Preferential centrality - a new measure unifying urban activity, attraction and accessibility

Citation for the original published paper (version of record):
Hellervik, A., Nilsson, L., Andersson, C. (2019). Preferential centrality - a new measure unifying urban activity, attraction and accessibility. Environment and Planning B: Urban Analytics and City Science, 46(7): 1331-1346. http://dx.doi.org/10.1177/2399808318812888

Preferential centrality - a new measure unifying urban activity, attraction and accessibility

Accepted for publication in Environment and Planning B: Urban Analytics and City Science
DOI: 10.1177/2399808318812888

Alexander Hellervik[1], Leonard Nilsson, Claes Andersson
Chalmers University of Technology, Gothenburg, Sweden

Abstract

The fact that accessibility shapes the geographic distribution of activity needs to be addressed in any long-term policy and planning for urban systems. One major problem is that current accessibility measures rely on the identification and quantification of attractions in the system. We propose that it is possible to devise a network centrality measure that bypasses this reliance and predicts the distribution of urban activity directly from the structure of the infrastructure networks over which interactions take place. From a basis of spatial interaction modelling and eigenvector centrality measures we develop what we call a *preferential centrality* measure that recursively and self-consistently integrates activity, attraction and accessibility. Derived from the same logic as Google's PageRank algorithm, we may describe its operation by drawing a parallel: Google's PageRank algorithm ranks the importance of networked documents without the need to perform any analysis of their contents. Instead it considers the topological structure of the network and piggybacks thereby on contextualized and deep evaluation of documents by the myriad distributed agents that constructed the network. We do the same thing with regard to networked geographical zones. Our approach opens up new applications of modelling and promises to alleviate a host of recalcitrant problems, associated with integrated modelling, and the need for large volumes of socioeconomic data. We present an initial validation of our proposed measure by using land taxation values in the Gothenburg municipality as an empirical proxy of urban activity. The resulting measure shows a promising correlation with the taxation values.

Keywords

Accessibility, urban activity, centrality, Eigenvector centrality, preferential attachment, PageRank, transportation, spatial regression, land value, spatial interaction

[1] alexander.hellervik@chalmers.se

1 Introduction

Spatial interaction is essential for urban activity and is ultimately afforded by the transportation network. Can the geographical distribution of urban activity thereby be inferred directly from some measure of centrality derived from the transportation system? In this paper we combine theories from spatial interaction modelling (e.g. Wilson, 2000), and network centrality (e.g. Newman, 2008) to develop a model to test this hypothesis with encouraging results. As a framing, we begin by subdividing the problems faced by planners and theorists into: *a planning problem* that carries with it *a modelling problem*, and *a data problem*.

The planning problem concerns the need to integrate transport and land use to handle dynamical consequences of change. At its heart, the planning problem stems from the essential unpredictability of complex interactions within and between domains. For example, a newly constructed road may itself increase traffic by inducing new development attracted to improved accessibility along its extent.

Computational models are attractive as tools for studying these dependencies, which leads us to *the modelling problem.* If we begin unpacking the transportation and land use domains, many levels of fine-grained subsystems appear (e.g. Iacono et al., 2008). To make matters worse, these subsystems are not as internally integrated and externally separated as we may wish. Integrated models are near-decomposable (Simon, 1962) in a *complicated* machine-like manner, while urban systems are *wicked* (Andersson et al., 2014). Integrated model systems and urban systems are not complex in the same way (Timmermans, 2003).

However, even if we were to solve the modelling problem, we would still be left with a *data problem*. Attempting to improve realism by integrating as much theoretical and empirical detail as possible (e.g. Waddell et al., 2003) leads to a two-fold problem. First, suitable and consistent data must be obtained. Second, empirical patterns must be expected to remain valid even as planning parameters are changed, which is particularly problematic for long term forecasts.

Our approach is to strike at the modelling and data problems simultaneously by exploring an alternative approach. We aim to infer the distribution of urban activity, by modelling only the physical characteristics of geographical zones and their interactions, i.e. without reliance on *any* demographic data. Our centrality measures are derived from the same basis as Google's PageRank algorithm (Brin and Page, 1998), but in our case the main input is the transportation network, which is used to infer the importance – or centrality – of the zones that it links. Our hypothesis is that this centrality concept is intimately linked with the concept of urban activity. The result is an expandable, scalable and portable model based on new principles that bypasses some of these key modelling and data problems in planning. The model may be re-applied anywhere in the world, and, with regard to data availability, it may be scaled up to the global level, opening up new vistas of possible applications besides those of traditional planning.

The first part of the paper concerns theoretical background and derivation of centrality models for predicting urban activity. We then present our data sources, followed by methods and results sections where the model implementation and empirical validation processes are described.

2 Theory

2.1 Background

From the common wisdom that cities tended, from early on, to be established on trade routes, natural ports or river crossings stems the fundamental assumption of all spatial economic theories: a location with good accessibility is more attractive than locations with bad access. This is a fundamental assumption that theoretically goes back to von Thünen (1826). A break-through study by Hansen (1959) demonstrated that locations with high accessibility were developed earlier and more densely than less accessible locations. On the same path, Alonso (1964) formulated a theory linking accessibility and land use. Following Krugman (1996) and Fujita et al. (1999), a great part of spatial development can be explained by the interplay between two major driving forces, (i) economies of scale and (ii) spatial factors such as transport costs and land prices.

To take the leap from these concepts towards an urban centrality measure, we propose to use a simplified model of urban economic activity in combination with a much more detailed spatial representation. This makes it possible to view the urban system as a network of interacting locations (Barthélemy, 2011; De Montis et al., 2013; Andersson et al., 2006).

2.2 Urban activity

A central concept in this paper is the notion of urban activity (denoted a_i, for zone i). In our definition, urban activity is fundamentally tied to a location and to interactions. We do not differentiate between activity types but leave it as an aggregated intensity measure[2] corresponding to the sum of all interactions between a location and all other locations. Since it includes both social and economic interactions, it cannot be easily measured in total, which means that any modelling and empirical studies must resort to studying some relevant proxies. The monetary part of urban activity can be understood as a concept close to GDP, so that activity can be approximated by the sum of the market value of all (value-adding) production of goods and services taking place at a location at a certain point in time.

2.3 Local characteristics

A fundamental property of a location is its capacity to be adapted to human activity, determined by basic usability such as local access to buildable land and infrastructure. These local characteristics (denoted R_j) correspond to the attractivity of a zone "in itself". Details about how we have calculated the local attractivity characteristics are described in the Methods section.

2.4 Accessibility and centrality

Consider the accessibility to attractions as defined by Hansen (1959); $A_i = \sum_j W_j f(c_{ij})$, where W_j is the index of attraction of j, c_{ij} is a measure of distance or travel time of moving

[2] Different activity intensities however, do make a location more or less suited for different activity types, which means that a change of intensity sometimes goes together with a change of type. These type changes, however are assumed to be implicit in our modelling framework. This also means that we assume land improvements such as buildings are assumed to be an effect of activity – not a source of it.

between i and j, and f is a decreasing function. One way of describing centrality is by stating that a location is central if it has strong accessibility to other central locations, which can be formalised by replacing attraction W_j with accessibility A_j itself, to arrive at a recursive eigenvector centrality definition, $A_i = \sum_j A_j f(c_{ij})$.

This concept is powerful and forms the basis for the measures that we elaborate in this paper. One outcome of such a centrality concept is the famous page-rank algorithm used by Google (Brin and Page, 1998), which enables a ranking of web documents with regard to their importance. Documents on the internet are given a higher ranking if they are linked to from other pages with high ranking. Notably, at no point, the search engine has to analyse the semantic contents of the documents – which is exactly what it seeks to rank the importance of. This approach has also been applied to physical road networks by e.g. Jiang (2006) and Chin and Wen (2015), with the main objective to describe human movement. El-Geneidy and Levinson (2011) have tackled the centrality calculation from a different direction, by using data on actual flows as a starting point. Our proposed centrality measures are also based on flows of interactions, but without any requirements of specific travel data. Instead, the computations are performed by modelling these flows using a general interaction function with infrastructure network data as input (although modelling accuracy could likely be improved by using detailed empirical interaction data).

Using centrality measures based on the road network to predict urban flows and activities is not a new idea, see for example Hillier and Hanson (1989), Porta et al. (2009), Sevtsuk and Mekonnen (2012) and Gao et al. (2013). However, the measures that have been mostly in focus (closeness and betweenness centrality) cannot easily be incorporated into a spatial interaction modelling framework, which is our main reason for instead exploring extensions of eigenvector centrality.

