Formalization and Automated Detection of Tourist City Center Location

Ruslan Akhundov[1,2], Andrey Filchenkov[1], Vladimir Gorovoy[2]
[1] ITMO University, St. Petersburg, Russia
[2] Yandex, Moscow, Russia
akhundov2@gmail.com, afilchenkov@corp.ifmo.ru, vladimir.gorovoy@gmail.com

***Abstract*—Nowadays, more and more people tend to plan and book their vacation using services such as Yandex.Travel, Booking.com or TripAdvisor. When most of the tourists are booking a hotel, they want to know, where the touristic center of city is located. However, it is an informal concept. In this work we conducted a survey to obtain the ground truth on what people thought to be a touristic city centre of their city, and then approximated it with the mixture of Gaussians (ellipses). Finally, we have suggested an algorithm to predict a tourist city centre location given information on hotels in that city.**

I. Introduction

When people are planning their vacation in another city they have never been to before, most of them would like to know, where the centre of that city is located. It would help, for example, to avoid possible inconveniences such as huge everyday transport fees to get there. An option to filter hotels by location in the city center or by distance to it is one of the users' most demanded features in Yandex.Travel https://travel.yandex.ru, an online service helping users to find the right hotel and compare prices from different providers. Because the number of cities is extremely high and its very costly to use expert knowledge about each city's centre, an interesting task arises: how can one automatically locate a city centre using some available information about that city.

The main hardness of the problem of automated detection of city centre location is that the concept of city centre is informal itself despite being used in a large number of studies. City center is an object of research in urban studies [3], [26], [25], [14], [2]; economic studies [1], [16], [7], [24], [23], [13], [29] due to the high impact of urban centers in trading and its major role in retail (interestingly, all these papers are devoted solely to European cities); climatology [20], [19], [17], [9], [10], [8], [12], [15] due to impact of city centers on climate is different than impact of other areas; or even public safety studies [18], [21], [22], [27], due to high economic impact of the centres and high present of public, making it an attractive venue of various crimes.

Typically, authors of those studies do not define what a city center is. We believe that it is due to they used "city centre" as a linguistic but not scientific term. The only definition we found was given in [28], in which city center was defined as "an area, central to the city as a whole, in which the main land uses are commercial". This definition is obviously biased towards the object of their study. Also none of the authors explains how they specified that concept for data collection, which makes us believe they did it in an expert way. A study, which is worth to be discussed with a little more attention, is a paper [6], the authors of which suggested several descriptive models of city structure (including its center) with respect to their retailing functions. This model is applicable for expert analysis, however, it cannot be used for automated city center detection due to lack of precise definitions for concepts used in these models.

Thus, a solution of a city centre automated location determination problem requires the formalization of the city centre concept. Because we focus on tourists, we are interested in touristic city centre (TCC). In this paper, we propose an approach to formalize this concept in a computationally suitable manner and present an approach to determine its location automatically given information about hotels in this city.

The rest of the paper is organized as follows. In Section 2, we describe requirement that a definition of TCC should met and infer an approach how it can be formalized. In Section 3, we describe data we collect using a survey and results of approximating it. In Section 4, we propose an approach for automated TCC location detection. In Section 5, we describe data we use and experiment setup, results of which is presented in Section 6. Section 7 concludes the paper.

II. Formalization and evaluation of TCC

In this section, we provide mathematical formalization of a TCC and motivation under such formalization, as well as the algorithm for evaluating TCC location.

A. Concept of TCC

First of all, we need to restrict the domain of application of the formal TCC definition we present in this work. We claim that TCC is not a point, but an area, which we want to detect.

Thus, we need to define TCC in a easily computable way, because it is the aim of our prediction and it will be used as the ground truth for learning predictive models. The only source where we can find a city center formal definition is an administrative division of cities. The key problem with this source is not the fact that in many cases there is no such a district as "Central" (an example is Samara, Russia), but that even if such a district exists, it is very imprecise with representation of TCC (an example is St. Petersburg, Russia).

We assume that we the only valid source for representing TCC is opinion of citizens. Thus, we conclude that a survey is the only method to locate TCC. However, under this approach, a new question arises: what is the form of this opinion, which

should be collected? TCC is a very informal concept with very vague bounds, and this is why there is no purpose in asking about the precise bounds of TCC. Thus, the survey has an only question: "Where is located touristic centre of your city?". An answer may be any area of the city (figure on the city map) marked by a respondent.

