography with accuracy at least equal to the GPS device's accuracy or in other words sufficient accuracy to place the car on the right street. On the other hand, geospatial data for civil engineering sometimes requires an accuracy of less than one centimetre (Lima et al., 2006). In other words, this parameter depends on its fitness for use (Devillers et al., 2010).

VGI is generated by users by tracking roads and getting points coordinates using usually GPS device or referencing landmarks manually from digital cartography. With a GPS device, positional accuracy depends on the GPS's receiver. Nevertheless, this accuracy could be annotated when information is introduced directly from the GPS receiver (Meng, 1998). If data is exported using the GPX[1] format, signal accuracy could be annotated within the file, but this is not mandatory. In the cases where there is no possibility to get the GPS accuracy it is assumed that GPS accuracy is +/-100 meters(Kong, 2007). Commercial GPSs such as Garmin[2]have an accuracy average of 15 meters. A study by Wing et al. (2005) tested different consumer-grade GPSs and their accuracy was from 5 to 10 meters. GPS information may also be saved in RINEX format (Gurtner, 2007) which is used to post process GPS information to estimate its positional accuracy.

[1] http://www.topografix.com/gpx/1/1/
[2] http://www8.garmin.com/aboutGPS/

Table 4.2: Version and corresponding amount of features (ways) in Valencia City.

Version	Features	Percentage
1	1912	23,06%
2	3365	40,59%
3	1547	18,66%
4	687	8,29%
5	298	3,59%
6	180	2,17%
7	108	1,30%
8	52	0,63%
9	44	0,53%
10	26	0,31%
11	22	0,27%
12	8	0,10%
13	6	0,07%
14	12	0,14%
15	9	0,11%
16	1	0,01%
18	3	0,04%
19	3	0,04%
20	1	0,01%
21	2	0,02%
22	3	0,04%
24	1	0,01%
26	1	0,01%

Table 4.3: Comparing digitized road distance in meters to the official cartography BCN25 from IGN.

	Mean	Variance	Stand. Devi	Minimum	Maximum
Distance to IGN road	2.36	1.49	1.22	0.23	6.82

Moreover, an editor could be making human mistake manipulating the GPS device.

Another possibility is georeferencing using reference maps or cartography. In this case the map resolution is the highest reachable precision for georeferenced data (Goodchild, 2001). To acquire accuracy for VGI according to reference maps there are two possibilities:

1. With raster layers as a reference source restrict georeferencing resolution to fit the raster resolution. In this way digitalized accuracy is similar to the raster resolution.

2. Define the positional accuracy as it relates to the zoom level. To apply this method the system must know the spatial resolution of the reference geospatial information displayed at each zoom level.

Even if data digitalization is restricted, human errors are always possible, for instance confusing the map Datum. Human mistake is the error when the user is assigning a position to the data, not because of the map reference, but by user misplacement. Correction of human mistake could be done by user validation and correction or by topological testing (as is explained in section Logical consistency). There is an example of VGI project, Wikiloc, look at Figure 4.1, where digitalization of new data is done using reference maps, digitalization is constraint using an activated area where new nodes are able to be drawn.

Next experiment was done to test the accuracy in road digitalization by different users. The result were compared with "official data". In the experiment 8 people digitized a road that is 5 meters wide. The reference image is a PNOA[3] (National Plan of Aerial Orthoimagering) from the National Geographic Institute from Spain[4] (IGN) with a resolution of 0.5 meters. Fig. 1 shows the digitalization output from the experiment. The digitalization was compared with the road vector layer of BCN25 (this is the official cartographic map from IGN) for which the scale is 1:25000. The experiment attempts to reproduce the conditions of generating new data for OSM. The digitization has been done at a raster layer resolution that fits the screen resolution. The results are shown in Table 4.3.

[3] http://www.01.ign.es/PNOA/presentacion.html
[4] http://centrodedescargas.cnig.es/CentroDescargas/inicio.do

Figure 4.1: Digitalization of a track with distance restrictions in Wikiloc VGI web. Source: http://wikiloc.com.

The difference between the "official" data and the experiment data is not significant and as is shown in Figure 4.2. In some cases VGI will provide accuracy similar to the "official" source (Haklay, 2010). The data digitized by the 8 study participants have sufficient accuracy to be useful in a GPS navigation system

4.5 Atributte acuracy

This quality parameter is difficult to fit into the VGI quality assessment. Although it is possible to predefine a limited list of values to be properly assigned as an attribute to an object, this is not easily guaranteed. That issue that could be misunderstood as a disadvantage reflects the VGI adaptations to editor and user behaviour, e.g., an object with different set of attributes can fit to different usage purposes, cultural, leisure, scientific,...

In OSM a user is free to assign any value to an attribute. Nevertheless, a predefined list of attributes in some cases will be useful to reduce the variability of descriptions for a geospatial object. Once again a post-process validation may correct a mistake in a tag description, by assigning a new attribute value.

It is possible to improve attribute assignment. An assignment may be done after verifying the attribute with related geospatial information such as

Figure 4.2: Micro-experiment. The wide red line is the official data. The thinnest lines are the digitalized ones.

administrative boundaries, digital elevation models, or climatological data.

The next experiment examines whether or not OSM street names in Valencia have been tagged properly. To acquire the correctly tagged features, Cartociudad cartography is used (http://www.cartociudad.es). Cartociudad contains the names of streets being the official cartography from the IGN (National Geographic Institute from Spain). The steps followed in this experiment are as follows:

1. Convert OSM line features to equidistant points (5 meters between each point).

2. Perform a spatial join with Cartociduad.

3. Compare Cartociudad street names with OSM street names.

The result in Figure 4.3 shows the coincidence of OSM street tagging with official Cartocidudad tagging. The result shows a small percentage of coincidence in the names. The difference could be explained partially because of the language used to tag the street names (Catalan/Spanish). This result is only valid to check and certify the coincidence in street name annotations. The percentage in street name coincidence will vary and could be checked manually street by street in order to avoid language annotation difference and typography mistakes. If Cartociudad is used as a reference tool for tagging the streets accuracy will increase. Moreover, if there is any difference between official data and user knowledge it will be detected by a user when tagging the street's name.

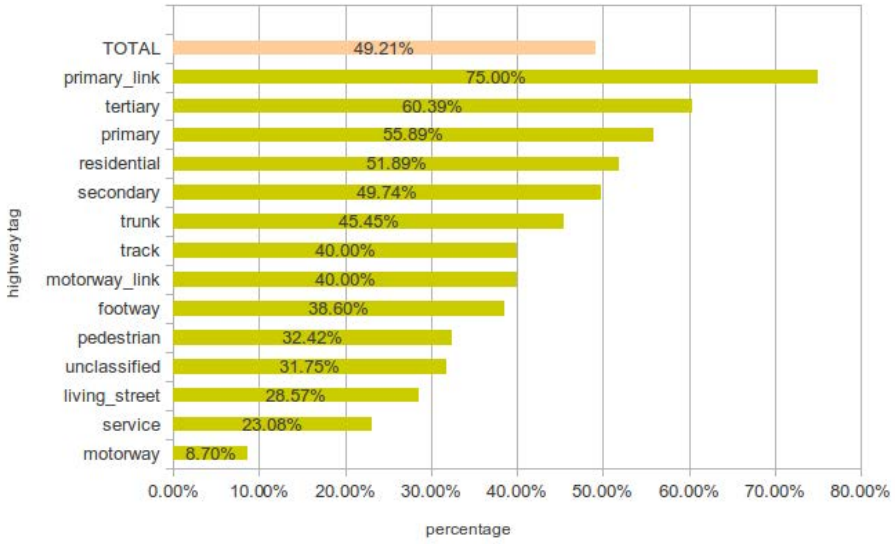


Figure 4.3: Coincidence in street tagging between OSM and Cartociudad.

VGI may be generated from different sources and one of those could be social networks such as Twitter1, where a user can publish georeferenced, short messages. Some authors propose to exploit VGI after it has been processed thus gathering raw VGI and consuming the results of analysis such as the density of similar messages (Schade et al., 2010). In the case of Twitter the attribute of a spatial object is the text message and its accuracy is the number of similar text messages published in the same location or near.

4.6 Logical consistency

In VGI logical consistency is rarely checked. When a user georeferences a point, line or polygons, indicating an object such as a bus stop, house, or street, it could be misplaced in regard to its logical position. Logical consistency must be checked and errors detected for misplaced objects (Goodchild, 2008). However, not all misplaced objects can be automatically detected. When a house is placed on the wrong block only the user volunteering information would be able to correct this mistake Miss-junction error is a way to check logical consistency. A VGI project may run a topological test when a geospatial object is created in order to automatically address logical consistency.

One example of a tool for verifying logical consistency is the OSM Inspector, look at Figure 4.4, which checks the geometry of the OSM data by identifying long segments and ways, duplicate nodes, self-intersecting nodes,

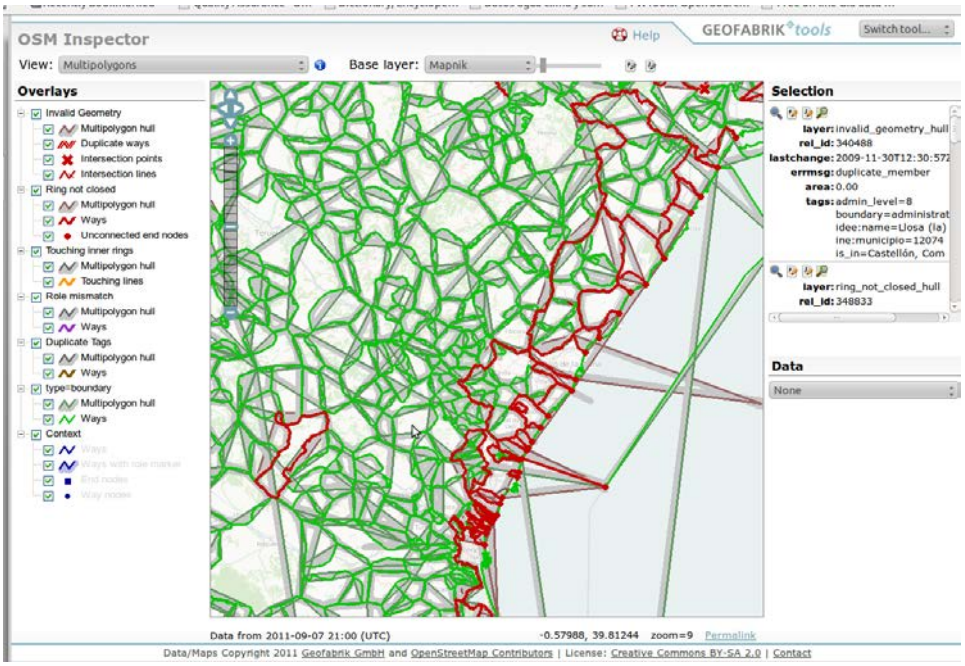


Figure 4.4: OSM Inspector Screenshot.Source: http://tools.geofabrik.de/osmi/.

and invalid geometries. OSM has more initiatives1 to increase quality that concern missing tags and duplicate nodes.

4.7 Completness

Completeness is one of the parameters that differences the most VGI projects and traditional GI (Haklay et al., 2010; Haklay, 2010). When is a VGI project complete? Projects such as OSM could be considered partially completed at the country level or even the continent level, making it quite useful for road navigation purposes. But how do we know when Open Street Map is completed at world-wide level? Moreover, the amount of geospatial information in OSM is growing and growing with bus stops, hospitals, and more information constantly being added. Completeness may be addressed by dividing the coverage area of a VGI project into subareas (quad-trees) depending on the variability number for objects represented in these areas (Maué and Schade, 2008). From a sub-area, variability representation is calculated using a ratio (see 4.1).

$$C = \frac{No}{Mo} \tag{4.1}$$

Where C is completeness, No is number of objects of the regionand Mo

is maximum number of objects of any region.

In this way the completeness is always related with the VGI project evolution and not dependent upon any external data source. Nevertheless, completeness also could be done by doing comparisons with official reference data or other reference data (Haklay et al., 2010; Zielstra and Zipf, 2010; Girres and Touya, 2010). The testing of the completeness ratio in 4.1 is left for future work.

Previous explanation of completeness is useful for a mono-thematic VGI project, a project which describes only one kind of feature. On the other hand, within an OSM project it is possible to find roads, house numbers, natural parks, bus stops, benches, hospitals, among other features. In OSM there is information built on top of previous information. In order to reference a bus stop it is necessary to georeference the street beforehand, this is similar for a drugstore and other features. Indeed, OSM project will be more complete as more layers of information are added.

4.8 Semantic accuracy

In a VGI project a user is able to freely annotate an object attribute. However, the user's interpretation could be different from the creator's interpretation. Even when restricting the user's freedom, the semantic accuracy will be diverse, because it depends on the user's training and knowledge. One approach to the semantic accuracy is the use of folksonomies. "A folksonomie is a collaboratively generated, open-ended labelling system, that enables users to categorize content" (Bishr and Kuhn, 2007). By using folksonomies it is possible to take advantage of multiple contributions. A user can tag a geospatial object he/she has created or other users may tag it and add meaning to the object. Hence the more tags an object has the more semantic accuracy it has to provide for users. Nevertheless, there is a possibility of vandalism or pollution by ill-intentioned users, this vandalism1 can be also detected and corrected by users. The objective is to take advantage of a large number of users within a VGI project to tag an object. This way an object receives a greater number of descriptions and interpretations.

Next experiment was design to extract the number of attributes annotations (tags) from an object. The Valencia OSM data only contains a core of recommended OSM's set of tags[5]. From this the amount of tags per feature has been extracted. Figure 4.5 shows a map of Valencia's city. Darker lines signify more tags for a feature. Hence, it is possible to visually distinguish which feature has more tags describing it.

[5] http://wiki.openstreetmap.org/wiki/Map_Features

Figure 4.5: OSM map of Valencia's old town. Number of tags per feature. Graduate color scale representation from yellow (lower values) to purple (higher values).

4.9 Temporal quality

This describes the temporal resolution, e.g., how fast a spatial object is captured and represented. Temporal quality depends on how fast the real world is changing and how fast those changes are translated to digital cartography. There are things which have not changed during years such as a cathedral or a church. Some objects change frequently in a short time such as new building areas. For instance, during the construction boom in Spain, the appearances of cities were changing so fast that official institutions were unable to keep up with the changes. In only a minute a forest fire can change ground vegetation from a mature forest to a desolated burned area. Burnt forest areas later change slowly back to a tree-filled forest. VGI projects are constantly being edited, published and reviewed, whereas "official" data revision depends on its cost (Goodchild, 2008) and most of the time it is slower than VGI. Indeed, the amount of changes of an object could be a measure of the quality. The more changes an object representation has the more close to real object could be.

Regarding temporal quality in VGI, as a bare minimum we know the object creation date. It is also possible to record the number of times an objec's attributes have been revised or modified. Therefore it is possible to calculate a ratio of graphical changes and attribute changes for a geospatial object, see 4.2.

$$Cr = \frac{Nu}{Lu - Cd} \tag{4.2}$$

Where Cr is change ratio, Nu is number of updates, Lu is last update, Cd is creation date.