2.5 Closing the loop from activities to flows and back again to activities

Our modelling approach departs from classical spatial interaction modelling (Wilson, 2000; Batty, 2013), where local activity levels a_i are exogenous variables, appearing as specific aspects of local activity, such as population or purchasing power. We then ask whether we may instead infer the distribution of activity from knowledge about the other variables, in particular the information embodied by infrastructure networks. The causal rationale for this belief is, first, that large-scale infrastructure change is a relatively slow process, which implies that land use, activity levels and interaction flows have enough time to adapt to a semi-static infrastructure network. Second, even to the extent that the time scales of road and land use change do overlap, actual planning practices link according to ideas of need and geographical importance, so the effect also of the reciprocal dynamics goes in the same direction.

2.5.1 From activity to spatial interaction

Spatial interaction models arise by subjecting the logic of the gravity model to local constraints on the size of flows in the system. Flows of interactions between zones can then be estimated, by distributing economic flows from origins to destinations in proportion to their relative attractions, see Figure 1. As noted by Wilson (2000) such a model formulation will take into account the competition between different locations for attracting incoming flows.

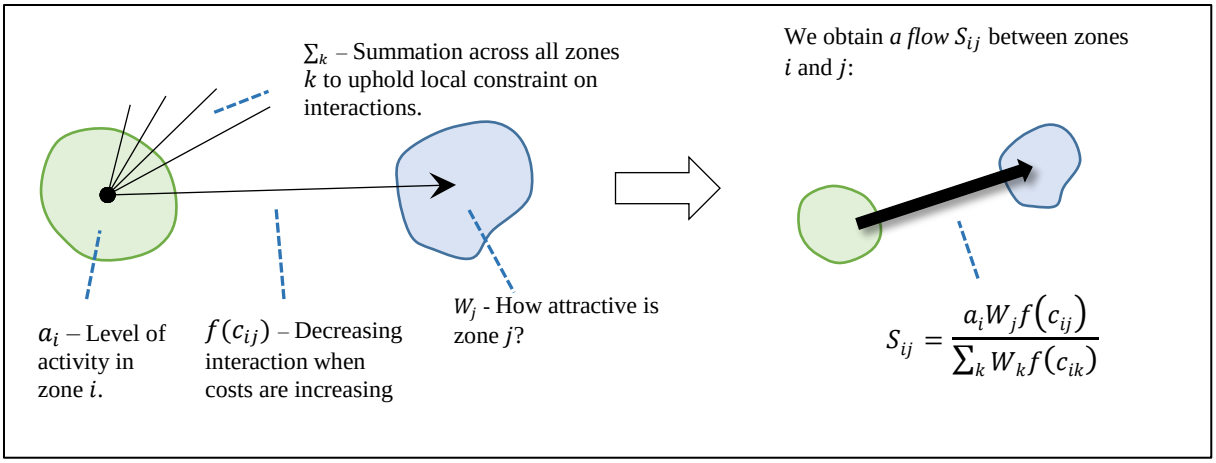


Figure 1. Deriving flows from activity and attractivity. The flow is shown as one-directed, but a flow in the opposite direction is also present and can be computed analogously. See supplemental material for a detailed derivation of the interaction model.

2.5.2 From spatial interaction back to activity

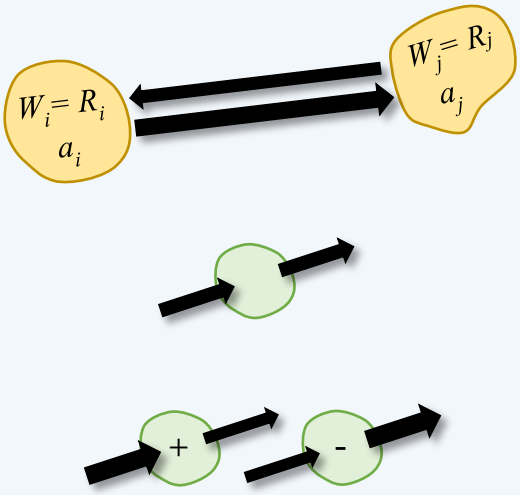
In many cases, the distribution of activities in the system is of interest in itself. Salient questions include how infrastructural change affects things like urban extent, patterns of interaction, housing, jobs and so on. Infrastructural data is considerably more widely available, complete and consistent than demographic and economic data on the nebulous concepts of activity and attraction, which we must approach via its rich flora of expressions such as buildings, land value and population. If we can tease most of the information we need out of the infrastructure of interactions, we are in a much better shape with regard to data supply but also with regard to model design. We may then circumvent the need to figure out how various sub-models interact, and we are at least less exposed to the ontological mismatch between models and reality.

In spatial interaction modeling, activity a_i represents demand while attraction W_j represents supply. Interactions are thereby directed, going from demand to supply.

In the basic eigenvector centrality model, activity a_i represents *any* activity. Most types of activity generate both supply and demand on an aggregated level. Attraction W_j is refined into an intrinsic property R_j of the zone, reflecting suitability for development.

We posit that activity is in equilibrium when total interaction from a zone is in balance with total interaction to the zone. Our task is then to find such a configuration of a_i to fulfil this for all zones.

To achieve this, we iteratively adjust a_i across the zones. If interactions in and out are not in balance, the current estimate must be adjusted. We repeat until a convergence criterion has been reached.



In our preferential centrality model, we refine the definition of attraction to reflect a dynamic coupling with activity. Development suitability R_j now figures as one aspect of attraction W_j together with activity a_j. A parameter α is used to set the balance between these aspects.

Figure 2. From spatial interaction to activity modelling.

In Figure 2 we outline the logical sequence in which we develop our preferential centrality model by using a "quasi-growth model" – *quasi* since it embodies a growth logic but is really used in an iterative process to find a stable equilibrium distribution of activity. First, we assume that activity quasi-growth is proportional to the sum of flows entering the zone. Second, attraction W_j is refined into an intrinsic property equal to our measure of local characteristics, $W_j = R_j$. Now, if we begin with activity uniformly distributed across the system, and we redistribute it according to this logic we arrive at an iterative algorithm,

$$a_j(t+1) = C\left[a_j(t) + \epsilon \sum_i S_{ij}(t)\right], \qquad (1)$$

with the equilibrium distribution

$$a_j = R_j \sum_i \frac{a_i f(c_{ij})}{\sum_k R_k f(c_{ik})}, \quad (2)$$

independently of the quasi-growth constants C and ϵ. See supplemental material for the full derivation of this self-referring equilibrium condition, that can be restated as $a_j = \sum_i a_i M_{ij}$, where $M_{ij} = \frac{R_j f(c_{ij})}{\sum_k R_k f(c_{ik})}$. The adjacency matrix M_{ij} corresponds to a transformation of the physical network and the activity will correspond to the eigenvector centrality of this weighted, transformed network. Thus, we can infer the structure of urban activity from the physical linking of places, similar to how the PageRank centrality algorithm can infer the relative importance of pages from the hyperlink structure.

The model may be substantially improved by positing that activity in itself stimulates attractivity, $W_j = a_j + \alpha R_j$, which results in a modification of the equilibrium formulation:

$$a_j = \left(a_j + \alpha R_j\right) \sum_i \frac{a_i f(c_{ij})}{\sum_k (a_k + \alpha R_k) f(c_{ik})}. \quad (3)$$

We call this new non-linear measure *preferential centrality*, because the activity-dependent attraction can be thought of as a continuous version of preferential attachment (Barabási and Albert, 1999) for the activity interaction network. The resulting equation can be solved for a_j by iteration. However, unique or positive solutions are not guaranteed for low values of α.

2.6 Interaction function

The most common choices for interaction functions are the exponential function $f_e(c_{ij}) = e^{-\beta c_{ij}}$, and power law decay, $f_p(c_{ij}) = c_{ij}^{-\beta}$. If we were studying a single type of activity it would be reasonable to assume a specific spatial scale of interaction, which is something that the exponential function captures well. However, our generalised concept of urban activity implies a mix of interactions on all scales which makes it more reasonable to use the power law function. Generally, the choice of interaction function is of course an empirical question.

3 Data

The data used for this study are of three kinds; road network, property polygons and land taxation values. The road network is used for three purposes; finding accessible areas within the polygons, finding connections from the polygons onto the road network and finally performing the distance calculations between zones. The property polygons are assigned a land taxation value from the taxations database according to a common identifier. They are thereafter aggregated into zones based on area and type code. In this study, the municipality of Gothenburg is chosen as a prototype area to develop, test and validate the model.

Roads and streets are imported with preserved topology and attributes from Open Street Map (OSM). OSM has been subject to questions about its quality, but studies have found that the data quality is on pair with other data sources (Haklay, 2010; Dhanani et al., 2012). The reasons for choosing OSM are several; it is readily available to download, it contains the necessary attributes for the calculation, it has worldwide coverage for future expansions of the model, and the data is open.

The entire extent of Sweden is partitioned into “properties”. Properties are either owned by individuals or juridical entities, or they can be jointly owned in the form of associations. The

precision and quality of this data is high, since the purpose is to establish and prove ownership (which needs to be precise and just). Properties are of different types and usages; therefore, they are classified and assigned a type code based on usage by the Swedish taxation authority. The extent and borders of these properties are obtained from the Swedish land survey.