The next question we need to answer is how the resulting answer should look like. Working with polygons is not very convenient. Also, as we mentioned before, there is no need in a delicate handling with shapes, because the initial figures are imprecise, and a certain shape retrieved via intersecting will be thus biased towards the sample. It motivates us to work not with random figures, but with their smooth approximations. We found that an ellipse is a very good option to be used for that purpose, because of the following reasons:

- an ellipse provides a certain shape complexity, being simple to parameterize. It is a bit more complex than a circle, but much simpler, than any other smooth figure.
- an ellipse is a very convenient figure for prediction due to its relatedness to the Gaussian distribution.

Thus, we consider each TCC to be an ellipse or a union of a small number of ellipses.

B. Evaluating of an citizens' opinion ellipsoidal approximation

In the previous subsection we motivated that a TCC should be represented by a union of ellipses, which should approximate citizens' opinion on what the tourist center of their city is. In this subsection we describe the algorithm to evaluate the most suitable union of ellipses given a set of figures representing citizens' opinions.

After we obtain several citizens' opinions on what is a city centre in a form of set of figures, we need it to be approximated with a set of ellipses. To find and draw ellipses, we use EM algorithm for Gaussian mixtures, and also on the initialization step we use k-Means [5] to learn the initial means, as described in [4]. For each grid point, we calculate the number of covering figures and normalized it.

In order to determine the ellipse size for the particular city, we need to introduce several definitions: *Central point* (C) is a point such that it lies within more than a half of citizens' figures. *Peripheral point* (P) is a point such that it lies within less than a half of citizens' figures. Let $C(E)$ denote the number of central points lying within ellipse E and $P(E)$ be the number of peripheral points lying within it. Ellipse size can be determined as a Mahalanobis distance between its center and a point on its border. We want to choose a border point that maximizes $C(E) - P(E)$ for resulting ellipse. It seems clear that the zone covered with central points is an optimal candidate to be a touristic city center.

We also introduce function $f(e, E)$, which we will be maximized. This function for specified ellipse e and set of ellipses E :

- should increase when e covers more central points,
- should decrease when e covers more peripheral points,
- should not change its value when covering points that have already been covered with another ellipse from E.

Such function can be defined by the loss matrix presented on Table I with a parameter λ corresponding to how fast the function will decrease when the ellipse covers more peripheral points.

TABLE I. LOSS MATRIX FOR DIFFERENT TYPES OF POINTS

	Central	Peripheral
Uncovered Point	1	$-\lambda$
Point already covered with another ellipse	0	0

Let us take a look at the iterative algorithm. At the very beginning, we start with the empty set of ellipses. We create a grid with a fixed step size covering the city and compute for each point of this grid if it is peripheral or central. After that we find such ellipse e that it maximize $f(e)$. An ellipse can be parametrized by five variables: two center coordinates (x, y), two lengths of half axes (r_1, r_2) and angle. For each point of the grid, we assume that it is an ellipse center and maximize $f(e)$ by changing r_1, r_2 and $angle$ with coordinate descent method. Then we choose the ellipse with maximum $f(e)$ value. We will add ellipses to resulting set while $f(e') > f(e) + \varepsilon$, where ε is some constant greater than zero. As the result, we obtain a set of ellipses that approximates citizens' opinions.

III. EVALUATION OF THE GROUND TRUTH TCC

A. Data collecting

For each city, we asked its citizens to draw a territory on the map, where they think the city tourist center is located. Each of the citizens was asked to draw his/her city tourist center as a polygonal (he/she can pick points of this polygonal). We created a specified web service and used it to ask people from different cities to draw their TCC. To implement this web service, we used JavaScript, HTML and Google maps API on the front-end, and Java, Spring Boot, MongoDB, and Gradle on the back-end side. We chose MongoDB as database, because it had geospatial index, which provided possibility to find nearest points easily.