Object changes described above in 4.2 are updated changes in an object feature. The validation and correction change descriptions are explained in Positional accuracy and Atributte acuracy sections . A more general approach to the temporal quality is the number of volunteer contributions in a time period 4.3. This contribution ratio is a measurement of VGI project activity. This measurement is relative to how fast change happens in the real world.

$$Ci = \frac{Nc}{Tp} \tag{4.3}$$

Where Ci is contribution index, Nc is number of contributions, Tp is time period.

OSM keeps a log with the changeset (log of every version of a spatial object). This change set is accessed via an API[6]. An application has been developed to collect the data from the OSM changeset history. The application has been developed to work with PostGIS for input data and to deliver

[6] http://wiki.openstreetmap.org/wiki/API_v0.6

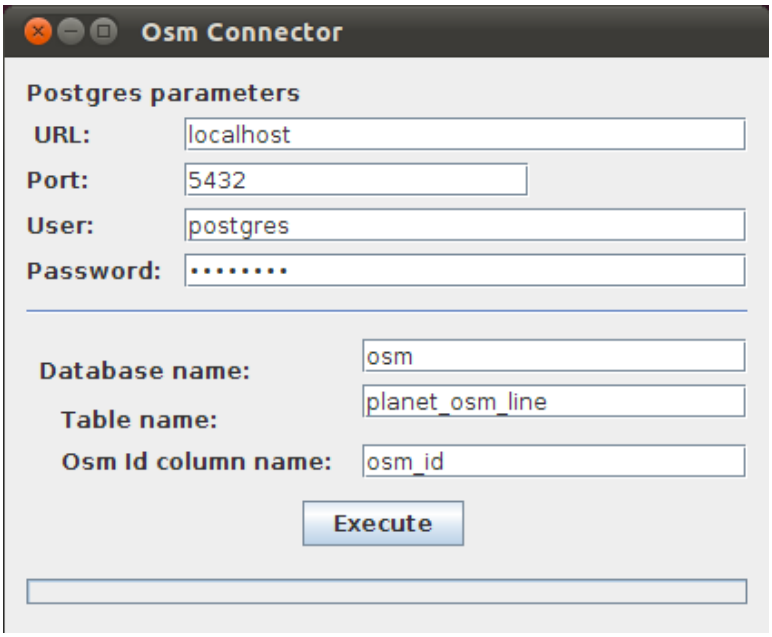


Figure 4.6: OSM Connector application screen-shoot.

output data. Figure 4.6 shows the application OSM Connector screen-shoot. The application extracts from OSM the changeset history, number of users that have contributed to describe an object, number of changes, days since last change and the change ratio describe in 4.3.

Figure 4.9 shows a spatial representation of these temporal measurements obtained with the application using graduated colors. Yellow indicates a low value whereas purple indicates high values. Of particular interest are the measurements for the temporal quality and change ratio, 7, and days from last change shown in 8. This approach is helpful for having an idea about how dynamic a VGI project is and which areas are most active and which remain unchanged.

There is a flaw with the change index tested with OSM data. A user can create an object and publish it then some minutes later visualize the object and realize that there is a mistake and quickly fix it. The change index in this case is very high not because a lot of users have contributed to it (e.g., by fixing mistakes or increasing the number of tags) but instead it is high due to the small time period between unique user updates.

The version number of the object and the number of unique users (see Figure 4.9) are complementary information for evaluating the meaning of the change ratio and days from last change. A fast analysis of the images shows how the city centre is a hotspot where the changes are faster and the surroundings have a smaller number of unique users contributing to updating the objects.

Visualization for Change Ratio for the city of Valencia.

Visualization for Days from last change for the city of Valencia.

Visualization of the version of the object.

Visualization of the unique users.

Figure 4.7: spatial representation of these temporal measurements obtained with the application using graduated colors, yellow denotes low values and purple high values.

4.10 Quality summary

In this chapter GI quality has been discussed with regard to VGI features. VGI projects have peculiarities that require the adaptation of the geospatial data quality parameters and assessment methodologies. Quality in VGI project is attained depending upon the volunteered contributions to the project. We have demonstrated that it is possible to measure contribution quality and constrain contributions using and adapting GI quality parameters. The lineage quality could be extracted by using log of spatial object changes history. Additionally, the freedom in contributing VGI could somehow be constrained by requiring the annotation of a reference source for information. Lineage information should be retrieved automatically for beginners. Expert users should be able to manually modify this information. It is possible to improve a volunteer's contributions by providing them with tools to assist them in the data generation process such as digitization within a predefined zoom level or scale. The method of digitization is useful for providing information about accuracy. Nevertheless avoiding human error during digitization is a difficult task. Providing tools that constrain digitization under some limitations will be the best approach to prevent a human mistake. Uploaded information using a GPS device could provide accuracy information but when that is not possible the accuracy should be within a standard limit. Logical consistency for topology could be checked automatically. When a geospatial object is placed where there is no reference cartography the VGI community should verify the object in order to enhance the quality. One of the strengths of VGI is the freedom in tagging an object. The information in a VGI project is created in a collaborative way. In fact it allows semantic accuracy to be improved by using folksonomies and adding tags by other users who are not the object creator. Temporal quality requires creating logs at least with a time stamp of the changes and reference source. This is a parameter that gives an index of VGI project dynamism and how frequently it is updated. Depending on the VGI project topic this index means a faster or slower change. Some adjustment should be done to ensure the appropriateness of this quality parameter, so as to address issues such as repetitive uploading by the same user within a short time.

Chapter 5

Descriptive Credibility

Chapter 4 describes challenges and possibilities to fit quality elements of GI to VGI idiosyncrasy. Quality doesn't take care of editor background or GI user's feedback. Credibility or trust is a key concepts for VGI data (Flanagin and Metzger, 2008; Bishr and Mantelas, 2008) . *"Credibility of a source or message is a receiver based judgement which involve judgement of information quality and accuracy"* (Metzger, 2007). Credibility involves perception of source's trustworthiness, expertise, and attractiveness (Freeman and Spyridakis, 2004). Trust concept is related with credibility defined by Mcknight and Chervany (1996) as "the extent to which one party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible" (Mcknight and Chervany, 1996). VGI data have an advantage comparing with other online information or commercial transactions, it is the possibility to check the correspondence of the georeferenced information with its real situation on earth or its geographical context.

Somehow credibility evaluation of VGI means to evaluate user trustworthiness, expertise, and attractiveness of GI regarding final use purpose. It is here and not in quality where fitness for use has wither meaning. The goal is to have a volunteer contribution credibility which could be transfer to his/her contributions. This transfer of credibility depends on the overall contributions done by the user. How credibility will be evaluated? Parameters to evaluate volunteer credibility or reputation are:

- User background profile or thematic knowledge of its contribution.
- Proximity to the event reported.
- Quantity of information reported
- Information source.
- Validation and correction of GI by other users.

Part of the credibility depends on the user. If we take as an example Twitter, user credibility depends on if the user has a verified account, number of the followers or the number of favorites tweets. Those are credibility factors that doesn't depend on the location of its contribution, depends on its reputation, credibility.

5.1 Volunteer background profile

Users difference is related to his/her background is his/her expertise. There are to different types of contributors; an expert or scientific ones which have a professional relationship between geospatial information contributed and an ordinary contributor, which have a proximity or relationship with the geospatial information contributed. Scientific contributor is mainly based on credibility-as-accuracy. Meanwhile, ordinary citizen is based on credibility-as a perception (Flanagin and Metzger, 2008). The expert or scientific and citizen contributors have the same start level ranking. Nevertheless, the expert has an academic or professional profile which accredit his expertise. Some projects do not validate such expertise A geospatial information created by both type of users is born without being validated.

5.2 Amount of geospatial information reported

A user with more geospatial information reported maybe is more involved with the VGI project and with the time will be more trained. Indeed, the credibility could be positive if we think that a volunteer is reporting a lot of information, maybe is because he/she have a better knowledge. Nevertheless, this could be negative if this data is reported in a short period where a possibility of human mistake could be high. A ratio like the equation 5.1 allows to compare users giving a ranking index to each one. An approach is the use percentiles from results of equation 5.1 to extract from results those highest values taht could be suspicious. As an example a result close 1 means a very participative one but a result close to 0 means low participation level. A percentile rank could available whenever is needed for VGI project or analysis using this equation,

$$Ra = \frac{\sum_{i=1}^{n} Ge}{P} \tag{5.1}$$

Where R is the event contribution ratios, Ge i a geospatial object uploaded, and P is the percentile to calculate the ratio.

5.3 Information source and complementary data

GI credibility will increase if there is a contrasted source used a a base. This is more important for the GI attributes than for georeferenced object. Answering the question, what source (maps, articles, books,..) has been consulted to create an object? Increase object credibility. Nevertheless, self-source content could be penalized if this criteria is measured.

A better approach is a fulfilment evaluation of the geospatial object. An object can be added to a geospatial database just by placing it on a map. Nevertheless, this object will increase its value if it is described by complementary data. Moreover, a user will have a better reputation when its geospatial information have a better description. For instance if an upload just contribute with a location of a bus station with no other complementary data the value of this data is minor. In the other hand, if is added the buses that have a stop in this station and its timetable its value will increase. The difference in both cases reflects user knowledge and involvement in the goodness of the geospatial information. The complementary data associated to an object could be predefined. Therefore, user reputation could be measured according to complementary data completeness, as is show in

$$Uc = \frac{\sum_{i=1}^{n} Fe}{F} \tag{5.2}$$

Where Uc is user complementary data, Fe filled fill and F is number of fields.

5.4 Validation and correction of geospatial information by other users. Volunteers reputation

The GI reported has an initial value. One part of its value comes from its contributor credibility and other one coming from the content and the way the object has been created. Nevertheless ,in some VGI projects a user can validate and correct the GI other users has contributed to. A positive validation increase user and object credibility . A correction or negative validation is a reputation devaluation

To develop a reputation system according to Resnick et al. (2000) is required at least three properties:

- Long-lived entities that inspire an expectation of future interaction.
- Capture and distribution of feedback about current interactions (such information must be visible in the future).
- A use of feedback to guide trust decisions.

Long-Lived depends on the impossibility to change contributions record history (Jøsang et al., 2007). If a volunteer wants to change its reputation grade or rank it must be done with new actions. Of course it is difficult to avoid volunteer double registration, identity change, after a mistake in a contribution. Nevertheless, to avoid a profile change it should be possible to evaluate positively its own corrections. The second properties requires users participation in the reputation evaluation, allowing feedbacks about what they have found in the GI checked. The simplest approach is correction or validation, but could be improved with other ways of feedback. The third property will work if the previous two are working. This property's means that the user of a VGI data should know a prior indicator of the object credibility before check it or use it.

An algorithm for specific validation and trust system between users (Bishr and Janowicz, 2010), could be a solution for trustworthiness assessment. Nevertheless, a user which will use the information to get accurate data for a research will check and asses geospatial information depending credibility and reliability more rigorously than another user that only want to have look at it (M. J. Metzger 2007).

Jøsang et al. (2007), describe methods to get reputation within a commercial transaction environment. VGI is not a commercial transaction environment, but some methodologies described in the article could be applied. Volunteers assignment reputation by a given party is done indirectly across his/her GI. Taking into account that VGI projects are mostly collaboratives (e.g. OSM), a geospatial object or evaluated area could contain more than one contributor, therefore reputation assignment is done globally to all object's contributors .

Volunteers reputations assignment should be carried on by validating geospatial objects, avoiding direct assignment of reputation by users, see 5.3. In other words, if a contribution is validated by other users, his/her reputation will increase. Meanwhile, if geospatial object is corrected by other user its volunteer reputation will decrease. Nevertheless, consecutive corrections of a geospatial object will increase object's credibility. In other words a user reputation will decrease with a correction but object credibility will always increase with users corrections because the object is being improved. Next equation could be applied

$$Uv = \frac{\sum_{i=1}^{n} v - \sum_{i=1}^{n} c}{\sum_{i=1}^{n} Nv} \tag{5.3}$$

Where Uv is user validation index, v is positive validations, c number of corrections, Nv is number of validations and corrections.

Chapter 6

Automatic classification model for Volunteered Geographic Information, applied to birds observer credibility. Spatial Credibility

The internet has become a place to store personal data or to contribute to collaborative projects such as Wikipedia built by users. Users from Web 2.0 (Oreilly, 2007) can contribute freely to a project. However, this contribution may be good or bad, there is a lack of credibility and quality assurance (Flanagin and Metzger, 2008). Sometimes it is checked and corrected by the community whereas at other times the volume of contributions make it an arduous task or impossible. In a geographical context, volunteered geographic information (VGI) (Goodchild, 2007) has become an important source of geospatial information. There have been several approaches to check and measure data quality. The first approach has been to check quality of VGI data using methodologies designed for geographical information done by professionals (Girres and Touya, 2010; Haklay et al., 2010). In other words, where professionals are the only ones contributing to the final product, it has a guaranteed or certified level of quality. This approach is not useful for VGI, because spatial data is created from a variety of users or purpose (Coleman et al., 2009), and where a user is not required to be certified to start his contribution. This first approach measures lineage,

Chapter submitted to Ecological Modelling http://www.journals.elsevier.com/ecological-modelling

positional accuracy, attribute accuracy, logical consistency, completeness, semantic accuracy, temporal quality, metaquality, variation in quality and usage purpose.

For example, OSM is a project which has become the Wikipedia for maps and has been compared with maps done by official geographical agencies from different countries done by professionals. There are several papers about Openstreetmap such as (Haklay and Weber, 2008; Fan et al., 2014; Girres and Touya, 2010; Haklay et al., 2010). Nevertheless, this approach requires something to compare to. OSM is made by volunteers but to get some results of data quality, a comparative map produced by a goverment o company is needed and there is spatial data available to be compared with.

The objective of this chapter is to create a model based on VGI, capable of filtering or classifying future contributions based on previous data coming from the same source at different or same locations.

A recent approach has been a classification of VGI based on a previous check list of conditions and the use of external data to identify useful information among VGI contributions (Albuquerque et al., 2015). Another approach proposed by Ostermann and Spinsanti (2011) also retrieved geographic information from social media. Both approaches require a classification of the contributions based on semantic criteria. Our approach avoids this classification because data source comes from ebird observation initiative simplifying the context. Indeed there is no need of semantic classification, so we point the efforts to spatial relations and filtering based on a statistical approach.

MCMC methods are powerful computational tools for Bayesian Inference which are normally used. Nevertheless, a new Bayesian approach, Integrated Nested Laplace Approximations (INLA) allows the inclusion of prior information coming from an expert point of view or the knowledge of the context Blangiardo et al. (2013). The advantage of using INLA is that there is no need to do a semantic filtering of the VGI data or ground truth. Indeed, this is possible because INLA approach use Stochastic Partial Differential Equation (SPDE). SPDE creates a mesh which allows a spatial relation between the different points.

In this chapter section 6.1 explains the data used to create the models, where it comes from and which data manipulation has been done to extract the input for the model. Section 6.2 explains which statistical approach has been chosen to model the data. Section 6.3 shows the results from the model output. In section 6.4 the results are analyzed and discussed according to the objectives of the chapter. Finally, section 6.5 explains the conclusions of this chapter.