The Swedish taxation authority assigns to all properties a taxation value that should represent about 75% of the market value. This value is arrived at by a procedure that takes several characteristics into consideration such as area, closeness to water, building type, sales values of the neighbouring properties etc. The quality of this data is also very good in the sense that it is done according to a legal criterion, although the values for industries is a bit uncertain due to the fact that they are seldom sold. Therefore, these few sales have a disproportionately big impact on the industrial properties taxation values. This has to be taken account for in the regression analysis. All the taxation values and type codes are acquired from the Swedish taxation authority.

4 Methods

The procedure for model exploration and validation is roughly composed of three steps; 1) data preparation in order to create the input for the activity model as well as preparing the empirical data used in the last step, 2) running the activity model and 3) finally using the results from the models in a multiple spatial regression analysis with the empirical values.

For the activity model we compare four different versions; the local model, the monocentric model, the iterative eigenvector model and the iterative preferential model. Our aim is to assess whether or not the more elaborate iterative models provide any additional predictive capabilities compared to the simpler versions. To find out whether the models are capable of capturing all of the spatial dependencies, we have performed spatial testing (Anselin, 1988) in the regression analysis.

4.1 Data preparation

4.1.1 Spatial entities

The spatial entities used in the activity model and the multiple regression analysis are chosen to be realized as zones, defined as one or more aggregated properties. All properties smaller than 3000 m^2 are aggregated to zones by dissolving common borders, if they have the same taxation type code.

Geographical analysis of polygon features are subject to the MAUP (Openshaw and Taylor, 1979). The way of spatial partitioning of land must therefore be carefully chosen. The justifications for using zones as spatial units are that; properties are readily available, have a designated usage and can provide useful output in planning applications. Property-based zones also simplifies the empirical comparisons, since model and data will have the same spatial representation.

4.1.2 Connection between road network and zones

We do not use detailed data about physical connections between zones and the road nework. Instead approximate “virtual” connections are created in the road network model by choosing the shortest Euclidean lines between zonal centroids and connection-permissible roads.

Motorways, trunk roads and other roads with high speed limits are not considered permissible for these virtual connections.

4.1.3 Zonal weights – local characteristics

A zonal weight (R_i) is assigned to every zone i based by accessible, buildable and permitted areas. Generally, the weight can also be modified with different types of (physical) attractivity factors.

Accessible areas are here stipulated as land that can be accessed from roads. Therefore, the assumption in the model is that only the area within a certain distance from a road is possible to develop. These areas are created by buffering the roads (30 m in the baseline case) and doing a union overlay onto the properties.

Buildable areas are hereby defined as firm ground suitable for buildings. Areas used by (or very close to) road or rail infrastructure are not considered as buildable.

Permitted areas are those that, according to planning restrictions, are allowed for development. In our current model implementation, productive forestry, agricultural land and areas used for special purpose buildings are considered as not permitted.

A basic attractivity factor is closeness to open water, which can have a large effect on land value and land taxation. Since our study area (Gothenburg) is situated by the coast we must include some approximation for this effect. We have chosen to include the water attraction as a multiplicative factor of 1.5 for the zonal weights for zones with centroids within 500 m of the coast-line.

4.2 Implementation of the activity model

To arrive at zone-to-zone impedances c_{ij}, Dijkstra's algorithm is used to identify the shortest paths in the road network weighted by segment travel times (taking into account speed limits). A constant impedance penalty (comparable to 1 minute in the baseline case) is added to all relations to reflect the cost of starting and ending an interaction. Zones are assumed to not interact with themselves, i.e. $f(c_{ii}) = 0$. As a baseline interaction function we have used the power law decay, $f(c_{ij}) = c_{ij}^{-\beta}$, with $\beta = 2$.

The eigenvector activity model is implemented by using simple iterative updating of the activity for all zones. Initial activity is chosen to equal local zonal weights, i.e. $a_i(t = 0) = R_i$. Zonal weights are then considered static during the iteration. For every iteration a new activity vector is computed using Equation (1). Total activity is kept constant in every iteration by a global normalisation. The relative vector norm of activity differences between subsequent iterations is compared to a predefined tolerance value (we have used 10^{-5}), to determine if a good enough approximation to the equilibrium is found.

The implementation of the preferential model is identical to the eigenvector model in all aspects except from the additional mechanism of activity dependent attractivity. This mechanism introduces the parameter α, for which we have chosen a value as low as possible, but that still results in a convergent iterative process. This principle gives the largest possible difference of activity configuration in comparison to the eigenvector model, since increasing values of α can

bring the results of the preferential model arbitrarily close to the eigenvector model. In the baseline case, the application of the principle resulted in $\alpha = 1.625$.

Compared to the iterative models, *the monocentric version* is simpler. It is derived by assuming that all zones only interact with the most central zone, defined in the implementation as the zone closest to Gothenburg Central Station. For a full description of this model version, see supplemental material.

Zonal weights are mainly used as input to the iterative activity models. However, for comparative purposes we also investigate a *local activity model*, without any interaction between zones. It is implemented using direct proportionality between zonal weights and activity.

4.3 Spatial regression

4.3.1 Preparation of the spatial regression analysis data

The two independent variables are; the prediction from the activity model and the amount of industrial area per zone. The reason to include the amount of industrial area in the regression model is that industrial properties have on average a lower taxation value due to the taxation process.

The dependent variable is the property taxation value. For some records in the taxation database there is not a 1:1 relationship to property polygons. We handle this by aggregation, de-aggregation and filtering. We start from 60137 property polygons and arrive at 27628 zones after aggregation. Out of these, we have empirical taxation values for 12062 zones, hence only they are used in the regression.

4.3.2 Weight matrix creation.

In order to specify a regression model with spatial diagnostics a spatial weights matrix has to be created. The weights matrix in this study is created by using the impedance of the road network between all places and then apply a cut-off value in order to determine which zones as treated as adjacent ones. We have chosen a cut-off value that is 3000 meters. To examine the robustness of the model a weight matrix based on Euclidian distance of 600 meters is also tested in the regression.

4.3.3 Investigating spatial dependencies

To examine the presence of spatial dependence, an analysis of Moran's I for the model values and empirical values is made (Moran, 1950; Haining, 2004). This test (see Table 1) shows that both preferential model values and taxation values are subject to a rather strong spatial autocorrelation while the local weights are not.

Variable	**Moran's I**
Land taxation value (dependent)	0.34
Local weights (independent)	0.04
Preferential model prediction (independent)	0.47
Industrial areas (independent) used as correction factor	0.24

Table 1. Indicators for spatial autocorrelation.

This finding indicates that spatial diagnostics needs to be evaluated in the regression analysis, to make sure that all spatial autocorrelation is taken care of. The finding that local weights are virtually not at all spatially autocorrelated tells us that they cannot sufficiently explain the variation in the empirical property taxation values.

4.3.4 Ordinary Least Squares (OLS) with spatial diagnostics

An OLS with both spatial and non-spatial diagnostics is performed in order to know whether the dependent variable's spatial autocorrelation is captured by the independent variables (which would mean that an ordinary OLS is sufficient). If not, the diagnostics are used as guidance for the next steps in order to take care of the spatial autocorrelation (Anselin, 1988). This results in a collection of diagnostics that need to be analysed:

- Diagnosis for non-normal error distribution, Jaque-Bera (JB) test.
- Diagnostics for heteroscedasticity, Breusch-Pagan and Koenker-Bassett tests (B-P and K-B).
- Diagnostics for spatial autocorrelation, Lagrange Multipliers (LM) tests and Moran's I on the residuals.

4.3.5 Comparative indicators for model fitness and validity

To evaluate and compare models, R^2 is commonly used but is not reliable when residual spatial autocorrelation is present. Therefore, the Schwarz information criterion is also used (Anselin and Rey, 2014).

When spatial autocorrelation is present in the residuals, the observations are not independent from each other, hence the regression model is not valid. This is investigated with the LM tests; if they are significant it indicates that some measure like using a spatial lag or spatial error model has to be taken in order to handle the remaining spatial autocorrelation (Anselin, 1988). If the LM (or robust LM) test for spatial error model is significant while the tests for lag model are not, a spatial error model is probably the right way to go, and vice versa. If both tests are significant, the regression analysis is not valid and there is no indication of any spatial model that can make it valid. In that case the model has to be re-specified (Anselin and Rey, 2014). This procedure has been used in this study for guidance in the search for a good and valid model.

4.4 Software

For the data preparation, cleaning and aggregation, FME was used. The activity models were implemented in python, using the packages OSMnx (Boeing, 2017) and NetworkX (Hagberg et al., 2008). The spatial statistical analysis were performed in GeoDa (Anselin, 2006).