For this study, we chose only cities from the former Soviet republics (mostly, from Russia), because they corresponds to the Yandex.Travel audience interests. We worked with the following twenty cities:

- Almaty (Kazakhstan),
- Baku (Azerbaijan),
- Chelyabinsk (Russia),
- Ekaterinburg (Russia),
- Kazan (Russia),
- Kiev (Ukraine),
- Kharkiv (Ukraine),
- Minsk (Belarus),
- Moscow (Russia),
- Nizhny Novgorod (Russia),

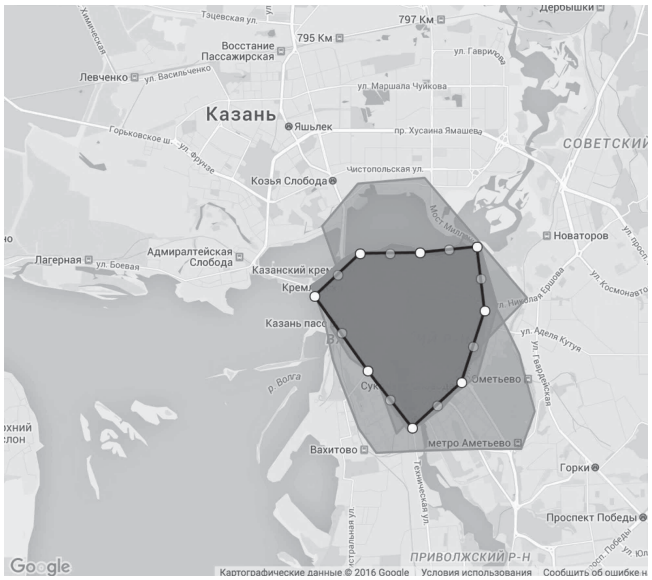

Fig.1. Results of merging opinions on Kazan TCC

Fig. 2. Results of merging opinions on Nizhny Novgorod TCC

- Novosibirsk (Russia),
- Odessa (Ukraine),
- Omsk (Russia),
- Rostov-on-Don (Russia),
- Samara (Russia),
- Sochi (Russia),
- St. Petersburg (Russia),
- Tashkent (Uzbekistan),
- Volgograd (Russia),
- Yerevan (Armenia).

For each city from this list, we collected at least 15 opinions of its For each city from this list, we collected at least 15 opinions of its citizens. Examples of merged opinions are presented on Fig. 1 and Fig. 2. The more intensive the grey color of a point on the map is, the more polygons cover this point.

B. Results of ground truth evaluation

The results of the algorithm application are listed in Table II. For the most of the cities, one ellipse is enough to cover citizens' opinions. However, for Kazan, two overlapping ellipses were required.

Examples of algorithm application are presented in Fig. 3–6.

Table II. RESULTS OF TCC EVALUATION

City	Number of answers
St. Petersburg	114
Moscow	61
Kiev	15
Tashkent	16
Baku	19
Minsk	17
Novosibirsk	15
Almaty	15
Kharkiv	16
Ekaterinburg	16
Nizhniy Novgorod	19
Samara	17
Sochi	16
Kazan	17
Omsk	15
Chelyabinsk	15
Erevan	15
Rostov-na-Don	16
Volgograd	17
Odessa	15

IV. DETECTION OF TCC LOCATION

A. TCC as a Gaussian mixture

As we described in the previous section, we will try to work with TCC as a composition of ellipses. A very straight way to achieve this is to use the Gaussian mixture model (GMM) due to the Gaussian distribution is ellipsoidal in geometric sense. An ellipse is described by the covariation matrix Σ and the vector of mean values μ of Gaussian distribution.

In order to apply GMM, we need to work with a density function. We use a grid with a fixed step size. For each grid node x and set of touristic objects O, we define density function as:

$$d(x) = \sum_{o \in O} K(dist(x, o)),$$

where $dist(x, y))$ is a distance between two objects and K is a kernel function.

Using the density function, we calculate the density relative to tourist objects for each grid node. Then we applied EM algorithm for GMM:

$$p(x;\theta) = \sum_{i=1}^{k} w_i p_i(x;\theta_i), \sum_{i=1}^{k} = 1, w_i \geq 0, \theta = \{\theta_k, w_k\}_{i=1}^{k}.$$

EM algorithm returns k Gaussians, for each of them, we know its covariation matrix and mean value.