6.1 Data settings

The data of this paper comes from ebird, http://ebird.org, coordinated by Cornell Laboratory of Ornithology and National Audubon Society. This is a Citizen Science project to collect in real-time ebird data and make it accessible online using a check-list program. We have used the data from Spain, only those within the Iberian Peninsula, for the year 2013 (Cornell Lab of Ornithology, 2013), look at Figure 6.1. The total number of bird observations used in this chapter are 44123, it is a large number of volunteered contributions to this project, taking into account it is only for the Spanish peninsula. Within ebird data columns for each feature we have selected:

1. Global unique identifier (a unique alphanumeric code assigned to each record)
2. Scientific identifier (a unique identifier for each scientific species within 2013 ebird data)
3. Observation count (the count of the individuals,(birds) at the time of observation if there is no count 0 is the indicator of presence)
4. Observer's identifiers (the individual identifier for the observer)
5. Number of observers (number of persons that are observing the birds)
6. Project identifier (there are several portals to report data, this field points to the portal from where the data has been reported)
7. Trip comment (binomial variable 1 for comment, 0 for no comment)
8. Corine Land Cover (2006 data[1]) is the variable added to the original ebird data.
9. Order identifier (an identifier for each unique species' order reported during 2013)
10. Number of observers (total number of observers reported participating in the sampling event)

Within this list these covariates have been used in the model, order identifier (id_order), observartion count (observatio) and number of observers (MUMBER_OBS). During birds observation the number of observers is important because other observers are somehow validating the birds identification. Bird species could be an important covariate because some species are easier to identify or observe than others.

[1] http://sia.eionet.europa.eu/CLC2006/

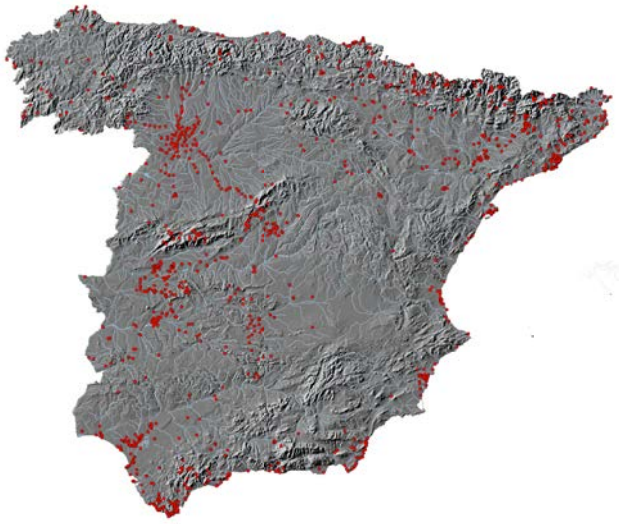


Figure 6.1: Location of the ebird data used in this chapter. Some points are masked they are in close proximity to each other.

Corine land cover data is a land use layer. Bird distribution depends on the landscape; forest, scrubland, pasture, wetlands and so on. Therefore, an observation of a bird species outside of its land use should be reviewed to check its credibility.

The response variable is binomial and its value is zero when number of observers or trip comment or birds observed is zero, in other cases it is one. The global unique identifier, project identifier and observer identifier are kept as a link to the original data.

6.2 Methodology

There are mainly two ways to get data about bird species and their migratory patterns directly from fields. One is to have a well trained staff capable of developing campaigns with different techniques for bird census, with this option it is impossible to cover a country, so you go to strategic points to census or observe as many birds as you can. The other one is to get volunteers capable enough to observe birds and record their observations, with this options it is possible to cover more territory and get fortuitous observation records, the only handicap is the credibility and quality of these observations. The data gotten from ebird are a spatial point pattern, in other words, birds observations are done in concrete places with their coordinates,

it is not a measurement of a continuous variable. In our case we have different covariates included in the point process modelling as was explained in section 6.1.

A new Bayesian approach allows the inclusion of prior information coming from an expert point of view or the knowledge of the context (Blangiardo et al., 2013), for VGI data this approach allows definition of some rules for data classification according to the source or to the final use of this data or both. Within the Bayesian approach Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009; Lindgren and Rue, 2015) has been used to model the ebird data. INLA is a different approach to work to Hierarchical Bayesian Models which traditionally have relied on Markov Chain Monte Carlo (MCMC). Whereas MCMC is based on simulation techniques which are computational consuming, INLA uses approximation for inference, avoiding large computation demand to get a result (Rue and Martino, 2007).

In this chapter point-georeference data are being used, which are being processed using the Stochastic Partial Differential Equation (SPDE) approach proposed by Lindgren et al. (2011). We used the proposal consisting in the representation of the Gaussian Field (GF) with Matérn covariance function as a Gaussian Markov Random Field (GMRF). We start from a GF with a Matérn structure this allows using SPDE approximation to transform the initial GF to a GMRF. GMRF has a simplest internal structure more simple, which is better for computation. In this approach the GMRF will be used through the SPDE (Cameletti et al., 2013; Lindgren et al., 2011). GMRF can model the spatial dependencies of observed data from a location in a space represented by regular grid, geographic region or spatial units.

We have used Matérn correlation function. This correlation function depends on a scale parameter $\kappa > 0$ and a smoothness parameter $\nu > 0$. Considering two locations s_i and s_j , the stationary and isotropic Matérn correlation function is :

$$Cor_M(X(s_i), X(s_j)) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\kappa||(s_i - s_j||)^\nu K_\nu(\kappa||(s_i - s_j||) \tag{6.1}$$

Where $||\cdot||$ denotes the euclidean distance and k is the modified Bessel function of the second order. The Matérn covariance function is $\sigma Cor(X(s_s), X(s_{sj})$ where σ_x is the marginal covariance of the process (Krainski et al., 2015).

Therefore GMRF could be useful to model the spatial relations of the observed ebird data, for further information see Rue and Held (2005). The SPDE approach builds a triangulation or a mesh starting from the location points of the data frame being the vertices of these triangles (birds observation points), the data location points and extending it to the areas where

there are no data. With this added triangulation it will be possible to get a spatial prediction in the study region.

As was defined by Blangiardo et al. (2013) the mean for the i-th unit is modelled by means of an additive linear predictor, defined on a suitable scale 6.2.

$$\eta_i = \alpha + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{m=1}^{M} f_l(z_{l1}) \tag{6.2}$$

Here α is a scalar representing the intercept, the coefficient $\beta = (\beta_{1,...m\beta_M})$ quantify the effect of some covariates $x = (x_1, ..., xM)$ on the response, and $f = f_1(\cdot), ..., f_L(\cdot)$ is a collection of functions defining a set of covariates $z = (z_1, ..., <_L)$. This formulation can be used in spatial and spatio-temporal models (Rue et al., 2009).

In order to model the data, Response variable, has been used as a binomial one, this variable is zero where number of observers or trip comment or birds observers is zero, in this case the value for response is zero, or else is one. There are several articles dealing with binary data to model with INLA such as Roos and Held (2011); Grilli et al. (2014)

We use a Bernouilli model, with i_1 (Krainski et al., 2015). Nevertheless, we have used a binomial as a generalization of the Bernouilli. We define the linear predictor to first component by:

$$logit(p_{i=\alpha_i - x_i}) \tag{6.3}$$

Where α_i is the intercept and x_i is a random effect modelled by a Gaussian field through the SPDE approach.

Finally the automatic classification is the difference between the VGI response and the model response. This result is used to do the automatic classification of the VGI input data.

6.3 Results

In order to build a model using INLA within a SPDE approach a mesh is needed, covering the study area and containing the observation points. Figure 6.2 shows the mesh build for the model. The triangles from where there is ebird data have a smaller triangulation, whereas in the areas with no data, the triangulation is bigger. This approach allows the model to have a mesh with different triangulation density, reducing the amount of nodes. Nevertheless, it is important to point out that boundary conditions is an open research topic (Lindgren and Rue, 2015).

Results from Table 6.1 show the summary for the estimated regression parameters. In Table 6.1 results for models with only one covariate or no covariates show how all the covariates are significant, because quantiles have all

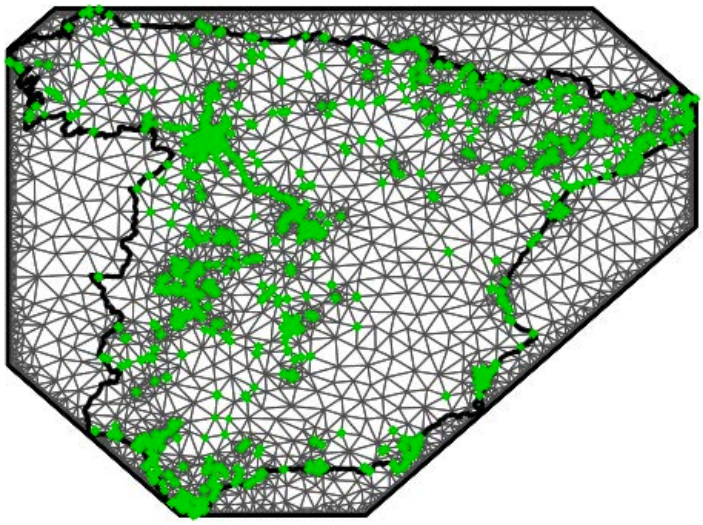


Figure 6.2: Mesh build to model ebird data.

the same sign, this means that where a covariate increases/decreases its value also increases/decreases the result along the model. For the model with all the covariates on the other hand, id_order and number of observers (MUMBER_OBS) and observatio are significant whereas corine and ID_ORDER has a different behaviour, for some quantiles it has a positive influence and for others it has a negative influences, its behaviour not being significant.

Table 6.2 results for the hyper-parameters show that, attending to the sign, no covariates, “corine”, “observatio” and “all” covariates are significant, whereas id_order and number_obs are not significant.

Table 6.1: Summary for the estimated regression parameters, first row only considers the spatial effect.

covariate	**mean**	**sd**	**0.25 quant**	**0.5 quant**	**0.975 quant**	**mode**
no cov.	7.34789	4.67523e-05	7.34782	7.34788	7.34802	7.34790
Corine	7.9569899	0.0008735	7.9552345	7.9570063	7.9586679	7.9570546
Observatio	6.5085188	0.0023356	6.5039325	6.5085187	6.5131103	6.5085188
ID_ORDER	7.36978718	0.00054201	7.36915514	7.369573	7.371148	7.369372
NUMBER_OBS	7.1703101	0.0011790	7.1680061	7.170305633	7.1726382	7.1702923
All covariates						
Corine	0.008715	0.01269592	-0.0161733	8.7014e-03	0.0336558	8.6751e-03
ID_ORDER	0.0009153	0.00260981	-0.0060394	-9.1541e-04	0.0042041	-9.1527e-04
observatio	0.0059191	0.00034394	0.0052453	5.9185e-03	0.0065952	5.9175e-03
NUMBER_OBS	-1.816447	0.0495932	-1.914934	-1.8160e	-1.720120	-1.8153

Table 6.2: Summary for the estimated regression hyper-parameters of the latent field. First row only considers the spatial effect.

covariate	**mean**	**sd**	**0.25 quant**	**0.5 quant**	**0.975 quant**	**mode**
no cov.	-9.812201	4.892298	-19.42459	-9.809938	-0.2211418	-9.804979
Corine	-0.024423	0.0033018	-0.030912	-0.024421	-0.01795	-0.024418
Observatio	0.004899	0.0002845	0.004341	0.004899	0.005458	0.004898
ID_ORDER	-9.694086	5.061680	-19.639380	-9.691711	0.228948	-9.686512
NUMBER_OBS	-8.98596	5.016367	-18.84722	-8.98194	0.8440142	-8.97345
All covariates	5.6385065	0.0015355	5.6354912	5.6385064	5.6415251	5.6385065

Table 6.3: Validation results for the model with all the covariates compared with other models.

Models	**RMSE**	**Pearson correlation coeficient**
only spatial, no cov.	0.2401988	0.7749376
Corine	0.2509217	0.7514381
Observatio	0.2234917	0.8085386
ID_ORDER	0.2405872	0.7741071
NUMBER OBS	0.2131303	0.8275727
All covarialbes	0.1909584	0.8641934

Other descriptive results for ebird data modelling taking into account all the covariates are shown in Figure Table 6.3. All of them with their corresponding Gaussian structure. Their values correspond with the priors of the model by default. Table 6.3 shows the validation results for all the models tested in this chapter. This validation is important to choose the better model and also to validate model's suitability for data input. In Table 6.3 the best results are for the model with all the covariates with higher Pearson correlation coefficient, 0.864193, and with a lower Root minimum square error (RMSE), 0.1909584. Therefore, the model chosen to implement VGI data automatic classification is all covariates model including the spatial relationship.

Figure 6.3 shows only the results for the model with all the covariates for a distribution of the response and latent field along the study area. Figure 6.3 shows the latent field standard deviation as a negative of the response standard deviation. Response standard deviation for the All covariates model has a good approach to the data input compared with Figure 6.2. Response mean is close to one where there are more data input, nevertheless there are areas where there is data where the model gives a low mean value, for example central west of Spain.

Finally, after reviewing all the models, the best is that which takes into account all the covariates. Using this model an automatic classification map of the ebird data was built. This map was built based on the difference of ebird response value and model response value. The output value range goes from -1 to 1. Values close to 0 means that the difference between data and modelled response are minimum whereas values close to 1 or -1 means bigger differences. We have to point out that the model has a response value for all the study area but we only want to check those points with a VGI contribution. In order to improve user visualization of output result the data has been classified using traffic lights colours taking as a reference the values from Table 6.4 as an example. The final map is represented by Figures 6.5,

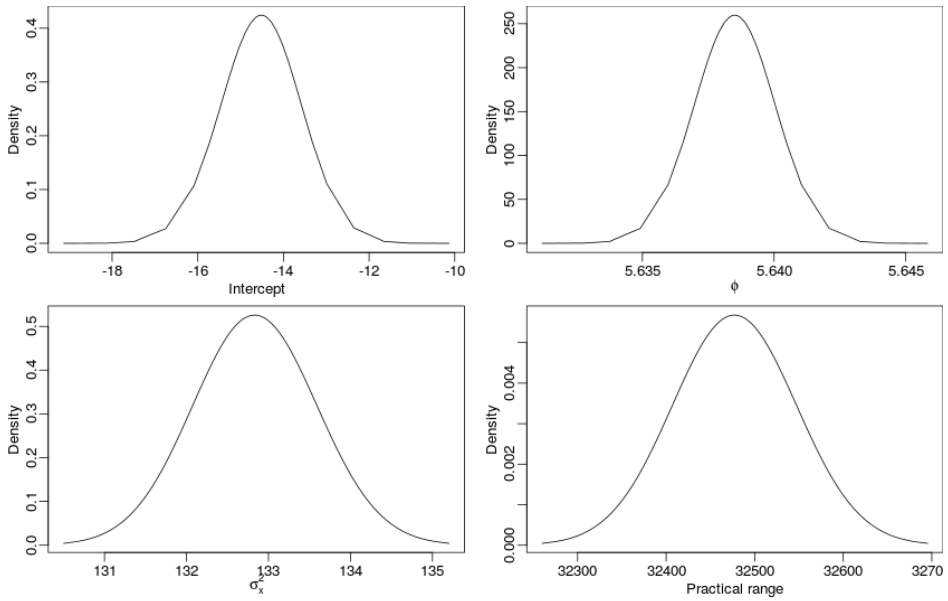


Figure 6.3: From left to right and up to bottom, figure shows marginals of the fix parameters, marginals for hyper-parameters, variance and nominal range.