5 Results

5.1 Model validity and fitness

All models except the preferential models have all the LM tests significant, which invalidates them due to untreated spatial autocorrelation. The local and industrial models are included just as control, to see that it is actually the activity model prediction that is responsible for the good results. The other indicators on model fitness shown in Table 2 implies that the preferential model is the best choice, even before considering and applying the spatial error model.

For the preferential model, the robust version of the LM test for error model was significant (0.00) while the robust version of the LM test for lag model was not (0.83). This suggested that using a spatial error model is the correct approach (Anselin and Rey, 2014). Therefore, only the preferential spatial error model is usable for inference and predictions, although its spatially clustered errors (Anselin, 1995) are hiding some unknown spatial factors (see Figure 3).

Model version	**R^2**	**Morans' I on residuals**	**Schwarz information criterion**	**Model valid?**
Industrial area coverage (as control)	0.00	0.34	20842	No, since all LM tests are significant.
Local	0.40	0.42	14644	No, since all LM tests are significant.
Monocentric	0.54	0.24	11329	No, since all LM tests are significant.
Eigenvector	0.54	0.24	11470	No, since all LM tests are significant.
Preferential	0.58	0.16	10297	No, not as non-spatial OLS, since LM tests are significant.
Preferential spatial error model	**(Pseudo)** 0.66	Not applicable (none)	7792	Yes, since remaining spatial autocorrelation is taken care of as error term

Table 2. Results from the spatial regression. A better fit is indicated by a lower Schwarz and a higher R^2. For Morans' I, low values indicate low spatial autocorrelation. The pseudo R^2 value in a spatial error model is computed differently than in a standard OLS, which means that the R^2 for the preferential spatial error model is not directly comparable to the other R^2 values in the table.

5.2 Other statistical tests on the preferential spatial error model

The low multicollinearity number (12) indicates that there is no problematic multicollinearity among the explanatory variables. Values < 30 are usually considered as unproblematic (Anselin and Rey, 2014)

The JB test is significant, which indicates a non-normal distribution of error terms. However, this test is less relevant, since this dataset is large (Anselin and Rey, 2014).

According to the B-P and K-B tests there is a significant heteroskedasticity in the model results. There can be multiple reasons for this where one possible cause is the aggregation of properties (Haining, 2004). The effects are not that great in these specific models, since the standard errors are very low on their own. It is therefore not considered as crucial for the conclusions of this study.

5.3 Sensitivity analysis

We have explored many variations of the key parameters, such as the preferentiality parameter α, and the functional form and parameters of the interaction function. See supplemental material for details on these results. The main finding is that the preferential model seems to be robust with regard to changes in parameter values

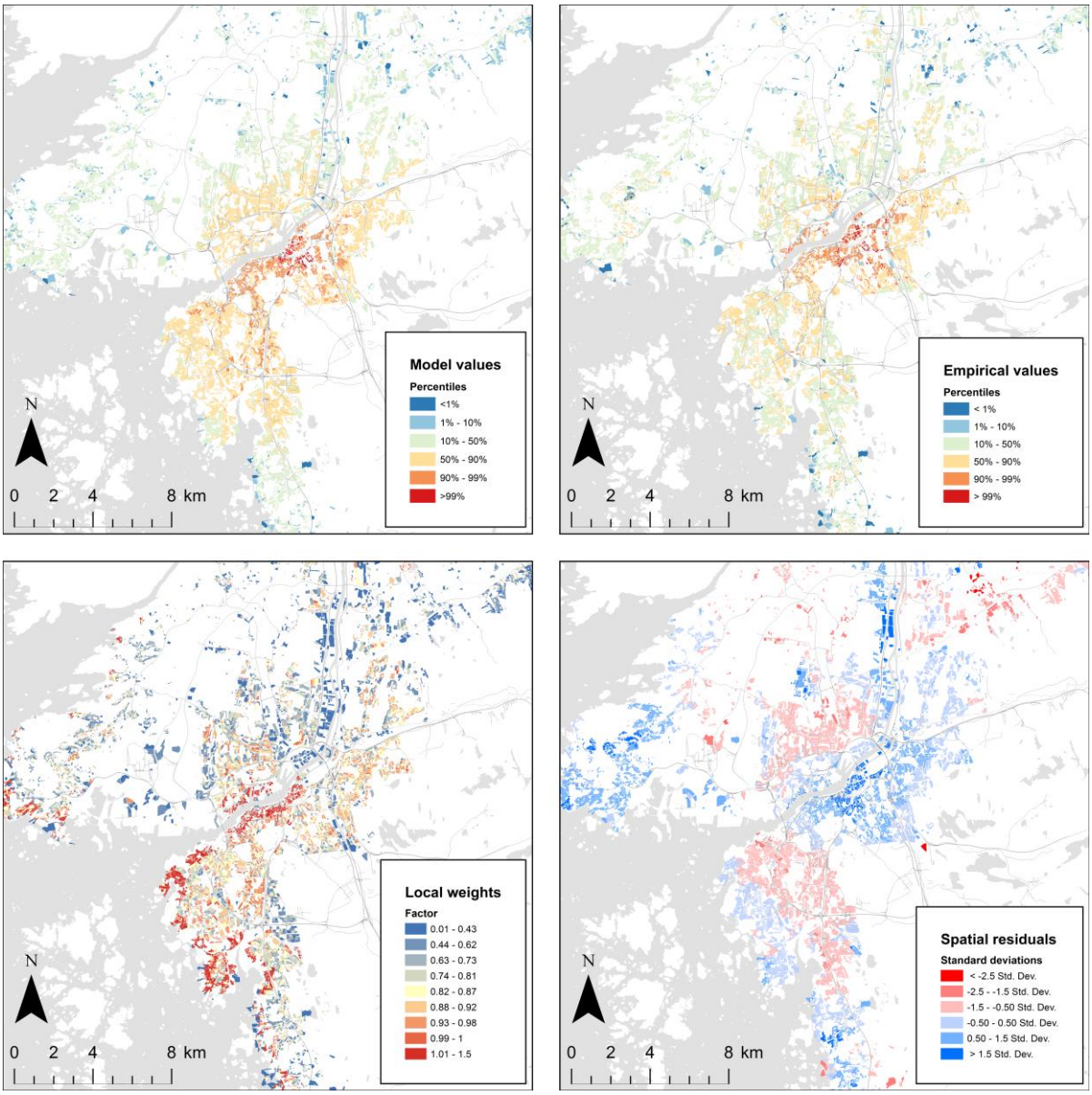


Figure 3. Preferential spatial error model: Predictions (top left), empirical land value (top right) and local weights (bottom left) are normalised with regard to zone area. Spatial residuals (bottom right) show the remaining spatially autocorrelated error term.

5.4 Discussion of results

5.4.1 Comparing the model versions

The eigenvector and monocentric models have decent performance; therefore, the interpretation of their results have been used as steps in the search for a valid model. The preferential spatial error model, besides being the only valid model, also performs well in absolute numbers with a pseudo $R^2 = 0.66$. Considering the small number of input data sources used, and the simple underpinning model assumptions, this level of correlation indicates that the proposed preferential centrality measure is promising.

5.4.2 Remaining challenges

In this paper we have not aimed to present a full predictive model. Some improvements for moving in that direction are:

- To reduce uncertainty in the regression coefficients, heteroskedasticity should be sufficiently taken care of. Some more parameter variations as well as trying different levels of aggregation into zones might give some clues how to handle this problem.
- The preferential spatial error model still contains unknown spatial variables that are handled as a spatial error term together with standard residuals. To understand those errors can be helpful for further development of the model. Some ideas and suggestions for further investigation are:
    - Different kinds of properties (i.e. commercial vs. residential) might not be fully comparable in taxation terms.
    - Other transportation modes, such as pedestrian, bicycle and public transport are not captured in the current car-oriented implementation of the model
    - Truncation effects; this model is only investigating areas within the Gothenburg municipality, although the city also acts a regional centre for a larger surrounding region.
- In the preferential model, we have a parameter α for which model fitness improves as it is lowered towards the threshold of iterative divergence. Perhaps the empirical system state corresponds to a non-convergent model outcome? To explore this hypothesis, the convergence criterion in the model can be replaced by a minimisation target.

6 Conclusions and ways forward

By using a theoretical concept of interaction-based centrality we have demonstrated that it is possible to create an urban activity model with empirical validity, using only two data sources – road networks and property polygons. The empirical validation is based upon using land taxation values as a proxy for urban activity.

According to the comparative results from the spatial regression, local characteristics are far from enough to explain the geographical variation of land values. The activity intensity is also affected by the geographical ranking of the location; in the city and in the region. Including the distance to the city centre in a monocentric interaction model gives a seemingly better fit, but the spatial statistical tests shows this model to be invalid for the geographical area that we study, indicating that a more elaborate model is warranted. With the introduction of our concept of preferential centrality, where initial concentrations of activity are assumed to ignite local feedback-mechanisms that attract even more activity, we finally arrive at a valid regression model.