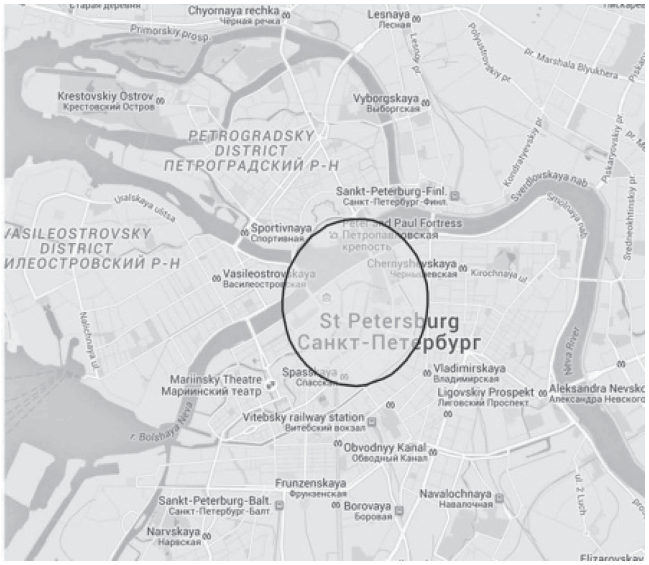

Fig. 3. Resulting TCC for St. Petersburg (union of two ellipses

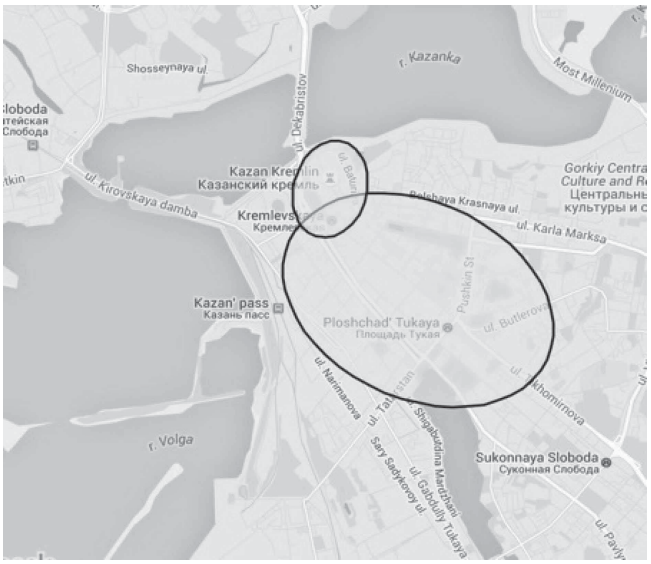

Fig. 4. Resulting TCC for Kazan

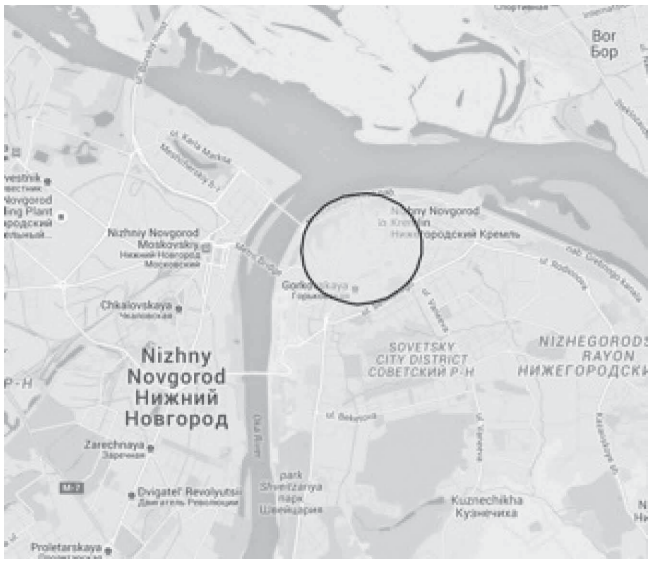

Fig. 5. Resulting TCC for Nizhny Novgorod some text here some text here

Fig. 6. Resulting TCC for Moscow

Let $\mathrm{likelihood}(k)$ be likelihood of data approximated with k Gaussians using EM-algorithm. We want to maximize it, which means that we need to find the number of Gaussian distributions, with which the density of touristic objects is well represented. In order to do so, we run the algorithm with different values of k and take a look at the resulting likelihood values. We will take such k value that the following inequality holds:

$$|\mathrm{likelihood}(k) - \mathrm{likelihood}(k+1)| < \varepsilon,$$

where ε is a positive constant.