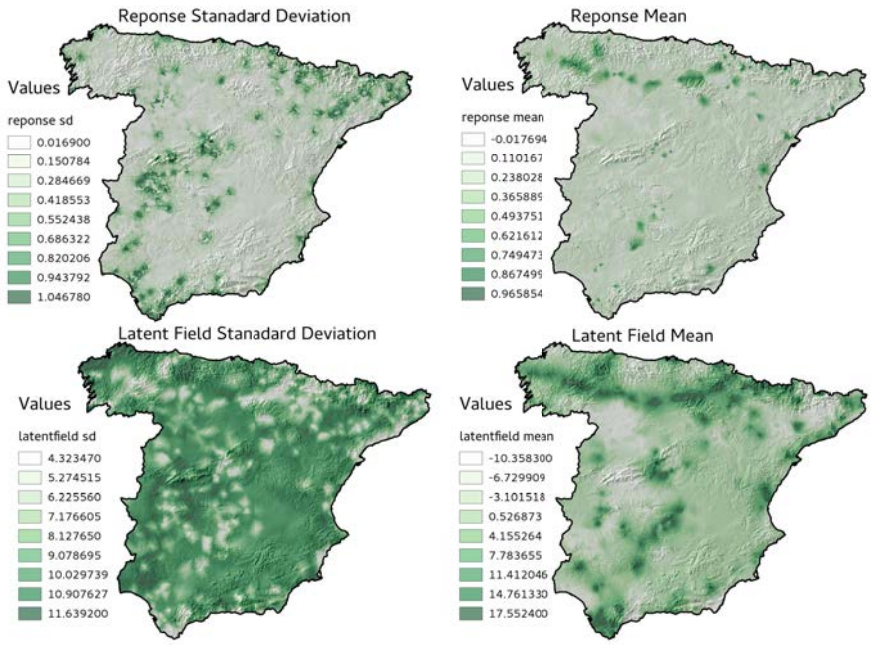


Figure 6.4: distribution of the response standard deviation, mean and Latent field standard deviation, mean.

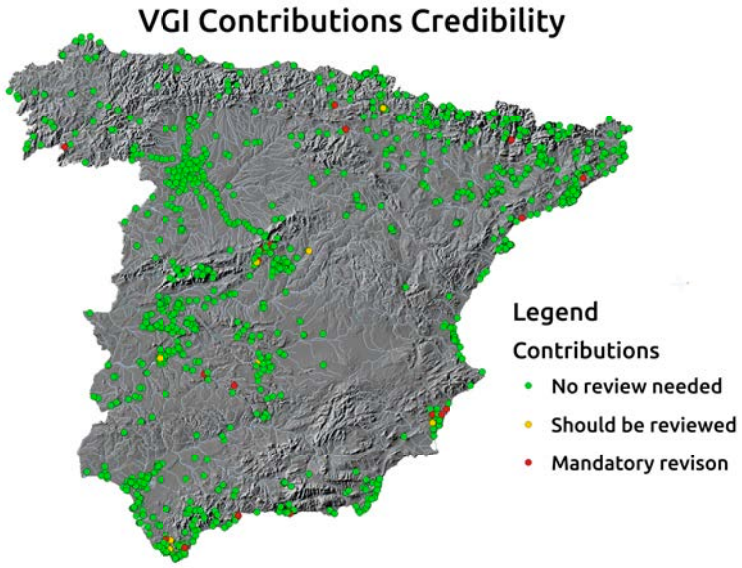


Figure 6.5: Automatic classification of ebird's volunteers contribution represented by a range of colors.

this map shows the same points locations as the map from Figure 6.1 but this time has a visual discrimination of which points should be reviewed or have less credibility.

6.4 Discussion

The approach for VGI credibility classification is based on a INLA GMRF and SPDE (Lindgren and Rue, 2015; Lindgren et al., 2011) approach. This approach allows us to relate a data point among the others based on its spatial relationship within the SPDE mesh (Cameletti et al., 2013). Therefore,

Table 6.4: VGI credibility example of color table classification according to VGI contribution credibility.

Color	Credibility classification range
red	-1 to -0.60
yellow	-0.60 to -0.45
green	-0.45 to 0.45
yellow	0.45 to 0.6
red	0.6 to 1

it is possible to build a model based only in the spatial relationship among points represented by the nodes from mesh vertex, indeed, it is possible to say that we are following Tobler's law, "everything is related to everything else, but near things are more related than distant things"(Tobler, 1970), as we take into account the neighbours in our model, close data have more impact.

In order to choose the right mesh several attempts have been made, and the criteria to choose it has been the triangulation density where there is data and the triangulation density where there is no data. At the moment there is no research done regarding mesh optimization.

Among all the covariates Corine Land Cover is the only one external to the ebird database and therefore is the one not done by VGI. The decision to include this data was because the land use is very important to bird observation, it is strange to observe a bird out of its habitat, especially if it is a wetland habitat bird. Nevertheless, the results for the model using only Corine land cover show that this covariate could be omitted from the model keeping the classification within VGI contributions.

ID_ORDER is not a covariate that has good results, look at Table 6.3, compared with the other ones, but we have used this covariate to check if there is any difficulty in birds recognition depending on bird specie. Finally, "observatio" and number of observers have become important covariates, see Tables 6.1, 6.2 and 6.3, even in a model alone or with all covariates model. Despite further research being needed, observatio and number of observers could be important because if there is more than one bird from the same species it is simple for the observer to compare the features among observed birds. In the same way number of observers has demonstrated to be an important covariate of the model, because it is more feasible to identify species correctly if there is more than one observer to avoid mistakes, in other words observers are reviewing the identification on the fly.

Finally a map with the ebird data credibility has been done, Figure 6.5. This map is based on the results of the model and its classification according to Table 6.4. The classification will be different depending on the objectives or fitness for use. For instance, if a reviewer has a limited time to do his work, he could classify the outputs of the model depending on the number of contributions to review, if the output is being published as is, a final user could have a reference of which data is more credible than others. Last case could be the use of this data for a scientific work, in which case the final user is forced to define a classification depending on the data quality parameters of the project or study. As we have explained there are different quality parameters depending on the usage of the data. A more complex classification could be done, for example Pôças et al. (2014) define quality indicators based on user requirements and expectations, in other words defining a fitness for use.

6.5 Conclusions

The chapter has shown the methodology approach to model VGI using INLA, GMRF and SPDE. This methodology as we have seen has some computational advantages. Regarding VGI data the principal advantage is the possibility of defining a response variable depending on credibility parameters defined by the modeller. Those credibility parameters could be different depending on data input or information usage. Moreover, credibility could be defined as a binomial, presence or absence of certain data simplifying the credibility criteria.

An advantage for this approach is the neighbour relationship considered in the model with the inclusion of the triangulated mesh. Indeed, the outputs of the model take into account the response values of its neighbours and the covariate values.

Regarding the covariates used in the model their number could be increased or in some cases eliminated to try to enhance the results and their consequences could be measured in the model outputs.

Further research is needed to choose the right mesh taking into account data input and is needed to define which parameters are better and in which context, to optimize the mesh.

Finally, the chapter has shown how it is possible to define a VGI automatic credibility filtering or validation based on predefined criteria using data coming from the same VGI project, a very important tool to manage spatio and spatio-temporal data in any kind of context, not only quality of bird observation.

Chapter 7

Tweet2r a package to capture from streaming, storing and describing large tweets data sets as spatio-temporal data

7.1 Introduction

Twitter is a microbloging service used by millions of people around the world. Twitter is used by people, companies and institutions as a tool to comment, support, criticizes, and publish information about what is happening to them or around them. Twitter also allow users to geotag their tweets. Scientific community has interest in analyses such phenomena, having access to tweets is a starting point.

StreamR package was developed by Barberá (2014), it allows the access to the twitter's stream API and parsing results from a JSON file in order to build a Data Frame. Despite its utility has some limitations to collect and work with large amount of tweets. One limitation is that a user can face on Twitter API streaming rate limits, second one is a limitation to manage large JSON files. Therefore, the workflow should change to work with large amount of tweets.

Tweet2r[1] introduces several changes in the workflow to enhance Twitter stream connection set up and results storage. In this package we have introduced the creation of two tables within a SQLite and PostGIS database and setting up parameters for an automating start and stop of tweets retrieving. Twitter results are exported to SQLite[2] and PostgresSQL[3] database. Post-

Chapter submitted to R-jorunal. https://journal.r-project.org/

[1] https://cran.r-project.org/web/packages/tweet2r/index.html

[2] https://www.sqlite.org/

[3] http://www.postgresql.org/

gresSQL has been chosen because it has a spatial database extension know as PostGIS[4], and is an open source database. SQLite has been chosen for its usage facility and integration with R. Therefore, postGIS and SQLite are the options chosen to store our data including geotag data. Afterwords it is possible to decide to work with geo and non geo data or export to GIS format (shapefiles, KML or GML).

Getting data from Twitter requires a definition of some keywords or a geographical bounding box, which results is a collection of tweets that match the search conditions. An advantage of this package is to make transparent the complexity of some parts of the procedure to the users, such as start and stop programming and postGIS and SQLite database connection to export tweets. Nevertheless, it is recommended to create your own Twitter application to get the secret keys in order to have access to Twitter streaming API. The package provide a default Twitter API access but its availability depends on the number of users connected.

The article is divided in three sections. Section 7.2 explains the requirements to work with this package. Section 7.3 explains how to work with tweet2r to retrieve tweets and to export them to SQLite, postGIS or GIS format. A comparison with another package is done section 7.4. Section 7.5 shows some examples about how to work with the data collected.

7.2 System requirements

Pacakge tweet2r works in connection with Twitter streaming API and SQLite or postGIS database. Working with SQLite database is the easiest option because the package can create it automatically, only needs a name for the database. In the other hand, PostGIS database requires access to a postGIS database. It is not mandatory to create a Twitter application it is possible to use the default one coming within the package. Nevertheless, for production usage it is recommended to create one in order to assure unique access to the API avoiding users rate limits . To create a Twitter application (App) go to `https://apps.twitter.com/` and click on the button "Create New App", you will fill a form and be agree with the developer application. With you new App you will get thise elements to set up the Twitter streaming connection:

- requestURL , `https://api.twitter.com/oauth/request_token`
- accessURL, `https://api.twitter.com/oauth/access_token`
- authURL, `https://api.twitter.com/oauth/authorize`
- consumerKey, which is personal for your App.

[4] `http://postgis.net/`

- consumerSecret, which is personal for you App.

Twitter API has some rate limits. The API only allows 15 requests per window. A window is 15 minutes interval. This limit means that only 15 requests are allowed every 15 minutes, in other words, only 15. Tweet2r requests per windows can be perform per Twitter App. This is enough for a single user. Streaming API is the one used by tweet2r, it has a limit of reconnections, that means if you keep open the connection for a long time, for instance a week or you are doing to much connections, streaming API connection could be limited by small number of minutes. Twitter do no specify the number of minutes. There is more informations about rate limits in Twitters' web for rate limits `https://dev.twitter.com/rest/public/rate-limiting`, and for streaming rate limits, `https://dev.twitter.com/streaming/overview/connecting`. As an example, the straming connection for the data retrieved in Section 7.3.4 hat got some problems with rate limits.

If a user choose a PostGIS database, the second step of the system requirements is to install Postgres an its postGIS extension. Postgres is a SQL databse and postGIS is a spatial SQL extension. Postgres is an open source software available to download at ,`http://www.postgresql.org/`, and is well documented. The version used for this article was 9.3. PostGIS is also well documented `https://trac.osgeo.org`. PostGIS version used in this article was 2.1. Once the database has been installed it is necessary to create a spatial database in which the tweets are going to be stored. This url `http://technobytz.com/install-postgis-postgresql-9-3-ubuntu.html` explains how to install and set up postgres 9.3 and postGIS 2.1 over a Ubuntu GNU-Linux distribution , also you can find more information in the posttGIS web page.Tweet2r requires:

- Active Twitter App, not mandatory but recommended.
- A SQLite or Postgres database with the postGIS extension

7.3 Package Workflow

The workflow is divided in three parts, look at Figure 7.1, Set up parameters for a connection using Twitter API, store JSON files and export tweets into SQLite or PostgresSQL/postGIS database. Nevertheless, the package has five functions. The two first steps of the workflow are done by tweet2r function. Some changes have been done in retrieving tweets, whereas JSON parsing functionality has been kept from streamR without modification. In the other hand, validate JSON files, export tweets into a SQLite or PostGIS database and export to GIS format is completely knew.

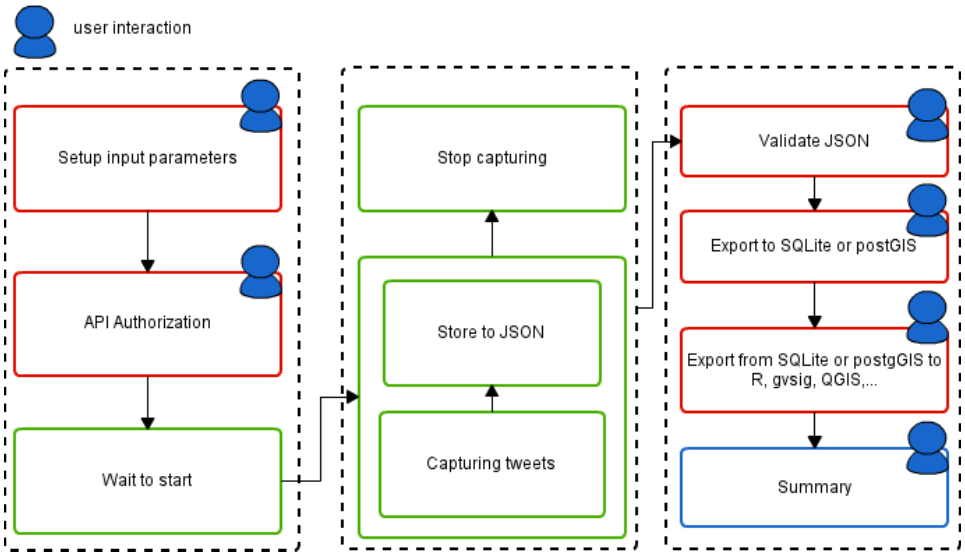


Figure 7.1: Tweet2r workflow.