The preferential centrality model has several additional advantages compared to a monocentric approach. First, we avoid the requirement of having to manually identify the most central location. Instead the centrality model will endogenously determine central places and their relative importance. In a polycentric setting this is a crucial model feature. Second, in a planning context it can often be an important question in itself how the location and strength of urban centres are affected by planning interventions, such as new infrastructure. For example, the

preferential model can be used to analyse the robustness of a city centre under the influence of suggested new road investments. Such an analysis is clearly not possible within a monocentric model framework.

Regarding data requirements, our approach is somewhat more demanding when compared to a basic monocentric model, since travel times must be computed between all zones and not only to the predefined centre. The number of zones needed (i.e. the spatial resolution) depends on context and further studies are needed to determine what levels of resolution that are adequate for different planning applications.

Our current model implementation is technically complicated and requires different pieces of software. This is however not a fundamental property of the approach and we aim in future work to achieve a work-flow within a single open source framework, to open up for broader testing and practical application.

Before using our modelling approach in a practical planning context, further validation is needed; both cross-sectional by studying other and larger areas, and longitudinal by investigating changes in urban activity over a time period where the road network also has changed. For the purpose of this validation, we cannot escape the need to use empirical activity data, such as taxation values or night light data. However, since our sensitivity analyses show that model outcomes are fairly robust, a validated preferential centrality model should be transferrable to applications in different geographical settings, without any need for local economic or demographic data.

7 Funding

This study was funded by the Norwegian Public Roads Administration, FORMAS - the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (grant number 2015-124), and the Swedish Transport Administration.

8 Acknowledgements

The paper has benefitted from discussions and support in the context of the Spatial Morphology Group at Chalmers University of Technology.

9 References

Alonso W. (1964) *Location and land use*: Harvard University Press.

Andersson C, Frenken K and Hellervik A. (2006) A complex network approach to urban growth. *Environment and Planning A* 38.

Andersson C, Törnberg A and Törnberg P. (2014) Societal systems – Complex or worse? *Futures* 63: 145-157.

Anselin L. (1988) Lagrange Multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis* 20: 1-17.

Anselin L. (1995) Local indicators of spatial association - LISA. *Geographical Analysis* 27: 93-115.

Anselin L, Ibnu Syabri and Youngihn Kho. (2006) GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis 38 (1).*

Anselin L and Rey S. (2014) *Modern Spatial Econometrics in Practice,* Chicago: GeoDa Press LLC.

Barabási A-L and Albert R. (1999) Emergence of scaling in random networks. *Science* 286: 509-512.

Barthélemy M. (2011) Spatial networks. *Physics Reports* 499: 1-101.
Batty M. (2013) *The New Science of Cities*: MIT Press.
Boeing G. (2017) OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* 65: 126-139.
Brin S and Page L. (1998) The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30: 107-117.
Chin WC and Wen TH. (2015) Geographically Modified PageRank Algorithms: Identifying the Spatial Concentration of Human Movement in a Geospatial Network. *PLoS One* 10: e0139509.
De Montis A, Caschili S and Chessa A. (2013) Recent Developments of Complex Network Analysis in Spatial Planning. In: Scherngell T (ed) *The Geography of Networks and R&D Collaborations.* Cham: Springer International Publishing, 29-47.
Dhanani A, Vaughan L, Ellul C, et al. (2012) From the axial line to the walked line: evaluating the utility of commercial and user-generated street network datasets in space syntax analysis. In: M. Greene JRaAC (ed) *Eighth International Space Syntax Symposium.* Santiago de Chile: Proceedings: Eighth International Space Syntax Symposium.
El-Geneidy A and Levinson D. (2011) Place Rank: Valuing Spatial Interactions. *Networks and Spatial Economics* 11: 643-659.
Fujita M, Krugman PR, Venables AJ, et al. (1999) *The spatial economy: cities, regions and international trade*: Wiley Online Library.
Gao S, Wang Y, Gao Y, et al. (2013) Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. *Environment and Planning B: Planning and Design* 40: 135-153.
Hagberg A, Swart P and Schult D. (2008) Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab, Los Alamos, NM (United States).
Haining R. (2004) *Spatial Data Analysis: Theory and Practice*, Cambridge university press.
Haklay M. (2010) How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B-Planning & Design* 37: 682-703.
Hansen WG. (1959) How Accessibility Shapes Land Use. *Journal of the American Institute of Planners* 25: 73-76.
Hillier B and Hanson J. (1989) *The social logic of space*: Cambridge university press.
Iacono M, Levinson D and El-Geneidy A. (2008) Models of transportation and land use change: a guide to the territory. *Journal of Planning Literature* 22: 323-340.
Jiang B. (2006) Ranking Spaces for Predicting Human Movement in an Urban Environment. *International Journal of Geographical Information Science* 23: 823-837.
Krugman P. (1996) Urban concentration: The role of increasing returns and transport costs. *International Regional Science Review* 19: 5-30.
Moran PAP. (1950) Notes on Continuous Stochastic Phenomena. *Biometrika* 37.
Newman ME. (2008) The mathematics of networks. *The new palgrave encyclopedia of economics* 2: 1-12.
Openshaw S and Taylor PJ. (1979) A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley N (ed) *Statistical applications in spatial sciences.* London: Pion, pp. 127–144.
Porta S, Strano E, Iacoviello V, et al. (2009) Street centrality and densities of retail and services in Bologna, Italy. *Environment and Planning B: Planning and Design* 36: 450-465.

Sevtsuk A and Mekonnen M. (2012) Urban network analysis. *Revue internationale de géomatique–n* 287: 305.

Simon HA. (1962) The Architecture of Complexity. *Proceedings of the American Philosophical Society* 106: 467-482.

Thünen Jv. (1826) Der isolierte Staat. *Beziehung auf Landwirtschaft und Nationalökonomie*.

Timmermans H. (2003) The saga of integrated land use-transport modeling: how many more dreams before we wake up? *Keynote paper, Moving through nets: The Physical and social dimension of travel, 10th International Conference on Travel Behaviour Research, Lucerna.*

Waddell P, Borning A, Noth M, et al. (2003) Microsimulation of urban development and location choices: Design and implementation of UrbanSim. *Networks and Spatial Economics* 3: 43-67.

Wilson AG. (2000) *Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis*: Prentice Hall.

Supplemental material A – mathematical detail

From activity to spatial interaction

We have an interaction model, for the economic flow S_{ij} from zone i to j:

$$S_{ij} = B_i P_i W_j f(c_{ij}),$$

where B_i is a balancing factor (to be determined below) for zone i, P_i is the economic output (total spending) from zone i, W_j is the attraction of zone j, c_{ij} is the cost of interaction (cost can be derived from network impedance) between i and j, f is a function decreasing with increasing cost.

We also have the accounting relation that the sum of flows from a zone should correspond to a_i, the economic activity in the zone:

$$\sum_k S_{ik} = a_i.$$

The accounting relation can be combined with the initial flow equation

$$B_i P_i \sum_k W_k f(c_{ik}) = a_i$$

to determine the balancing factor

$$B_i = \frac{a_i}{P_i \sum_k W_k f(c_{ik})}.$$

Thus, the resulting interaction flow model is

$$S_{ij} = \frac{a_i W_j f(c_{ij})}{\sum_k W_k f(c_{ik})},$$

which has the simple interpretation that the total economic flow from zone i is distributed between all other zones in proportion to their relative attractions modified by a function of interaction cost.

The total incoming economic flow D_j for zone j can now be found by

$$D_j = \sum_i S_{ij} = W_j \sum_i \frac{a_i f(c_{ij})}{\sum_k W_k f(c_{ik})}$$

As noted by Wilson (2000), this will take into account the competition between different locations for acquiring incoming flows.

From spatial interaction back to activity

To be able to infer activity levels from only attractions and costs of interaction, we use a "quasi-growth model" – "quasi" since it embodies a growth logic, but is really used in an iterative process to find a stable equilibrium distribution. The local "quasi"-growth of activity at a zone is assumed to be proportional to the sum of incoming flows, according to

$$a_j(t+1) = C(a_j(t) + \epsilon D_j)$$

Where C is a global factor controlling overall growth.

Assuming the condition of constant global activity and the relation $\sum_k D_k = \sum_k \sum_i S_{ik} = \sum_i a_i$ gives

$\sum_j a_j(t) = \sum_j a_j(t+1) = C \sum_j a_j(t) + C\epsilon \sum_j D_j(t) = C(1+\epsilon) \sum_j a_j(t).$

For this to hold true, $C = \frac{1}{1+\epsilon}$, and $a_j(t+1) = \frac{a_j(t)+\epsilon D_j}{1+\epsilon}$

The equilibrium condition $a_j(t+1) = a_j(t)$ yields the only solution $a_j = D_j$, i.e. that incoming flow must be equal to activity.