In order to visualize the resulting ellipses, we need to calculate for each of them its center, two radii and an angle. Coordinate vector of an ellipse center equals to its mean value vector. In order to calculate the angle, we take eigenvector $\mathbf{v}_1$ of the biggest eigenvalue and use the known formula:

$$\alpha = \arctan\frac{\mathbf{v}_1(y)}{\mathbf{v}_1(x)}.$$

Now we need to evaluate the ellipse size. In our case, we can express it in terms of standard deviations σ_x and σ_y, such that the ellipse is expressed by the following equation:

$$\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2 = s,$$

where s is the ellipse size and can have any value. Since we operate in terms of the normal distribution, the left-hand side of the equality given above is the sum of the squares of the independent normally distributed quantities. It is

known that this quantity has a chi-square distribution. The chi-square distribution is defined in terms of freedom degrees, in our case there are two degrees of freedom. Thus, the value of the variable s corresponds to certain values of the chi-square distribution. Accordingly, the lengths of the semiaxes of the ellipse can be expressed as follows: $r_1 = 2\sigma_x\sqrt{s}$, $r_2 = 2\sigma_y\sqrt{s}$.

The equalities above can be expressed in terms of the eigenvalues λ_1, λ_2 of the covariance matrix: $r_1 = 2\sqrt{s\lambda_1}$, $r_2 = 2\sqrt{s\lambda_2}$.

Note that the parameter s in this case is a critical point, such that for some probability p: $P(x < s) = p$, where $p \in [0, 1]$.

Thus, we reduced the problem of finding the lengths of two semiaxes of an ellipse to the problem of finding the value of p that is probability of the chi-square distribution, the range of possible values of which is from 0 to 1.

In order to determine the value of the parameter p, it is enough to iterate through all possible values from 0 to 1 with a small step. For each of the values, we construct the resulting ellipses and evaluate its accuracy in comparison with the reference ellipse, built on the opinions of the city dwellers.

V. Detection of TCC location using information about tourist objects

After we evaluated TCC for each of the city in our dataset, we can solve the problem of detecting its location using open information.

A. Experiments

To learn a model, we used an aggregated hotel dataset which was generously provided by Yandex.Travel company and contained all hotels in the selected cities, total number of which was 820,000. Each hotel was described by its location (longitude and latitude), number of stars and average price per night. We decided to exclude several other potentially useful features, otherwise it can limit the model application. For instance, hotel-specific features are not always available, or the number of reviews depends on both service and city popularity.

We used a grid with the fixed step size equal to 100 meters to obtain a discrete space. To assess the model efficacy, we used citizens' ellipses to mark the grid nodes. The resulting ellipses should cover only the ones that were covered by the citizens' ellipses. As a quality measure, we used F_1 score and accuracy.

For model learning and testing, we used 4-fold cross-validation. Samples (20 cities) were divided into two sets: 15 cities for the training set, and 5 for test set.

To learn a density function, we used different approaches. Initially, we had distribution of points with two characteristics: number of stars and price. We tried several models that used these characteristics, for instance, that gave bigger weights to the hotels with higher number of stars or weights depending on their price. We also tried to filter hotels according to their price. However, since none of these models outperformed a simple model we used, we will not describe them in details.

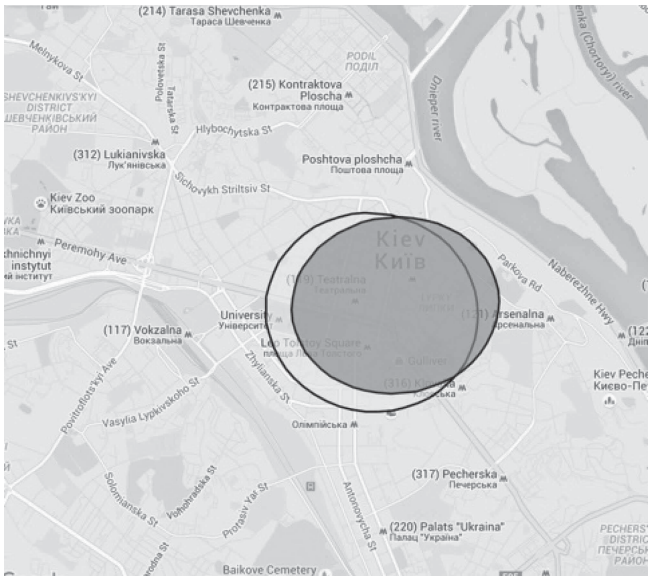

Fig. 7. Results for Kiev, the dark ellipse is built using data about touristic objects, and the light one is built using people opinions

The simple model used only the number of hotels. For each grid node it evaluated Epanechnikov kernel function [11].