7.3.1 Retrieve data from Twitter API, tweet2r function

Tweets are retrieved using the streaming API from Twitter https://dev.twitter.com/streaming/overview. The connection to Twitter service is done using the streamR package functionality. The change in this part has been a set up of start time and stop time to retrieve tweets. This is important when you need to get tweets for a certain event or time period, therefore there is no necessity of do it manually. Nevertheless, it is necessary to execute previously a connection with Twitter API service which keep waiting until the predefined start time comes. We have to point out that the stop time is not exactly the end time parameter, indeed this is an approximation and the last tweet will be retrieved when the number of tweets of the last JSON file arrives at the number of tweets set up previously. A user should set up the start and stop as follows:

```
#definition of the start time and end time
start<-"2015-09-11 9:45:00"
end<-"2015-09-11 23:59:59"
```

Tweets filtering could be done by bounding box or a list of comma separated key words (simple word or hashtag list). The list of key words ar filtered using the conditional 'OR', that means that all the tweets matching one of the key words listed will be captured. Twitter API doesn't make distinctions between a word with a hashtag or a word without hashtags, as an example #hashtag will be the same as "hashtag". Should be pointed out that the twitter API allows only a search by key word or by bounding box (bbox), never both at the same time. Therefore, tweet2r function checks

that only keyword search or bbox search has been set up, never both at the same time. A user set up API streaming filter as follow:

```
#set up a key word list
key=c("keyword1", "keyword2", "#hashtag")

#set up a bbox
bbox=c(-0.1644,39.8485,0.6916,40.0034)
```

Tweets are stored into JSON files, see Figure A.1 on section A. JSON file format is a lightweight interchange format human readable and easy to parse by machines. Despite there is a GeoJSON format this is not used by Twitter API, nevertheless coordinates are stored following the conventions described in GeoJSON . When the stream is open for a long time or has retrieved a thousand of messages or even millions the result is a large JSON file difficult to manage. Therefore, tweets are being captured in a loop and stored within several small JSON files. Indeed, if there is a problem with API restriction limit (see 7.2) the file will not be corrupted and previous and next files will be saved correctly.

It is possible to define the number of tweets per file. If number of tweets per file is not defined, 3000 will be set up as default. Another step to tweet2r function set up is to choose a file name prefix to store the tweets into a JSON file. A user defines the file name and the number of tweets per file as follow:

```
#definition of the file prefix
fileprefix="tweets"

#definition number of tweets per JSON file
number=3000
```

Finally a user can provide its own set up for Twitter API configuration according to the API definitions. This is done as follow:

```
#Configuration for twitter API connection
requestURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
consumerKey <- "your consumer key"
consumerSecret <- "your consumer secret"
```

It is possible to use the default Twitter API connection parameters, in this case there is no need to define a connection, the function takes the predefined one. Nevertheless, for large production work is recommended to create your own Twitter application as is explained in 7.2.

Once everything is set up correctly the funciton tweet2r can be run.

```
#running the function tweet2r, with keywords filter and your own
     Twitter application
tweet2r(start=start,end=end,ntweets=number,keywords=key,
    fileprefix = fileprefix,
    requestURL,accessURL,authURL,consumerKey,consumerSecret )

```

```
#running the function tweet2r, with keywords filter and by
    default Twitter Pi connection
tweet2r(start=start,end=end,ntweets=number,keywords=key,
    fileprefix = fileprefix)
```

Summarizing this section shows how to configure Twitter streaming API. This is the first step of the workflow. To set up tweet2r function is mandatory to define start and stop time, file prefix and keywords or bounding box, API configuration its not mandatory. The result is a collection of JSON files with all the information retrieved in each tweet (look at Figure A.1).

7.3.2 Validate JSON files

Regarding to Twitter API rate limits, sometimes the sistem sends some messages across the streaming API. This messages are stored in a single JSON file with the fileprefix name and its corresponding JSON file number (ex: twitter25.json). API messages are a problem for tweet2r functions working with JSON files, t2pgis and t2sqlite. Therefore, it is recommended to validate, delete JSON files containing API messages and rename the collections of JSON files. This is done by valjson function. This function delete JSON files containing API messages such as,"Exceeded connection limit for user" or"Easy there, Turbo. Too many requests recently. Enhance your calm". Afterwards next JSON file will be renamed using the correspondent file prefix and a consecutive file number.

7.3.3 Export JSON file to SQLite or postGIS, t2sqlite and t2pgis functions

Once the data is stored using JSON files it is necessary to parse the files, and export the data to SQLite or postGIS. Data retrieved could be to large. Therefore, it is better the usage of a SQL database to store and later on filter the data using SQL script, moreover when you are working with large amount of data. Within the data retrieved there are geotagged and non geotagged data. Apart from the geotagged tweets there are other geo data within a tweet but this package only stores coordinates of tweets geotagged by Twitter user. The code to export code to a SQlite or postGIS database is completely new. Whereas SQLite database is design to keep it in a single file and suitable for a single users. Postgres database allows multiple users connection using a web service to connect with it. Therefore, it could be useful to store large amount of tweets and update the database with new tweets.

We are using parse functionality from streamR and database connection functionality from RSQLite, RPostgreSQL and rgdal packages, (Wickham et al., 2014; Conway et al., 2013; Bivand et al., 2011). The steps to export tweets into database are, open database connection, parse JSON files, populate database, manipulate table structure and close connection. This is done

automatically to create a SQLite database with t2sqlite function where only is required a database name.

First step for t2pgis function is to create a Postgres database connection. This step is described in the documentation of RPostgreSQL. Second step is to parse JSON files and store it in the database. This is done by t2pgis function. The t2pgis function can manage with the collection of JSON files to parse it and export to postGIS.

It is necessary to point out that parsing and exporting tweets to other formats as shp has some problems regarding to character encoding. Indeed, this problem exporting tweets to shp from ***R*** has been on of the reasons to develop this package. Therefore, an important issue for JSON parsing is the character encoding, indeed to get a clean database it is mandatory to remove some special characters encoding. Nevertheless, the integrity of the tweet is keep safe and language special characters are not altered. Most useful information from tweets is exported to the database such as user id, text, retweets, location ,.... look at Figure A.1 and Figure A.2 in section A.3. The information stored in a SQLite or postGIS database is the next one:

- text, Tweet text,
- retweet_count, number of retweets,
- favourited, number of users that has favourite this tweet,
- truncated, if the tweet is truncated in several ones,
- id_str,identifier of the tweet,
- in_reply_to_screen_name, if there has been a reply to the tweet,
- source, source text,
- retweeted,if the tweet has been retweeted,
- created_at, creation time stamp,
- in_reply_to_status_id_str,if the tweet is contains the id of the original tweet
- in_reply_to_user_id_str if the tweet is contains the name of the original author,
- lang, text language,
- listed_count, the number of Twitter lists on which the author of a Tweet appears
- verified, if the account is a verified one,

- location, name of the place of the tweet,
- user_id_str, user identifier number,
- description, user description,
- geo_enabled, if the tweet is geoenabled,
- user_created_at, user creation date,
- statuses_count, total number of Tweets and Retweets a Twitter user has posted
- followers_count, user number of followers ,
- favourites_count, number of Tweets a user has favourited,
- protected, if the tweet is protected,
- user_url, user web page,
- name, user name name,
- time_zone, time zone of the tweet,
- user_lang, user defined language,
- utc_offset,user utc offset,
- friends_count, user number of friends or followers,
- screen_name,user screen name,
- country_code, code for user country,
- country, user country name,
- place_type, tweet place type,
- full_name, place full name,
- place_name, tweet place name,
- place_id, place id
- user place latitude,
- user place longitude,
- place_lat, place latitude,
- place_lat, place longitude,

- lat, tweet latitude,
- lon, tweet longitude,
- expanded_url expanded url,
- url,
- t_tans.

Bellow code example shows functionalities of the third step of the workflow. After tweet2r function stops from retrieving tweets, according to the stop time previously defined and after validate de JSON files, t2sqlite, t2pgis functions can be used. This functions creates and populate two tables (one with all the tweets and other with only the geotagged ones). Note that your working directory should be the same where the JSON files are stored. tweet2r and t2sqlite and t2pgis functions could be used separately in two different ***R*** sessions, therefore you can export JSON files whereas new ones are being created. Next example shows how to export tweets to postGIS:

```
1
2 #set up postgres conection
3 connection <- dbConnect(PostgreSQL(), host="database url",port
      =5432,
4 user="user name", password="password", dbname="tweets")
5
6 #name of the file prefix to  be parsed and export
7 fileprefix="tweets"
8
9 #run the function
10 t2pgis(fileprefix, connection)
```

Next example shows how to export tweets to SQLite:

```
1 #define the name of JSON file prefix to validate
2 fileprefix="tweets"
3 #export tweets to json an import to R as Data Frame
4 tweets<-t2sqlite(fileprefix, import=TRUE)
```

A JSON tweet has a time stamp and coordinates to locate the tweet. A timestamp is important to work with time models. SQLite and Postgres deals with several timestamp formats but not with the one stored in the JSON file. Therefore, t2pgis and t2sqlite deals with this issue. We have decided to create a SQLite table and postGIS view to store only the geotagged tweets. A postGIS view instead a table saves disk memory space, this means that each time data from a view is requested a virtual table is created and R can import the data to work with it. This procedure could be important when you are dealing whit large amount of tweets. Finally, the data is available within SQLite or postGIS database and we can use all SQLite and postGIS functionalities to manage, query and export this data.

7.3.4 Import geotweets an export to GIS format, t2gis

Having a function to export geotagged tweets to a GIS format allows to work with the data using several GIS software. This function exports geotweets from SQLite database created with t2sqlite function to a GIS format and import geotweets to ***R*** as a SpatialPointDataFrame. The example bellows this line shows how to export tweets to several GIS formats:

```
1
2 #database name
3 dbname="castnov"
4
5 #export as kml
6 export="kml"
7 geotweets<-t2gis(dbname,export)
8
9 #export as shp
10 export="shp"
11 geotweets<-t2gis(dbname,export)
12
13 #export as gml
14 export="gml"
15 geotweets<-t2gis(dbname,export)
```

7.3.5 Summary of the tweets

T2summary can perform a summary of the data retrieved with this package. As is point out in 7.3.3 deal with timestamp is quite difficult, t2summary deals with the complexity of date and time data to offer this information group by day and hour. The summary of a spatial data should include a map. This is done using ggmap functionalities, (Kahle and Wickham, 2013). Summary provides next information:

- Number of tweets (geo and non geo) as 'ntweets', number of tweets with geotag as 'ngeotweets', number of tweets whit no geotag as diftweets, percentage of geotweets as 'pergeotweets',
- a tweets location Map,
- a plot of the number of tweets distributed by hour (UTC +000),
- a plot of the number of tweets distributed by days of the week (UTC +000).

Next example show how to get this summary:

```
1 summary<-t2summary(tweets, geotweets)
2
3 #show output
4 summary
```

7.5

7.4 Differences tweet2r between twitteR package

There is a package on R that also deals with Twitter, twitteR. Whereas twitteR (Gentry, 2015) is focus in searching users and timelines, tweet2r is focused in the streaming API. TwitteR package has functions to deal with user public information such as relations with other users, explore Twitter trends, search twitters, get Twitter timelines among others, there is no functions that deals with Twitter streaming API. TwitteR deals with Twitter user information and search tweets focused on relevance, for instance, when a search is done the result is a timeline of tweets according to relevance and not completeness. Twitter recommends the use of the Streaming API to match for completeness `https://dev.twitter.com/rest/public/search`, an advantage of Tweet2R is to search for pass events and get the more relevant tweets.On the other hand Tweet2r keeps and open connection in order to get tweets according to completeness and not relevance using streaming API. Indeed, the advantage of tweet2r is to keep a long live request to get live tweets according to search parameters as it is explained in section 7.3.1.

7.5 Anaysis of data captured and stored by tweet2r package

In this section we are analyzing the data captured with tweet2r package. Once the data is in SQLite or postGIS database it can be imported to ***R*** or to GIS format.

We are going to use the data retrieved from a streaming connection. The streaming connection was working from 30/10/2015 at 11:00 a.m. (first tweet on Fri Oct 30 11:18:35 +0000 2015) to 02/10/2015 at 11:00 a.m. (timestamp of the last tweet was Mon Nov 02 11:07:13 +0000 2015). Whereas the connection was open, data from a bounding box (-0.1644, 39.8485, 0.6916, 40.0034) was retrieved, Corresponding to Castelló de la Plana (Spain) and surroundings. To test the code for this section look at section A.3, there is a test.R file with the code to test the package functionalities.

7.5.1 Data summary

We have seen in section 7.3.5 it is possible to get some description summary from the tweets retrieved with the package tweet2r. We are going to illustrate it with the data described at the beginning of this section. the procedure is simple; first runt the function t2summary to get description and plots.

```
summary<-t2summary(tweets, geotweets)
#show output
summary
```

The results from the code are the Figures 7.2,7.3 7.4. Table 7.1 show tweets description where: ntweets is the number of tweets, ngeotweets is the number of geotweets, diftweets is de difference between tweets and geotweets and pergeotweets is the percentage of geotweets . Figure 7.2 shows the maps of the tweets. As you can see in figure 7.2 there are points out of the bbox, most clear ones are from South America, this is because a user has its location within the defined bbox but is posting geotagged tweets from other places outside the bbox. This is something to have into account where you are analysing tweets and could useful to filter tweets to get only those from within the bbox.

Table 7.1: Description of the data retrieved.

ntweets	ngeotweets	diftweets	pergeotweetts
9730	410	9320	4.213772

Figure 7.2: Map of the tweets.

Figure 7.3 is the distribution of the tweets along the day group by hours. Tweet hour is stored in UTC +000 whereas the Castelló de la Plana UTC hour es +002. In order to interpret Figure 7.3 is necessary to sum 2 hours. As we can see the tweets are published mainly in the afternoon-night with an increase of publications during 13:00-15:00 Spanish launch time, ther is another increase at 18:00 end of the working time and a maximum number of tweets at 22:00. Figure 7.4 shows how the tweets are increasing during

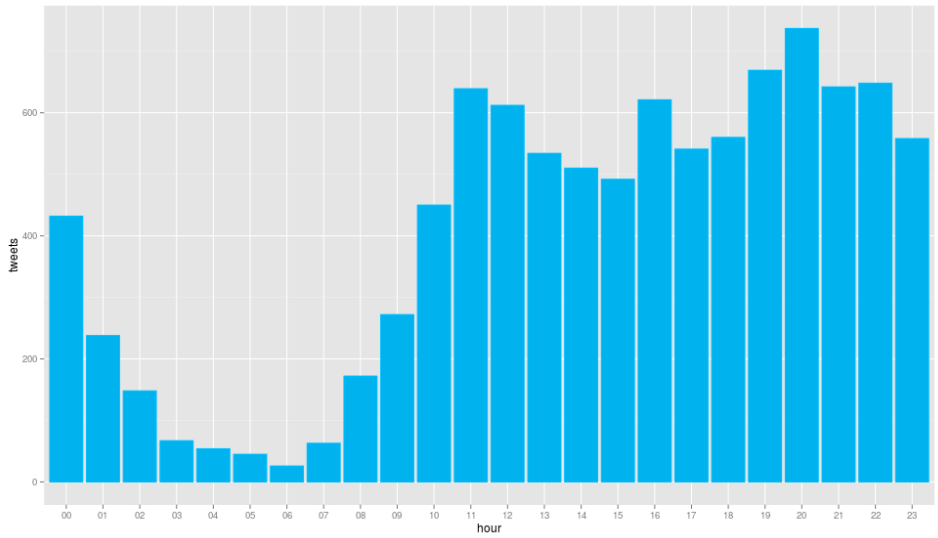


Figure 7.3: Distribution of the tweets per hours.

the weekend being the maximum on Sunday. This distribution of time is an overview about when the users are tweeting and the results shows that users mainly tweet during its free time.