And then we can state the equilibrium condition

$$a_j = W_j \sum_i \frac{a_i f(c_{ij})}{\sum_k W_k f(c_{ik})}$$

or in a briefer version:

$$\frac{a_j}{W_j} = \sum_i f(c_{ij}) \frac{a_i}{A_i},$$

where $A_i = \sum_k W_k f(c_{ik})$ is the accessibility to attractivity from zone i. Thus, we can interpret the equilibrium condition as a system state where the ratio of activity and attraction must equal the accessibility to normalised activity, where the normalisation is with regard to accessibility. This means that for a zone to have high activity relative to its attractivity, it must have high accessibility to other zones which themselves have low accessibility to attractivity.

The right-hand term can be thought of as a relative accessibility, or spatial fitness $\eta_i = \sum_i f(c_{ij}) \frac{a_i}{A_i}$. This term captures everything related to the spatial propensity of a location for attracting new activity. The left-hand term contains only localised variables. This means that in equilibrium, the local activity and attractivity of a zone must be in balance with the zone's relative place in the spatial system, described by the right-hand term. In short, $\frac{a_j}{W_j} = \eta_i$.

To find specific solutions, an additional model component is needed, to describe how attractions develop. Two obvious alternatives are:

1. Describe static attractions based on specific data of the studied system.

2. Create a dynamic economic model for the evolution of different types of attractions.

Since our aim is to achieve a simple model, with minimal dependence of data and economic specifics of different types of businesses, we have chosen a somewhat different approach.

Our first attraction-model is akin to suggestion 1 above in that we only consider local, static properties of zones. If attractions are only dependent on constant local characteristics, $W_j = R_j$ (see main paper for an explanation of how these are determined), an eigenvector equation is obtained:

$$a_j = \sum_i \frac{a_i R_j f(c_{ij})}{\sum_k R_k f(c_{ik})} = \sum_i a_i M_{ij} \,,$$

where $M_{ij} = \frac{R_j f(c_{ij})}{\sum_k R_k f(c_{ik})}$. Note that M_{ij} is only determined by local characteristics and the impedance structure of the underlying transportation network. $S_{ij} = a_i M_{ij}$, which means that the matrix described by M_{ij} reveals the relative flow from i to j, and that the equilibrium activity corresponds to the eigenvector centrality (Bonacich, 1972) of a weighted network of relative flows between zones.

Our second model version incorporate dynamics by using the simple assumption that attraction is linearly related to generalized urban activity, $W_j = a_j + \alpha R_j$, which results in the modified equilibrium formulation:

$$a_j = (a_j + \alpha R_j) \sum_i \frac{a_i f(c_{ij})}{\sum_k (a_k + \alpha R_k) f(c_{ik})}$$

More elaborate functions g of attraction $W_j = g(a_j)$, are of course also conceivable within the same formalism, but we have not yet further investigated this.

Starting from a standard interaction model we have now arrived at a self-consistent non-linear centrality measure. We call this new measure preferential centrality, because the activity-dependent attraction can be thought of as a continuous version of preferential attachment (Barabási and Albert, 1999) for the activity interaction network.

The resulting equation can be solved for a_j by iteration. However, positive solutions are not guaranteed for low values of α. At the limit of large α, the preferential centrality corresponds to the eigenvector centrality, since

$$(a_j + \alpha R_j) \sum_i \frac{a_i f(c_{ij})}{\sum_k (a_k + \alpha R_k) f(c_{ik})} = \left(\frac{a_j}{\alpha} + R_j\right) \sum_i \frac{a_i f(c_{ij})}{\sum_k \left(\frac{a_j}{\alpha} + R_j\right) f(c_{ik})}$$

$$\rightarrow \sum_i \frac{a_i R_j f(c_{ij})}{\sum_k R_k f(c_{ik})}, \text{ when } \alpha \rightarrow \infty.$$

We can summarise some observations about the equilibrium condition in the preferential model, which must hold true for every zone:

- The sum of outgoing interactions is equal to activity

- The sum of incoming interactions is equal to activity
- Attraction is linearly related to activity
- Relative accessibility (fitness) equals the ratio between activity and attraction.

Derivation of the monocentric model

For comparative purposes the same formalism can be used for creating a monocentric model,

$$a_j = C_{mono} R_j f(c_{0j}),$$

by making the assumption that all zones only interact with the most central zone, that we name zone 0. I.e. $f(c_{ij})$=0, for $i \neq 0$. C_{mono} is a global constant regulating the total activity in the system.

The monocentric model can be derived from the eigenvector equation:

$$a_{j\neq 0} = \sum_i \frac{a_i R_j f(c_{ij})}{\sum_k R_k f(c_{ik})} = \frac{a_0 R_j f(c_{0j})}{\sum_k R_k f(c_{0k})} = \frac{a_0 R_j f(c_{0j})}{\sum_k R_k f(c_{0k})} = C_{mono} R_j f(c_{0j}),$$

with $C_{mono} = \frac{a_0}{\sum_k R_k f(c_{0k})}$.

The value of a_0 will be directly related to the total activity a_{tot} according to:

$$a_{tot} = \sum_j a_j = a_0 + \sum_{j\neq 0} a_j = a_0 + a_0 \frac{\sum_{j\neq 0} R_j f(c_{0j})}{\sum_k R_k f(c_{0k})} = a_0 + a_0 \frac{\sum_j R_j f(c_{0j})}{\sum_k R_k f(c_{0k})} = 2a_0,$$

if we also make the assumption of no self-interaction within the central zone, $f(c_{00})$=0.

In its formulation the monocentric model only embodies information about the cost of travel to the city centre in combination with local characteristics. This means that relative activity a_j/R_j must decrease monotonously with increasing cost of travel to the centre.

One straight-forward interpretation of the monocentric model is that the periphery provides services (such as housing/labour) exclusively to the central zone, where all other production takes place, as well as all commercial activity. Increasing cost of interaction with the centre will make fewer services profitable and as an effect activity will decrease as we move further into the periphery.

References

Barabási A-L and Albert R. (1999) Emergence of scaling in random networks. *science* 286: 509-512.

Bonacich P. (1972) Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology* 2: 113-120.

Wilson AG. (2000) Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis.

Supplemental material B – Sensitivity analyses

Parameters

Table 3 shows an overview of the implementation-specific parameters. For all of these parameters a single typical value has been used for all baseline results reported in the main paper.

Parameter	**Unit**	**Typical value**	**Theoretical effect of change**
Waterfront factor	None	1.5	A higher value gives zones with centroids within the distance cut-off increased baseline attractivity (R_j).
Waterfront distance cut-off	Meters	500	A higher value will cause the waterfront factor to be applied to more zones.
Constant impedance penalty	Meters	1000	A higher value reduces the relative difference of interactions between nearby zones.
Buffer distance for accessible land	Meters	30	Higher local weights due to larger percentage usable for development.
Iteration break tolerance	None	10^{-5}	Convergence criterion, should be between 0 and 1.
Interaction function, $\mathrm{f}(\mathrm{c_{ij}})$	None	$c_{ij}^{-\beta}$	A change toward a more strongly decaying function reduces the interaction between farther zones.
Travel time decay exponent, β	None	2.0	An increase corresponds to a relative shift from long-range towards more local interactions.
Local characteristics weight, α	None	1.625	An increase corresponds to a smaller effect of activity on attraction, which means lower centralisation.

Table 3. An overview of the model parameters. The typical values correspond to the baseline case.

Method and results

For all sensitivity tests, one free parameter at a time is varied in combination with a changed value of α that is chosen according to same principle as in the baseline case: as low α value as possible that still results in a convergent iteration (according to the iteration break tolerance, that is held constant). All other methodology is the same as described in the main paper. Results are shown in Table 4 for the preferential activity model only. We have not used any spatial error models for the sensitivity tests.

The main findings are that the preferential model seems to be robust to changes in parameter values. Some cases such as extremely strong exponential distance decay seem to make the model perform badly, but all other variations are in general valid (according to LM tests) and not too far from the baseline, in terms of model fitness comparisons.