Having a density function, we applied Expectation Maximization algorithm for Gaussian Mixture. We got set of Gaussians as an output, after that we built ellipses using the proposed algorithm. Next, we selected the coefficient p to calculate the lengths of the semiaxes which maximized model quality (F-measure) on test set.

B. Results

The model quality is presented in Table III. The difference between F_1 measure and Accuracy are almost similar on both train and test, so we model is well-trained for such a small sample. The resulting parameter p value is 0.017.

Table III. Results of TCC evaluation

	F_1	Acc
Train set	0.792	0.797
Test set	0.769	0.773

Example of results for Kiev and St. Petersburg are presented in Fig. 7 and Fig. 8.

We presented the obtained results to content experts in Yandex.Travel, and they confirmed that it is more accurate than such naïve approaches as the union or intersection of all the experts' opinions.

VI. Conclusion

In this paper, we have suggested an approach to define formally, what the TCC of a city is. Also, we proposed a method to compute it as a union of ellipses. This method is confirmed by content experts to show acceptable quality of results. We also described an idea how to use expectation maximization algorithm for points considering that it could intersect more than one polygon, by repeating this point in the test set.

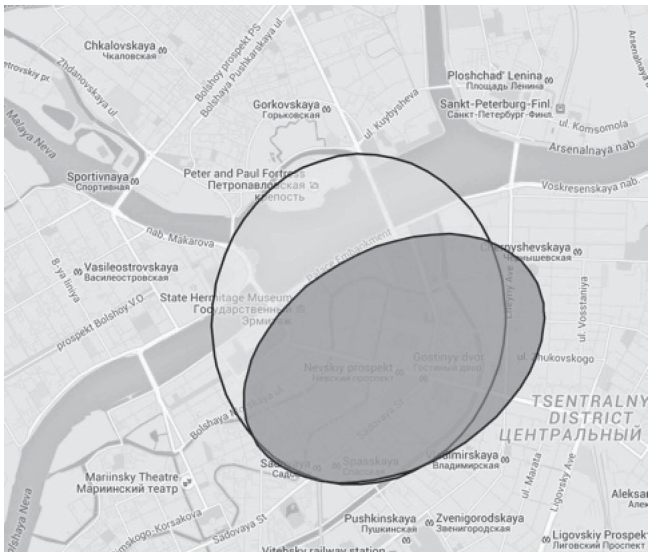

Fig. 8. Results for St. Petersburg, the dark ellipse is built using data about touristic objects, and the light one is built using people opinions

Described results can be used in the future research as a ground truth for the task of automatically determining TCC when expert opinions are not available. They will form a learning dataset and machine learning approach could be applied to detect TCC location based on factors such as tourist objects densities.

ACKNOWLEDGMENT

Authors would like to thank Sergey Muravyov for useful comments. This work was financially supported by the Government of Russian Federation, Grant 08-08.

REFERENCES

[1] Aloys Borgers and HJP Timmermans. City centre entry points, store location patterns and pedestrian route choice behaviour: A microlevel simulation model. *Socio-economic planning sciences*, 20(1):25–31, 1986.

[2] Rosemary DF Bromley, Andrew R Tallon, and Colin J Thomas. Disaggregating the space–time layers of city-centre activities and their users. *Environment and Planning A*, 35(10):1831–1851, 2003.

[3] Stuart Cameron. Housing, gentrification and urban regeneration policies. *Urban Studies*, 29(1):3–14, 1992.

[4] Yihua Chen and Maya R Gupta. Em demystified: An expectation-maximization tutorial. In *Electrical Engineering*. Citeseer, 2010.

[5] Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural networks: Tricks of the trade*, pages 561–580. Springer, 2012.

[6] RL Davies. Structural models of retail distribution: analogies with settlement and urban land-use theories. *Transactions of the Institute of British Geographers*, pages 59–82, 1972.