7.5.2 Space analysis

Tweet2r main objective is to facilitate the extraction of geolocated tweets to make a posterior use of them. The total amount of tweets can provide a trending topic (most popular topic among community), a opinion of a tweeter community about a topic. With geotag tweets one can have the distribution of this geotwitter community.

Work with Twitter data is interesting but should be point out the geolocted tweets is a question of users self-selection. Users by default don't allow Twitter to use their location it is a matter of self-selection to geotag their tweets, as a result only a small proportion are geotagged tweets, about 0.85 % to 2 %, (Dredze et al., 2013; Hawelka et al., 2014) . Indeed, this should be the case when tweets are get by keyword, nevertheless, Table 7.1 shows a geotag percentage close to 4.2 %, in this case the tweets has been filtered by bounding box and not by keyword. Indeed, some results show how the geolocation is different depending on language, place of residence or even gender, (Sloan and Morgan, 2015). Some research has been done to increase the number of located tweets taking the information from tweet metadata or even from the text, nevertheless this location could be useful only where the scale of work is small enough to avoid misslocation, for instance country level or region level where tweet location could be inferred

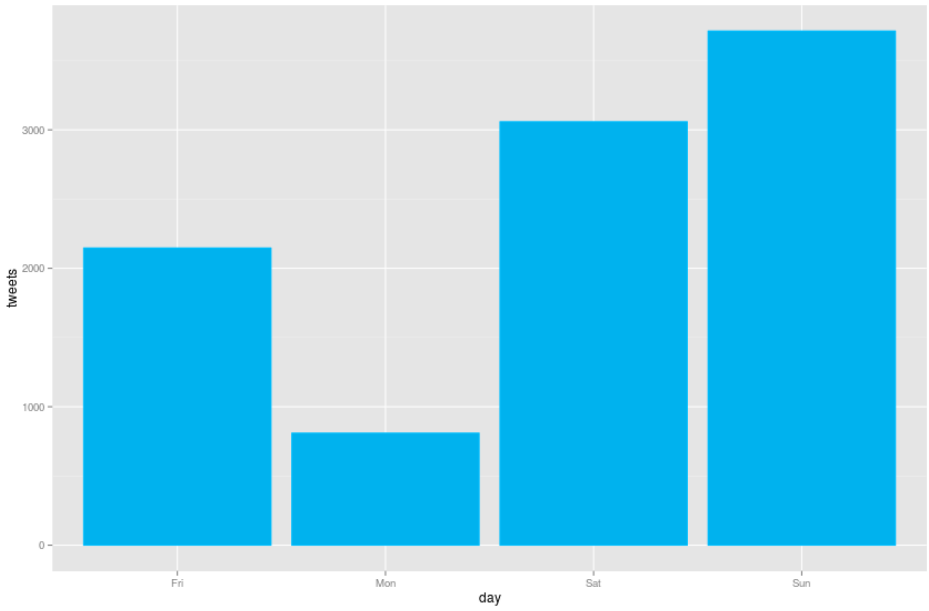


Figure 7.4: Distribution of the tweets per days.

from user metadata (location or place) ,(Schulz et al., 2013), other approach could be done base on temporal association with other geolocated tweets, (Ueda et al., 2015). Despite all this efforts to improve location tweet location enabled rely on users and those could represents around 4%.

Work with only the 4 % of the tweets could be a pour sample to get the trending of the twitter community, nevertheless it could be the only source of data to deal with spatial location, work done by Lenormand et al. (2014) found that geotag tweets covers 77% of highways from, Spain.

In order to show visual examples of space tweets distribution a subset from data has been chosen. From the data described in Table 7.1 a subset from a bounding box (-0.02265,39.9694,-0.8048,40.0032) has been selected. This bounding box is the city center of Castelló de la Plana (Spain). This subset correspond to a total number of 163 tweets. Figure 7.5 shows a map with the location of 10 clusters get with kmeans() function. The tweets are separated by neighbourhoods, in the city center there are two big clusters (cluster number 1 and 3) separated by one of the main avenues, Casalduch. There are another cluster at west of the map corresponding to a shopping center (cluster number 7).

Despite Figure 7.6 is an example about how the tweets could be divided into clusters, doesn't provide to much information about which parts of the city are more active within a geotwitter community. Nevertheless, an approach can be done by ggmap using stat_density2d function , this function compute a kernel estimation density base on the location of the points.

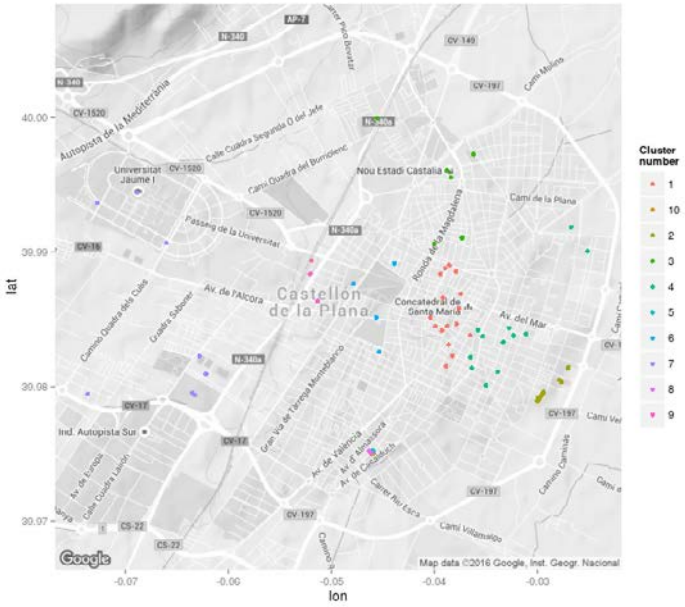


Figure 7.5: Distribution of the tweets. Colors shows the different clusters.

Figure 7.6 shows the result for stat_density2d, in the map is clearly visible three spots where tweets density lever is higher. This spots are north-east close to the train station (cluster number 9), east where is the city center and commercial area (cluster number 1 and 3) and a concentration in the south of the city in a residential area(cluster number 2).

7.6 Summary

This package has been created to facilitate data collection from Twitter, for instance the task of start and stop of Twitter streaming, this is very useful because it can be programmed before the start hour, therefore only tweets after the programmed start will be captured. The second advantage is the possibility to export the data to a SQLIte or postGIS data base and create two tables one for all the tweets and other one only for geotweets. SQLite and PostGIS are spatial relational databases, therefore it has SQL functions, PostGIS has also all the geoprocesses of an spatial database. Indeed, it simplifies filtering tasks and spatial tasks. This approach, using relational databases to store tweets and import to ***R*** functions or export functions, is very useful for a researcher because in the end the data is ready to use on ***R*** as an Spatial Data Frame or on a GIS software using the postGIS connection build on a GIS software or export from postGIS as a shp file. The function t2gis is in charge of the geotagged tweets, and export them to

Figure 7.6: Tweets density a darker color means higher density of tweets.

a GIS format. This function allows the user to work with this data using a GIS software.Tweet2r provides a summary function, t2summary which describes the data retrieved from Twitter.

Collect data from Twitter require a programming effort by a researcher, which tweet2r try to solve. Indeed, data collected from Twitter is very useful for researchers. Twitter data has been used in research fields such as crisis management (Albuquerque et al., 2015; Pultar et al., 2009), spatial crowd behaviour (Lee and Sumiya, 2010) or political ideology studies (Barbera, 2015) among others.

Chapter 8

Conclusions and future work

This Thesis address the challenge of delivering with VGI, citizen science and big data. Each one is different but somehow are related. All are related according to data creation. Data is created following a bottom-up approach, where ordinary citizen and non expert professionals are generating spatial data sometimes anarchically or non intentional, just uploading something with a geotag added to it. Indeed, decision makers face to a great amount of data coming from a variety of users with different professional profile, aptitude, motivation and so on. Therefore, it is necessary to classify, separate and analyse this amount of information as efficient as possible.

A starting point for this Thesis was forest fires historical database, collected by forest fires managers to be an historical record rather than to be use for spatial statistical analysis. Forest fires modelling traditionally has focus on physical fires models but that models doesn't take into account human behaviour or other variable not directly related with fires physics. Nowadays there are historical databases of forest fires. Here is proposed spatial point pattern process models with interactions, a well known methodology that acknowledges the non independence between wildfire events. As a result is possible to take advantage of a historical forest fires database to extract influence of other spatial features such as roads, landscape use, weather...

Physical models are clear and well studied, forest fires start where all the elements of the fire triangle are present, heat, fuel and oxygen. A forest fire spread faster on windy days and high slopes, that is what physical models says (Vélez, 2000). Results show that there is an interaction between wildfires, the expected number of wildfires will decrease close to another burned area, this could be explained regarding to fuel decrease. Results also show how forest fires has a human interaction where risk increase in coastal areas and close to roads, decreasing where population density is smaller. Forest arson are clearly related with human activity and spatial point pattern models could be very helpful for decision makers and risk disaster management.

The general form of the model used allows its application in wider areas as long as information about wildfire incidence are recorded in a historical database and covariates layers are available. Notice that is possible to increase the number of covariates by using crowd-sourcing information coming form participatory projects or social networks.

Spatial technologies from a digital world has become usual by the hand of experts, but before digital revolution, paper maps, made by cartographers or sketch maps where there. Participatory projects allow non experts to share his ground knowledge and his experience to professionals and experts.

Working from paper maps or easy to use applications, is fundamental to brake the digital divide. Facilitate contributions from bottom-up is not enough, a feedback should be ensured. An ant works for the colony but is receiving a place to live, meal and protection,as a feedback. To cooperate in a participatory project something should come in return, a reputation, a like, a retweet, an advice... A participatory model should be circular where information is being created bottom-up by participants, experts or not, but with knowledge in its neighbourhood. Later on experts could enhance these contributions or create a base data (Up-bottom) where participants another time can improve, check or complete it. This creates a community profit circle where experts get detailed information difficult to digitalize or to expensive to collect by themselves and non experts get and expanded information going beyond its expertise, capabilities or neighbourhood. Experts can get a more detailed data and participants can get a wider overview.

Projects build within a participatory perspective, VGI or citizen science, nowadays are able to involve participants to collect data. There are a lot of examples OSM, Geo-Wiki, Panoramio, or eBird. Once you get data, the challenge is to define quality and credibility parameters. Try to implement quality parameters for spatial data as is done within a professional GI project its difficult. Requires restrictions about volunteers participation depending on its profile or an ardours task to review contributions. In the first case the project potential contributors falls down, losing the opportunity to get more detailed or complete information in the best case the projects will go on but data will grow slowly. Second case requires to have an army of experts reviewing each contribution, which in some cases will be unaffordable.

Why don't use alternatives to standard quality measurements?. The point is to adapt quality measurement to the participatory projects (VGI citizen science). Each contribution could have a quality tag or could be check by the community. OSM have several tools to check quality or look for errors. Participatory projects base on Volunteers should provide tools to assist users in content creation in order to avoid simple mistakes or fill all required data. Quality concept has changed, there is no more standard quality for all the data, there is a distributed quality of each data subset or even single feature. At this point the quality somehow is not defined by the provider rather than the final user, it is fitness for use. In other words

quality depends on the final use of the data. Quality level will be different if the final use is for car navigation, outdoor tracking or scientific work.

Data quality depends in concepts such as, lineage, positional accuracy, Completeness, temporal quality... but, what about user credibility?. OSM project sometimes face with "trolls" , users that are destroying or creating fake streets or data, of course this is done on purpose and is anecdotal and quickly solved by the community. Nevertheless, there area a variety of contributors not all as passionate or precise in its work. Therefore, a description of user credibility could be interesting to evaluate user's work. Users credibility will be useful to review only a subset of data. Moreover, this credibility is also useful for users. A contributor could have a feedback about his contribution. A user cold have an idea about contributor reputation. As an example Wikiloc is a VGI project for outdoor tracks user get a rank depending on some parameters and the evaluation of the community,in this way other users can decide if the track is good enough for them. Another time here is the concept of fitness for use.

It is possible to go further in the concept of credibility, data is evaluated descriptively depending on amount of contributions or other users evaluation. Using INLA and SPDE it is possible to model data credibility base on a group of covariates and its neighbourhood relationship. INLA and SPDE can be applied using covariates coming from the same data source or metadata or complementary data source. VGI data used in the model was the eBird project. Bird observers over the world contribute to the project reporting birds position. Observers annotated some data relative to a bird observation, apart from bird and location, what is call metadata. This data is used in the model to evaluate a record, to give a credibility. Moreover, apart from this metadata is possible to relate a bird with external spatial covariates, like landscape. A wetland bird is difficult to be seen outside his ecosystem (landscape). Indeed, INLA and SPDE allows relating bird records spatially, so the output is a "spatial credibility" represented as traffic lights. This is a very useful tool for a citizen science project where data need a posterior revision. This procedure reduces the amount of data to be reviewed.

As the reader may notice all the projects used to evaluated the methodologies are coming form VGI or citizen science projects such as OSM or eBird. What Happens with Social Media Networks? The projects mentioned are open data or freely accessible, so is easy to get data to work on it. Social Media Networks mainly have an API as an interface to retrieve data from their database. Deal with an API could be relatively easy for a user with programming skills but an insuperable obstacle for those that do not have these skills. I propose a methodology for the science community to facilitate tweets retrieving and storage for a posterior analysis. Tweet2r package for *R* is a tool to deal with Twitter API able to get tweets using the streaming interface in real-time and store them in difference formats ready

to use.

Summarizing, this Thesis has approached the challenge of deal with VGI to enhance its value. To do this task spatial statistics is basic. Point pattern processes allow the analysis from point data depending on related covariates. INLA and SPDE allows data filtering base on credibility parameters and represent them as a traffic light model easy to understand by managers and users. VGI data also could be defined according a predefined set of quality and credibility parameters. These parameters could be different depending on participatory projects features or data collection methodologies. Nevertheless, a VGI project is based on people, mainly volunteers. Volunteers are the richness of these projects. Project design should take care on facilitate tools to improve contributions quality and avoid human mistakes. Managers should provide continuous feedback to the users, individually or collectively.

8.1 Future work

I started this Thesis some years ago thinking that I will get response to some research challenges. As soon as I started I realised that after finishing an issue a new challenge will open after it.

I have been following news and articles about digital divide, access to internet has been increasing since I started, thanks to wireless networks and smart devices. Despite this good new digital divide is always there, as access restrictions or lack of knowledge. Future work should go in the direction to make technology transparent. Therefore, people can participate independently of their educational level, programming skills or internet connection (offline mode). I have proposed a methodology but this could be enhanced with research to improve accessibility and brake users contribution restrictions.

This Thesis has open my mind to spatial statistics. I knew before the importance of statistics, but I did not imagine the close relation and grate potential between spatial statistics an GIS. VGI and citizen science has grown a lot on implementation and quantity of data collected. Future work should point to improve big data analysis, research on better methodologies to separate important data from noise.