Kind of sensitivity test	α	R^2	**Schwarz infor-mation criterion**	**Moran's I on residuals, Euclidian distance weights matrix**	**Moran's I on residuals, network distance weights matrix**	**LM-test lag model, Euclidian distance weights matrix**	**LM-test lag model, network distance weights matrix**	**K-B**
Baseline case	1.625	0.58	10297	0.24	0.16	0.5892	0.8262	0.00
α = 2.0	2.0	0.57	10606	0.26	0.19	0.0006	0.0000	0.00
β = 1.0	0.602	0.59	10056	0.24	0.16	0.2422	0.0000	0.00
β = 1.5	1.055	0.59	10116	0.24	0.16	0.0003	0.0003	0.00
β = 2.5	2.280	0.58	10429	0.24	0.16	0.0968	0.0538	0.00
β = 3.0	2.988	0.57	10675	0.24	0.16	0.0174	0.0439	0.00
β = 4.0	2.988	0.54	11439	0.24	0.16	0.7781	0.0026	0.00
Exponential interaction function; $f(c_{ij}) = e^{-\beta c_{ij}}$, β = 0.001	3.586	0.53	11763	0.25	0.18	0.0456	0.2209	0.00
Exponential interaction function; $f(c_{ij}) = e^{-\beta c_{ij}}$, β = 0.0001	0.383	0.59	10019	0.23	0.16	0.0000	0.0000	0.00
Exponential interaction function; $f(c_{ij}) = e^{-\beta c_{ij}}$, β = 0.002	3.717	0.37	15193	0.38	0.31	0.0636	0.0000	0.07
Zonal self-interaction turned on	1.578	0.58	10257	0.24	0.17	0.3369	0.4862	0.00
Waterfront factor 1.0	1.574	0.57	10724	0.27	0.18	0.2265	0.1260	0.00
Waterfront factor 2.0	1.695	0.58	10302	0.24	0.16	0.3434	0.1569	0.00
Waterfront factor 3.0	1.844	0.56	10864	0.25	0.17	0.9156	0.0091	0.00
Constant impedance penalty 1 m	2.593	0.56	10691	0.25	0.17	0.0021	0.0000	0.00
Constant impedance penalty 5000 m	0.800	0.59	10053	0.24	0.16	0.7864	0.4622	0.00

Table 4. Sensitivity tests for the preferential activity model, with different parameter variations compared to the baseline case.

Supplemental material C – Maps

Introduction

This document contains detailed maps with local weights (*Figure 4*), baseline results for the different activity model versions, that are presented in the main paper (*Figure 5* to *Figure 7*), as well as empirical values (*Figure 8*) spatial (*Figure 9*) and non-spatial errors (*Figure 10*). All local weights, model values and empirical values have been normalised with regard to zone area. In *Figure 11* we show a close-up of the city centre to illustrate the zonal representation.

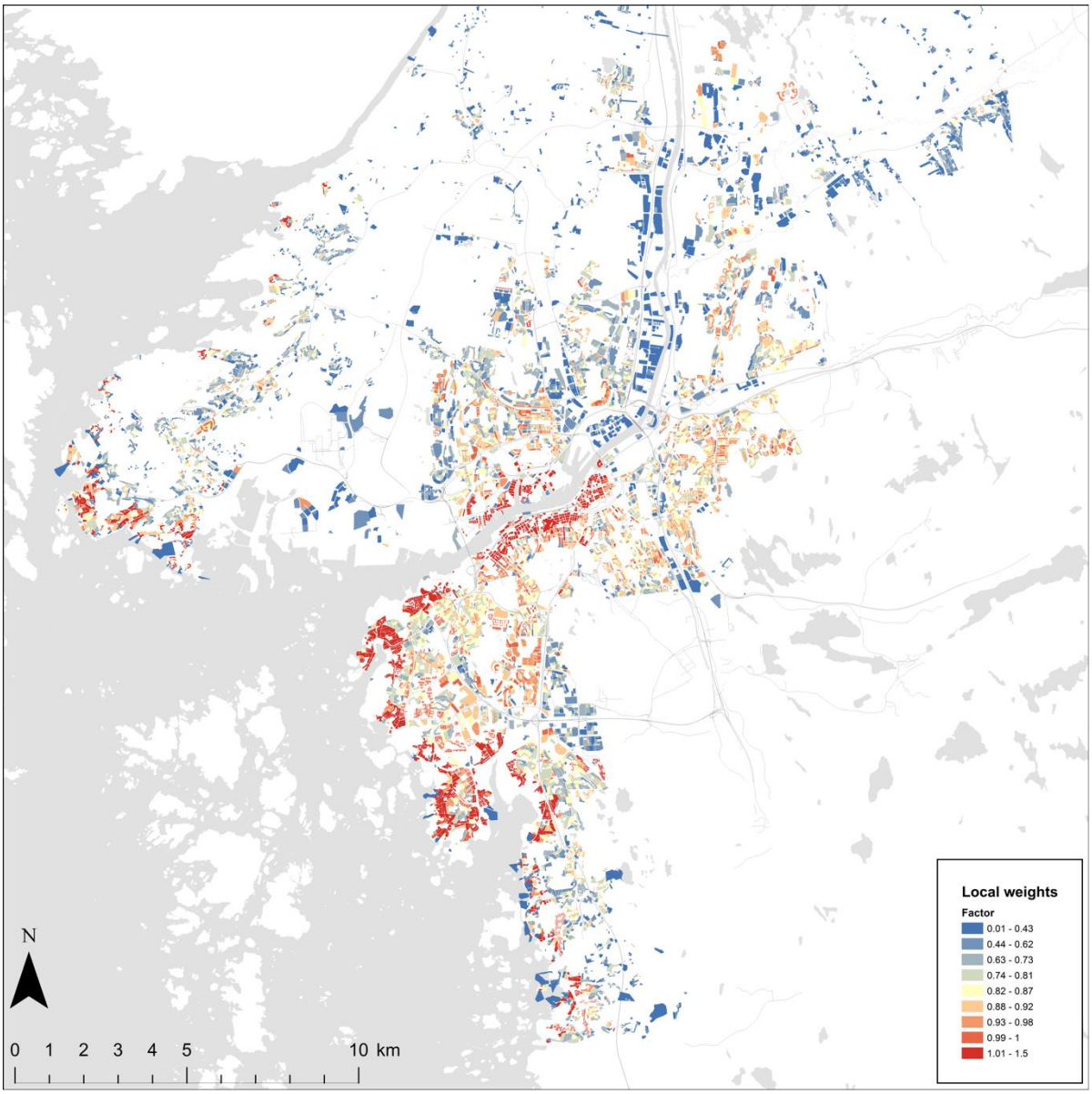


Figure 4. The local weights that are used as a starting point.

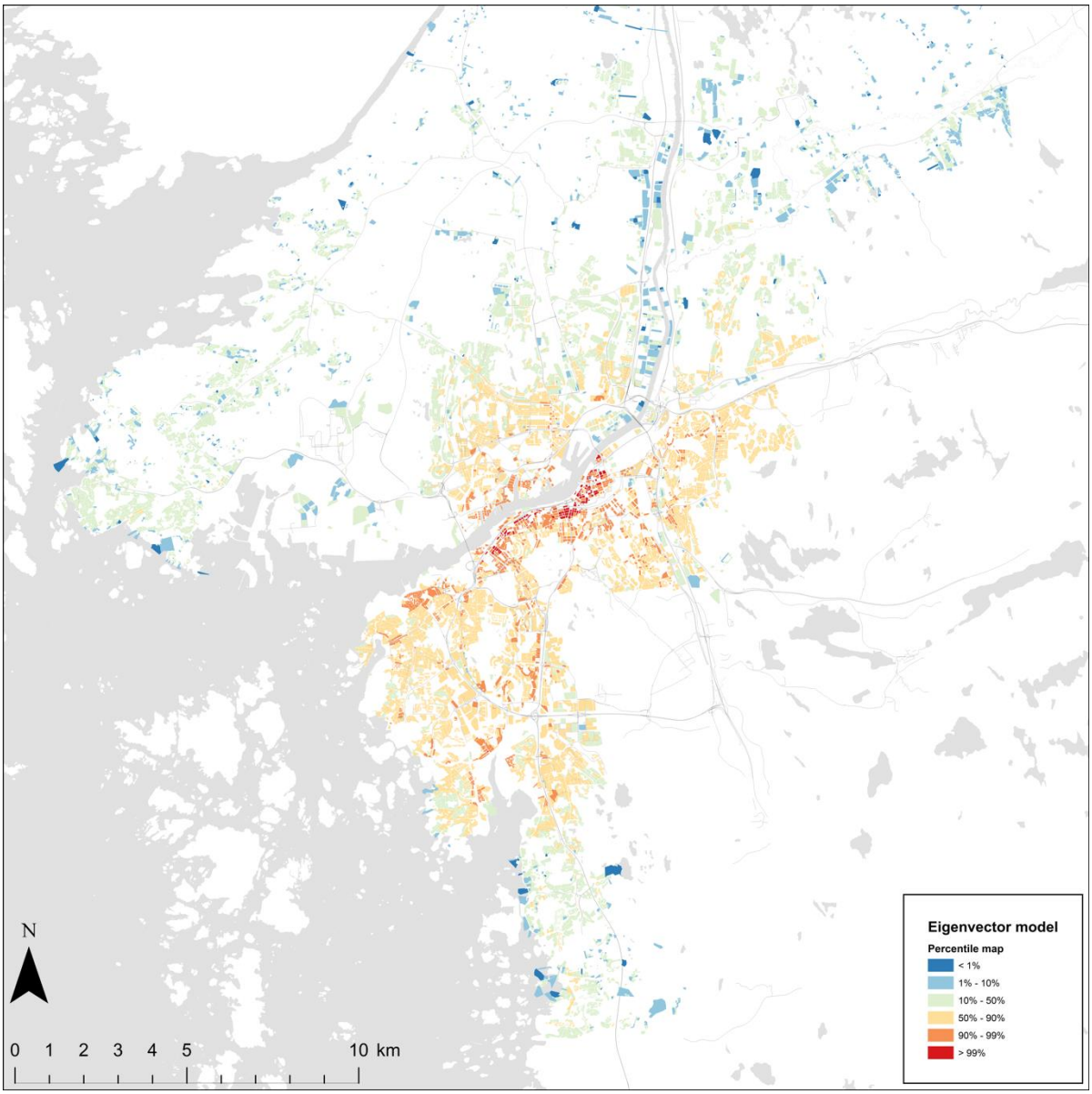


Figure 5. The eigenvector model results.