[7] Tim Dixon and Andrew Marston. Uk retail real estate and the effects of online shopping. *Journal of Urban Technology*, 9(3):19–47, 2002.

[8] Krzysztof Fortuniak, Włodzimierz Pawlak, and Mariusz Siedlecki. Integral turbulence statistics over a central european city centre. *Boundary-layer meteorology*, pages 1–20, 2013.

[9] Fumiaki Fujibe. Weekday-weekend differences of urban climates. *Journal of the Meteorological Society of Japan. Ser. II*, 65(6):923–929, 1987.

[10] Fumiaki Fujibe. Weekday-weekend differences of urban climates. *Journal of the Meteorological Society of Japan. Ser. II*, 66(2):377–385, 1988.

[11] Wolfgang Härdle, Marlene Müller, Stefan Sperlich, and Axel Werwatz. *Nonparametric and semiparametric models*. Springer Science & Business Media, 2012.

[12] PS Jackson. Wind structure near a city centre. *Boundary-Layer Meteorology*, 15(3):323–340, 1978.

[13] Colin Jones and Nicola Livingstone. Emerging implications of online retailing for real estate: Twenty-first century clicks and bricks. *Journal of Corporate Real Estate*, 17(3):226–239, 2015.

[14] Christopher M Law. Regenerating the city centre through leisure and tourism. *Built Environment (1978-)*, pages 117–129, 2000.

[15] Alastair C Lewis, Dorota Kupiszewska, Keith D Bartle, and Michael J Pilling. City centre concentrations of polycyclic aromatic hydrocarbons using supercritical fluid extraction. *Atmospheric Environment*, 29(13):1531–1542, 1995.

[16] Erika Nagy. Winners and losers in the transformation of city centre retailing in east central europe. *European Urban and Regional Studies*, 8(4):340–348, 2001.

[17] K Nakagawa and C Nakayama. The relationship between surface albedo and surface structure in the central parts of urban areas in the kanto plain, japan. *Geographical Review of Japan Series A*, 68:741–760, 1995.

[18] Amanda L Nelson, Rosemary DF Bromley, and Colin J Thomas. Identifying micro-spatial and temporal patterns of violent crime and disorder in the british city centre. *Applied Geography*, 21(3):249–274, 2001.

[19] B Offerle, Christine Susan B Grimmond, Krzysztof Fortuniak, Kazimierz Kłysik, and Timothy R Oke. Temporal variations in heat fluxes over a central european city centre. *Theoretical and applied climatology*, 84(1):103–115, 2006.

[20] Brian Offerle, Christine Susan B Grimmond, and Krzysztof Fortuniak. Heat storage and anthropogenic heat flux in relation to the energy balance of a central european city centre. *International Journal of Climatology*, 25(10):1405–1419, 2005.

[21] Rachel Pain and Tim Townshend. A safer city centre for all? senses of community safety in newcastle upon tyne. *Geoforum*, 33(1):105–119, 2002.

[22] Malcolm Ramsay. *Downtown Drinkers: the perceptions and fears of the public in a city centre*. Home Office, 1989.

[23] Orit Rotem-Mindali. E-commerce: Implications for travel and the environment. In *Handbook of Sustainable Travel*, pages 293–305. Springer, 2014.

[24] Bas Spierings. The return of regulation in the shopping landscape? reflecting on the persistent power of city centre preservation within shifting retail planning ideologies. *Tijdschrift voor economische en sociale geografie*, 97(5):602–609, 2006.

[25] Andrew R Tallon and Rosemary DF Bromley. Exploring the attractions of city centre living: evidence and policy implications in british cities. *Geoforum*, 35(6):771–787, 2004.

[26] Zilai Tang and Peter WJ Batey. Intra-urban spatial analysis of housing-related urban policies: The case of liverpool, 1981-91. *Urban Studies*, 33(6):911–936, 1996.

[27] CJ Thomas and RDF Bromley. Safety and shopping: peripherality and shopper anxiety in the city centre. *Environment and Planning C: Government and Policy*, 14(4):469–488, 1996.

[28] Jesse WJ Weltevreden. Substitution or complementarity? how the internet changes city centre shopping. *Journal of Retailing and consumer Services*, 14(3):192–207, 2007.

[29] Neil Wrigley and Dionysia Lambiri. British high streets: from crisis to recovery? *A Comprehensive Review of the Evidence. Southampton*, 2015.