Going beyond this Thesis I would like to explore tracking. Social media networks are here to be with us. It is possible to get space-time information. Being able to get tracking records. Spatial statistics will make possible to predict mass movements, but not only persons also natural hazards. Imagine forest fires that is being tracked by fire-fighters and citizens it will be possible to track a fire and to predict where the fire front will be and its implications to citizens security in advance. This future research also could be applied to animal migration, or to illness expansion.

Finally, I will stand up for open data, open source and open knowledge. The society is mature enough to be open, to change some productivity models. Science has been always open. The best scientists had shared to the world their new discoveries. **Science worth as wider is his accessibility**.

Appendix A

Appendix

A.1 Appendix chapter 4

List of available resource used at chapter 4:

- OSM data as shape data format.
- Data used to get the results in section 4.3 .
- OSMconnector application used in section 4.9, available at https://forja.uji.es/svn/osmconnector/osmconnector/.

Data sources are available at https://dx.doi.org/10.6084/m9.figshare.3112669.v1

A.2 Appendix chapter 6

List of available resource used at chapter 6:

- Input data for *R* code.
- *R* code .

Data sources and *R* are available at https://dx.doi.org/10.6084/m9.figshare.3112843

A.3 Appendix chapter 7

List of available resource used at chapter 7:

- Tweet2r package available at https://cran.r-project.org/web/packages/tweet2r/index.html.

- Data to test tweet2r package is available at https://dx.doi.org/10.6084/m9.figshare.2063529 .

Tweet JSON format stores tweet data. Here there are some examples how a tweet data is store in a JOSN files. The example is extended in Figure A.1.

- Tweet time stamp; "create_at":"Sun Sep 13 20:39:17 +0000 2015"
- User id;" id":643162053834842112
- Tweet text; "text":"Just posted a photo @ Nou Estadi Castalia https://t.co_HcTv2LOu7w"
- User's followers "followers_count":308
- Counter for favourite tweet tag "favourites_count":412
- Tweet coordinates; "geo":{"type":"Point","coordinates":[39.996053,-0.038792}}

The data coming from JSON file like the one shown in Figure A.1 is parsed ans stored in a postGIS database using the table structure of the Figure A.2, which shows the structure for the table that stores geo and non geo tweets. The previous data is stored into the table fields.

- Tweet time stamp as created_at; 2015-9-13 20:39:17+00
- User id as user_id_str ; 643162053834842112
- Tweet text as a text; "Just posted a photo @ Nou Estadi Castalia https:t.coHcTv2LOu7w"
- User's followers as followers_count
- latitude as lat, 39.996053
- longitude as lon; -0.038792
- Postgis geometry is stored in geom field the construction of this field is described in the postGIS documentation http://postgis.net/docs/manual-2.0/using_postgis_dbmanagement.html#OpenGISWKBWKT.

```
{"created_at":"Sun Sep 13 20:39:17 +0000 2015","id
":643162053834842112,"id_str":"643162053834842112","text":"
Just posted a photo @ Nou Estadi Castalia https:\/\/t.co\/
HcTv2LOu7w","source":"\u003ca href=\"http:\/\/instagram.com\"
 rel=\"nofollow\"\u003eInstagram\u003c\/a\u003e","truncated":
false,"in_reply_to_status_id":null,"in_reply_to_status_id_str
":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":
null,"in_reply_to_screen_name":null,"user":{"id":450992563,"
id_str":"450992563","name":"Lledo","screen_name":"LledoH","
location":"Los Angeles, CA","url":null,"description":"14. NV.
 Seny,pit i collons\u2764\ufe0f 14 de Junio","protected":
false,"verified":false,"followers_count":308,"friends_count
":275,"listed_count":3,"favourites_count":412,"statuses_count
":16044,"created_at":"Fri Dec 30 22:31:35 +0000 2011","
utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":"
es","contributors_enabled":false,"is_translator":false,"
profile_background_color":"DBE9ED","
profile_background_image_url":"http:\/\/abs.twimg.com\/images
\/themes\/theme17\/bg.gif","
profile_background_image_url_https":"https:\/\/abs.twimg.com
\/images\/themes\/theme17\/bg.gif","profile_background_tile":
false,"profile_link_color":"CC3366","
profile_sidebar_border_color":"DBE9ED","
profile_sidebar_fill_color":"E6F6F9","profile_text_color
":"333333","profile_use_background_image":true,"
profile_image_url":"http:\/\/pbs.twimg.com\/profile_images
\/630443721964814336\/xXPv1KnK_normal.jpg","
profile_image_url_https":"https:\/\/pbs.twimg.com\/
profile_images\/630443721964814336\/xXPv1KnK_normal.jpg","
profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners
\/450992563\/1439144471","default_profile":false,"
default_profile_image":false,"following":null,"
follow_request_sent":null,"notifications":null},"geo":{"type
":"Point","coordinates":[39.996053,-0.038792]},"coordinates
":{"type":"Point","coordinates":[-0.038792,39.996053]},"place
":{"id":"d687bdcf37d51c4f","url":"https:\/\/api.twitter.com
\/1.1\/geo\/id\/d687bdcf37d51c4f.json","place_type":"city","
name":"Castell\u00f3n de la Plana","full_name":"Castell\
u00f3n de la Plana, Comunidad Valenciana","country_code":"ES
","country":"Espa\u00f1a","bounding_box":{"type":"Polygon","
coordinates
":[[[-0.1641578,39.848549],[-0.1641578,40.0647026],
[0.6915727,40.0647026],[0.6915727,39.848549]]]},"attributes
":{}},"contributors":null,"retweet_count":0,"favorite_count
":0,"entities":{"hashtags":[],"trends":[],"urls":[{"url":"
https:\/\/t.co\/HcTv2LOu7w","expanded_url":"https:\/\/
instagram.com\/p\/7laGr1nPOG2lmAzI0qbLWoFVgyIDx8jec0Y0g0\/","
display_url":"instagram.com\/p\/7laGr1nPOG2l\u2026","indices
":[42,65]}],"user_mentions":[],"symbols":[]},"favorited":
false,"retweeted":false,"possibly_sensitive":false,"
filter_level":"low","lang":"en","timestamp_ms
":"1442176757664"}
```


Figure A.1: tweet stored using JSON format.

```
1
2 CREATE TABLE tweets
3 (
4   "row.names" text,
5   text text,
6   retweet_count double precision,
7   favorited boolean,
8   truncated boolean,
9   id_str text,
10  in_reply_to_screen_name text,
11  source text,
12  retweeted boolean,
13  in_reply_to_status_id_str text,
14  in_reply_to_user_id_str text,
15  lang text,
16  listed_count double precision,
17  verified boolean,
18  location text,
19  user_id_str text,
20  description text,
21  geo_enabled boolean,
22  user_created_at text,
23  statuses_count double precision,
24  followers_count double precision,
25  favourites_count double precision,
26  protected boolean,
27  user_url text,
28  name text,
29  time_zone text,
30  user_lang text,
31  utc_offset double precision,
32  friends_count double precision,
33  screen_name text,
34  country_code text,
35  country text,
36  place_type text,
37  full_name text,
38  place_name boolean,
39  place_id boolean,
40  place_lat double precision,
41  place_lon double precision,
42  lat double precision,
43  lon double precision,
44  expanded_url text,
45  url text,
46  int_id bigserial NOT NULL,
47  created_at timestamp with time zone,
48  geom geometry(Point,3857),
49  CONSTRAINT tweets_pkey PRIMARY KEY (int_id)
50 );
```

Figure A.2: SQL definition for the table that stores the tweets.

Appendix B

List of abbreviations

- API Application Programming Interface
- CSR Complete Spatial randomness
- DGPS differential GPS
- GF Gaussian Field
- GI Geospatial Information
- GIS Geographic Information System
- GMRF Gaussian Markov Random Field
- GPS Global Positioning Systems
- INLA Integrated Nested Laplace Approximation
- NHPP Homogeneous Poisson point process
- OSM Open Street Map
- PGIS participatory GIS
- PNOA Plan nacional de ortofotografia aerea
- R Free software environment for statistical computing and graphics
- RMSE Root Minimum Square Error
- SPDE tochastic Partial Differential Equation
- SDI Spatial Data Dnfrastructure
- VGI Volunteered Geographic Information
- WMS Web map service content.

Bibliography

Aggelopoulou, K. D., Wulfsohn, D., Fountas, S., Gemtos, T. A., Nanos, G. D., and Blackmore, S. (2009). Spatial variation in yield and quality in a small apple orchard. *Precision Agriculture*, 11:538–556.

Albuquerque, J. P. d., Herfort, B., Brenning, A., and Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 0(0):1–23.

Altieri, M. A. (2004). Linking ecologists and traditional farmers in the search for sustainable agriculture. *Frontiers in Ecology and the Environment*, 2(1):35–42.

Amatulli, G., Peréz-Cabello, F., and de la Riva, J. (2007). Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. *Ecological modelling*, 200(3):321–333.

Aragó Galindo, P., Díaz, L., and Huerta, J. (2011). A Quality approach to Volunteer Geographic Information. In *7th International Symposium on Spatial Data Quality (ISSDQ 2011). Raising awareness of Spatial Data Quality 2011, October, 12-14 Coimbra, Portugal*, pages 109–114, Coimbra, Portugal.

Avalos, C. and Alvarado, E. (1996). Space-time analysis of fire pattern in the blue mountains, oregon. In *13 th Fire and Forest Meteorology Conference, Lorne, Australia*, pages 413–420.

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns. *Australian & New Zealand Journal of Statistics*, 42(3):283–322.

Baddeley, A. J., Turner, R., et al. (2004). Spatstat: An r package for analyzing spatial point pattens.

Baddeley, A. J. and Van Lieshout, M. (1995). Area-interaction point processes. *Annals of the Institute of Statistical Mathematics*, 47(4):601–619.

Barberá, P. (2014). *streamR: Access to Twitter Streaming API via R.*

Barbera, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1):76–91.

Barbero, R., Abatzoglou, J., Steel, E. A., and Larkin, N. K. (2014). Modeling very large-fire occurrences over the continental united states from weather and climate forcing. *Environmental research letters*, 9(12):124009.

Barros, A. M. and Pereira, J. M. (2014). Wildfire selectivity for land cover type: does size matter? *PloS one*, 9(1):e84760.

Bastarrika, A. and Chuvieco, E. (2006). Cartografía del área quemada mediante crecimiento de regiones: aplicación en entornos mediterráneos con imágenes tm y etm+. *GeoFocus. Revista Internacional de Ciencia y Tecnología de la Información Geográfica*, (6):182–204.

Bishr, M. and Janowicz, K. (2010). Can we Trust Information?- The Case of Volunteered Geographic Information. In *Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, volume*, volume 640, Berlin, Germany.

Bishr, M. and Kuhn, W. (2007). Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In Fabrikant, S. I. and Wachowicz, M., editors, *The European Information Society*, Lecture Notes in Geoinformation and Cartography, pages 365–387. Springer Berlin Heidelberg. 10.1007/978-3-540-72385-1_22.

Bishr, M. and Mantelas, L. (2008). A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal*, 72(3-4):229–237.

Bivand, R., Keitt, T. H., Rowlingson, B., Pebesma, E., Summer, M., Hijmans, R., and Roualt, E. (2011). *rgdal: Bindings for the Geospatial Data Abstraction Library.*

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with r-INLA. *Spatial and Spatio-temporal Epidemiology*, 4:33–49.

Booltink, H., van Alphen, B., Batchelor, W., Paz, J., Stoorvogel, J., and Vargas, R. (2001). Tools for optimizing management of spatially-variable fields. *Agricultural Systems*, 70(2-3):445–476.

Bouma, J., Stoorvogel, J., and Booltink, H. (1999). Pedology, precision agriculture, and the changing paradigm of agricultural research. *Soil Science Soiciety of America Journal*, 63(6):1763–1768.

Caballero, D., Beltrán, I., and Velasco, A. (2007). Forest fires and wildland-urban interface in spain: types and risk distribution. In *En: IV Conferencia Internacional sobre Incendios Forestales. Sevilla*, pages 13–17.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97(2):109–131.

Carmo, M., Moreira, F., Casimiro, P., and Vaz, P. (2011). Land use and topography influences on wildfire occurrence in northern portugal. *Landscape and Urban Planning*, 100(1):169–176.

Coleman, D., Georgiadou, Y., and Labonte, J. (2009). Volunteered Geographic Information: the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, 4(4):332–358.

Coleman, D. J. (2010). The potential and early limitations of Volunteered Geographic Information. *Geomatica*, 64(2):209–219.

Conway, J., Eddelbuettel, D., Nishiyama, T., Prayaga, S., and Tiffin, N. (2013). *RPostgreSQL: R interface to the PostgreSQL database system (2010)*.

Cook, S. E. and Bramley, R. G. V. (1998). Precision agriculture — opportunities, benefits and pitfalls of site-specific crop management in australia. *Australian Journal of Experimental Agriculture*, 38(7):753.

Cook, S. E., O'Brien, R., Corner, R. J., and Oberthür, T. (2003). Is precision agriculture irrelevant to developing countries? In *European Conference on Precision Agriculture (4, 2003, Berlin, DE)*., page 6, Berlin.

Cornell Lab of Ornithology, . (2013). eBird Basic Dataset. Version: EBD_relNov-2013. Technical report, Cornell Lab of Ornithology, Ithaca, New York.

Craglia, M., Ostermann, F., and Spinsanti, L. (2012). Digital Earth from vision to practice: making sense of citizen-generated content. *International Journal of Digital Earth*, pages 1–19.

Cressie, N. (1993). Statistics for spatial data: Wiley series in probability and statistics. *Wiley-Interscience New York*, 15:16.

Cubo María, J. E., Enríqez Alcalde, E., Gallar Pérz-Pastor, J. J., López García, M., Mateo Díez, M. L., Muñoz Correal, A., Parra Orgaz,

P. J., and Jemes Díaz, V. (2012). Los incendios forestales en España. Decenio 2001-2010. Technical Report 280-12-210-8, Ministerio de Agricultura Alimentación y Medio Ambiente, Madrid.

De la Riva, J., Pérez-Cabello, F., Lana-Renault, N., and Koutsias, N. (2004). Mapping wildfire occurrence at regional scale. *Remote Sensing of Environment*, 92(3):363–369.

Del Hoyo, L. V., Isabel, M. P. M., and Vega, F. J. M. (2011). Logistic regression models for human-caused wildfire risk estimation: analysing the effect of the spatial accuracy in fire occurrence data. *European Journal of Forest Research*, 130(6):983–996.

Devillers, R. and Jeansoulin, R. (2006). Spatial Data Quality: Concepts. In Devillers, R. and Jeansoulin, R., editors, *Fundamentals of spatial data quality*, pages 31–42. ISTE, London ;;Newport Beach CA.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., and Shi, W. (2010). Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *Transactions in GIS*, 14(4):387–400.

Diagne, A. (2009). Technological change in smallholder agriculture: Bridging the adoption gap by understanding its source. In *Agriculture for Development in Sub-Saharan Africa*, Mombasa, Kenia. UC Berkeley: Center of Evaluation for Global Action.

Díaz-Avalos, C., Peterson, D. L., Alvarado, E., Ferguson, S. A., and Besag, J. E. (2001). Space time modelling of lightning-caused ignitions in the blue mountains, oregon. *Canadian Journal of Forest Research*, 31(9):1579–1593.

Diggle, P. J. et al. (1983). *Statistical analysis of spatial point patterns.* Academic press.

Dredze, M., Paul, M. J., Bergsma, S., and Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, pages 20–24. Citeseer.

Fan, H., Zipf, A., Fu, Q., and Neis, P. (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, 28(4):700–719.

Flanagin, A. and Metzger, M. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72(3):137–148.

Fountas, S., Aggelopoulou, K., Bouloulis, C., Nanos, G. D., Wulfsohn, D., Gemtos, T. A., Paraskevopoulos, A., and Galanis, M. (2010). Site-specific

management in an olive tree plantation. *Precision Agriculture*, 12(2):179–195.

Freeman, K. S. and Spyridakis, J. H. (2004). An Examination of Factors That Affect the Credibility of Online Health Information. *Technical Communication*, 51:239–263(25).

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., and Obersteiner, M. (2009). Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. *Remote Sensing*, 1(3):345–354.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, pages 256–274.

Gentry, J. (2015). *twitteR: R Based Twitter Client*.

Girres, J.-F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4):435–459.

GIS., T. C. O. a. L. (2005). *Specifications for 3D topographical cartography at scales of 1:1000 and 1:2000*. Diputació Barcelona Àrea d'Infraestructures Urbanisme i Habitatge, [Barcelona], 1st ed. edition.

Gonzalez, J. R., Palahi, M., Trasobares, A., and Pukkala, T. (2006). A fire probability model for forest stands in catalonia (north-east spain). *Annals of Forest Science*, 63(2):169–176.

Goodchild, M. F. (2001). Metrics of scale in remote sensing and GIS. *International Journal of Applied Earth Observation and Geoinformation*, 3(2):114–120.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.

Goodchild, M. F. (2008). Spatial Accuracy 2.0. In *Proceeding of the 8th int ernational symposium on spatial accuracy assessmen t in natural resources and environmental sciences*, pages 1–7, Shanghai, P. R. China.

Goodchild, M. F. and Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1:110–120.

Grabarnik, P., Myllymäki, M., and Stoyan, D. (2011). Correct testing of mark independence for marked point patterns. *Ecological Modelling*, 222(23):3888–3894.

Grilli, L., Metelli, S., and Rampichini, C. (2014). Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. *Journal of Statistical Computation and Simulation*, 0(0):1–9.

Gurtner, W. (2007). RINEX: The Receiver Independent Exchange Format Version 3.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4):682–703.

Haklay, M. and Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, 7(4):12–18.

Haklay, M. M., Basiouka, S., Antoniou, V., and Ather, A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *Cartographic Journal, The*, 47(4):315–322.

Hall, G. B., Chipeniuk, R., Feick, R. D., Leahy, M. G., and Deparday, V. (2010). Community-based production of geographic information using open source software and web 2.0. *International Journal of Geographical Information Science*, 24:761–781.

Hanewinkel, M., Hummel, S., and Albrecht, A. (2011). Assessing natural hazards in forestry for risk management: a review. *European Journal of Forest Research*, 130(3):329–351.

Hardy, C. C. (2005). Wildland fire hazard and risk: problems, definitions, and context. *Forest ecology and management*, 211(1):73–82.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

Hutchinson, T. F., Sutherland, E. K., and Yaussy, D. A. (2005). Effects of repeated prescribed fires on the structure, composition, and regeneration of mixed-oak forests in ohio. *Forest Ecology and Management*, 218(1):210–228.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.

ITU (2010). Measuring the information society 2010. Technical report, International Telecomunication Union.

James, J. (2008). The digital divide across all citizens of the world: A new concept. *Social Indicators Research*, 89(2):275–282.

Juan, P., Mateu, J., and Saez, M. (2012). Pinpointing spatio-temporal interactions in wildfire patterns. *Stochastic Environmental Research and Risk Assessment*, 26(8):1131–1150.

Jøsang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *R Journal*, 5(1).

Kong, X. (2007). GPS Modeling in Frequency Domain. In *Wireless Broadband and Ultra Wideband Communications, 2007. AusWireless 2007. The 2nd International Conference on*, page 61.

Koutsias, N. (2003). An autologistic regression model for increasing the accuracy of burned surface mapping using landsat thematic mapper data. *International Journal of Remote Sensing*, 24(10):2199–2204.

Koutsias, N., Kalabokidis, K. D., and Allgöwer, B. (2004). Fire occurrence patterns at landscape level: beyond positional accuracy of ignition points with kernel density estimation methods. *Natural Resource Modeling*, 17(4):359–375.

Koutsias, N., Xanthopoulos, G., Founda, D., Xystrakis, F., Nioti, F., Pleniou, M., Mallinis, G., and Arianoutsou, M. (2013). On the relationships between forest fires and weather conditions in greece from long-term national observations (1894–2010). *International Journal of Wildland Fire*, 22(4):493–507.

Krainski, E. T., Lindgren, F., Simpson, D., and Rue, H. (2015). The R-INLA tutorial: on SPDE models.

Lamb, D. W., Frazier, P., and Adams, P. (2008). Improving pathways to adoption: Putting the right p's in precision agriculture. *Computers and Electronics in Agriculture*, 61(1):4–9.

Lee, R. and Sumiya, K. (2010). Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection. In *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, LBSN '10, pages 1–10, New York, NY, USA. ACM.

Lenormand, M., Tugores, A., Colet, P., and Ramasco, J. J. (2014). Tweets on the Road. *PLoS ONE*, 9(8):e105407.

Lima, N., Casaca, J., and Henriques, M. (2006). Accuracy of Displacement Monitoring at Large Dams with GPS. In Sansò, F., Gil, A. J., and

Sansò, F., editors, *Geodetic Deformation Monitoring: From Geophysical to Engineering Roles*, volume 131 of *International Association of Geodesy Symposia*, pages 239–243. Springer Berlin Heidelberg. 10.1007/978-3-540-38596-7_29.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with r-INLA. *Journal of Statistical Software*, 63(19):1–25.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Ma, Y., Chen, L., Zhao, X., Zheng, H., and Lü, Y. (2009). What motivates farmers to participate in sustainable agriculture? evidence and policy implications. *International Journal of Sustainable Development & World Ecology*, 16(6):374.

Mann, K., Schumann, A., and Obreza, T. (2011). Delineating productivity zones in a citrus grove using citrus production, tree growth and temporally stable soil data. *Precision Agriculture*, 12:457–472. 10.1007/s11119-010-9189-y.

M.A.P.A (2003). *Libro blanco de la agricultura y el desarrollo rural.* Ministerio de Agricultura Pesca y Aalimentación Centro de Pûblicaciones, Madrid.

Martínez-Fernández, J., Chuvieco, E., and Koutsias, N. (2013). Modelling long-term fire occurrence factors in spain by accounting for local variations with geographically weighted regression. *Natural Hazards and Earth System Science*, 13(2):311–327.

Mateu, J., Uso, J., and Montes, F. (1998). The spatial pattern of a forest ecosystem. *Ecological Modelling*, 108(1):163–174.

Maué, P. and Schade, S. (2008). Quality of Geographic Information Patchworks. In *11th AGILE International Conference on Geographi c Information Science 2008*, Girona, Spain.

Mcknight, H. and Chervany, N. (1996). The Meanings of Trust. Working Paper Series 96-04, University of Minesota. Management Information Systems Research Center, Minesota.

McRae, R. H. (1992). Prediction of areas prone to lightning ignition. *International Journal of Wildland Fire*, 2(3):123–130.

Meng, T. (1998). Low-power GPS receiver design. In *Signal Processing Systems, 1998. SIPS 98. 1998 IEEE Workshop on*, pages 1–10.

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091.

Molin, J. (1997). Agricultura de precisao, parte 1: o que e o estado da arte em sensoriamento. *Engenharia Agricola (Brazil)*, 17(2):97–107.

Møller, J. and DÍaz-Avalos, C. (2010). Structured spatio-temporal shot-noise cox point process models, with a view to modelling forest fires. *Scandinavian Journal of Statistics*, 37(1):2–25.

Møller, J. and Waagepetersen, R. P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684.

Moreira, F., Viedma, O., Arianoutsou, M., Curt, T., Koutsias, N., Rigolot, E., Barbati, A., Corona, P., Vaz, P., Xanthopoulos, G., et al. (2011). Landscape–wildfire interactions in southern europe: implications for landscape management. *Journal of environmental management*, 92(10):2389–2402.

Oreilly, T. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, 1:17–38.

Ostermann, F. and Spinsanti, L. (2011). A Conceptual Workflow For Automatically Assessing The Quality Of Volunteered Geographic Information For Crisis Management. In *Proceedings of the 14th AGILE International Conference on Geographic Information Science*, page 6, Utrecht.

Pausas, J. G. (2004). Changes in fire and climate in the eastern iberian peninsula (mediterranean basin). *Climatic change*, 63(3):337–350.

Preisler, H. K. and Westerling, A. L. (2007). Statistical model for forecasting monthly large wildfire events in western united states. *Journal of Applied Meteorology and Climatology*, 46(7):1020–1030.

Pultar, E., Raubal, M., Cova, T. J., and Goodchild, M. F. (2009). Dynamic GIS Case Studies: Wildfire Evacuation and Volunteered Geographic Information. *Transactions in GIS*, 13:85–104.

Pôças, I., Gonçalves, J., Marcos, B., Alonso, J., Castro, P., and Honrado, J. P. (2014). Evaluating the fitness for use of spatial data sets to promote quality in ecological assessment and monitoring. *International Journal of Geographical Information Science*, 28(11):2356–2371.

Resnick, P., Kuwabara, K., Zeckhauser, R., and Friedman, E. (2000). Reputation systems. *Communications of the ACM*, 43(12):45–48.

Roberts, R. K., English, B. C., Larson, J. A., Cochran, R. L., Goodman, W. R., Marra, M. C., Martin, S. W., Shurley, W. D., and Reeves, J. M. (2004). Adoption of Site-Specific information and Variable-Rate technologies in cotton precision farming. *Journal of Agricultural and Applied Economics*, 36(1):148–148.

Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6(2):259–278.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. C&H/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC.

Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, 137(10):3177–3192.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Sassenrath, G., Heilman, P., Luschei, E., Bennett, G., Fitzgerald, G., Klesius, P., Tracy, W., Williford, J., and Zimba, P. (2008). Technology, complexity and change in agricultural production systems. *Renewable Agriculture and Food Systems*, 23(Special Issue 04):285–295.

Schade, S., Luraschi, G., De Longeville, B., Cox, S., and Díaz, L. (2010). Citizens as sensors for forest fires: sensor web enablement for volunteered geogrpaphic information. *ISPRS Workshop on Pervasive Web Mapping, Geoprocessing and Services XXXVIII-4/W13 (WebMGS 2010)*.

Schueller, J. K. (1992). A review and integrating analysis of Spatially-Variable control of crop production. *Fertilizer Research*, 33(1):1–34.

Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., and Mühlhäuser, M. (2013). A multi-indicator approach for geolocalization of tweets. In *ICWSM*.

Scott, A. C. (2000). The pre-quaternary history of fire. *Palaeogeography, palaeoclimatology, palaeoecology*, 164(1):281–329.

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Díaz-Ávalos, C., and Rue, H. (2014). Spatio-temporal log-gaussian cox processes for modelling wildfire occurrence: the case of catalonia, 1994–2008. *Environmental and ecological statistics*, 21(3):531–563.

Sieber, R. (2006). Public participation geographic information systems: A literature review and framework. *Annals of the Association of American Geographers*, 96(3):491–507.

Sloan, L. and Morgan, J. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PLoS ONE*, 10(11):e0142209.

Srinivasan, A. (2006). *Handbook of precision agriculture : principles and applications*. Food Products Press, Bringhamton NY.

Stoorvogel, J. (2006). Precision farming and smallholders.

Sui, D., Goodchild, M., and Elwood, S. (2013). Volunteered geographic information, the exaflood, and the growing digital divide. In *Crowdsourcing geographic knowledge*, pages 1–12. Springer.

Team, R. C. (2014). R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. 2013.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240.

Tulloch, D. L. (2007). Many, many maps: Empowerment and online participatory mapping. *First Monday*, 2(2).

Turner, A. (2006). *Introduction to neogeography*. " O'Reilly Media, Inc.".

Turner, R. (2009). Point patterns of forest fire locations. *Environmental and ecological statistics*, 16(2):197–223.

Ueda, S., Yamaguchi, Y., Kitagawa, H., and Amagasa, T. (2015). Tweet Location Inference Based on Contents and Temporal Association. In Wang, J., Cellary, W., Wang, D., Wang, H., Chen, S.-C., Li, T., and Zhang, Y., editors, *Web Information Systems Engineering – WISE 2015*, number 9419 in Lecture Notes in Computer Science, pages 259–266. Springer International Publishing. DOI: 10.1007/978-3-319-26187-4_22.

Van Oort, P. (2005). *Spatial data quality : from description to application*. PhD thesis, NCG Nederlandse Commissie voor Geodesie, Delft.

Van Wart, S., Tsai, K. J., and Parikh, T. (2010). Local ground:: a paper-based toolkit for documenting local geo-spatial knowledge. In *ACM DEV '10 Proceedings of the First ACM Symposium on Computing for Development*, page 1. ACM Press.

Vázquez, A. and Moreno, J. M. (1998). Patterns of lightning-, and people-caused fires in peninsular spain. *International Journal of Wildland Fire*, 8(2):103–115.

Vélez, R. (2000). *La defensa contra incendios forestales: fundamentos y experiencias.* Number 577.20946 D313. McGraw-Hill.

Veregin, H. (1999). Data quality parameters. In Goodchild, M. F., Maguire, D. J., and Rhind, D. W., editors, *Geographical Information Systems*, volume Principles and Technical Issues, pages 177–189. John Wiley & Sons Ltd.

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous neyman–scott processes. *Biometrics*, 63(1):252–258.

Whitney, J. D., Ling, Q., Miller, W., and Wheaton, T. (2001). A DGPS yield monitoring system for florida citrus. *Applied engineering in agriculture*, 17(2):115–120.

Wickham, H., James, D. A., Falcon, S., SQLite, Healy, L., and RStudio (2014). *RSQLite: SQLite Interface for R.*

Wiersma, Y. F. (2010). Birding 2.0: Citizen Science and Effective Monitoring in the Web 2.0 World. *Avian Conservation and Ecology*, 5(2).

Wing, M. G., Eklund, A., and Kellogg, L. D. (2005). Consumer-Grade Global Positioning System (GPS) Accuracy and Reliability. *Journal of Forestry*, 103(4):169–173.

Wittenberg, L. and Malkinson, D. (2009). Spatio-temporal perspectives of forest fires regimes in a maturing mediterranean mixed pine landscape. *European Journal of Forest Research*, 128(3):297–304.

Zielstra, D. and Zipf, A. (2010). A comparative study of proprietary geodata and volunteered geographic information for germany. In *13 th AGILE International Conference on Geographic Information Science 2010*, Guimarães, Portugal.