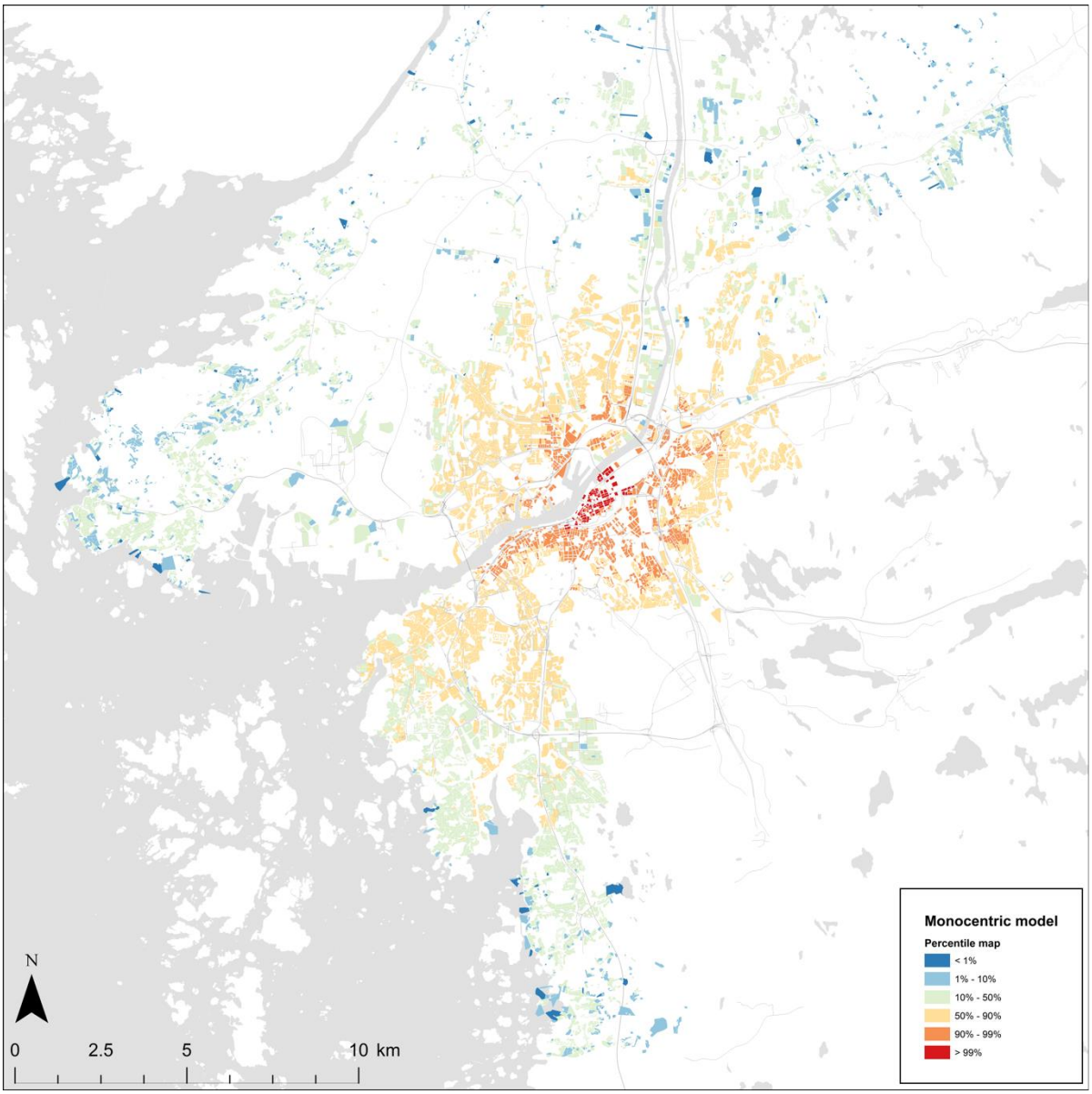


Figure 6. The monocentric model, where the central station is manually defined as the city centre.

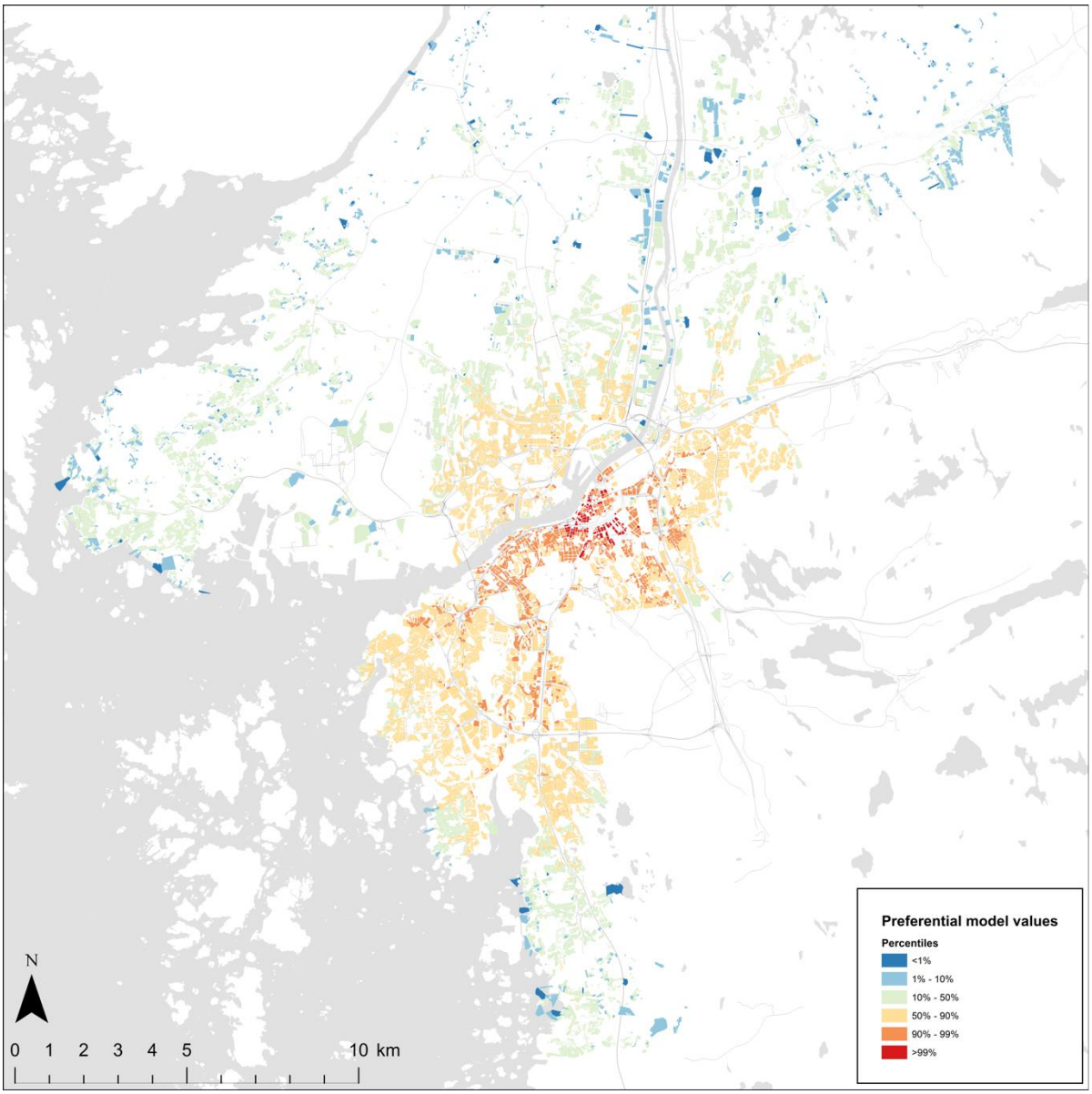


Figure 7. The preferential model, which is an elaboration of the eigenvector model, and therefore performs better compared to empirics.

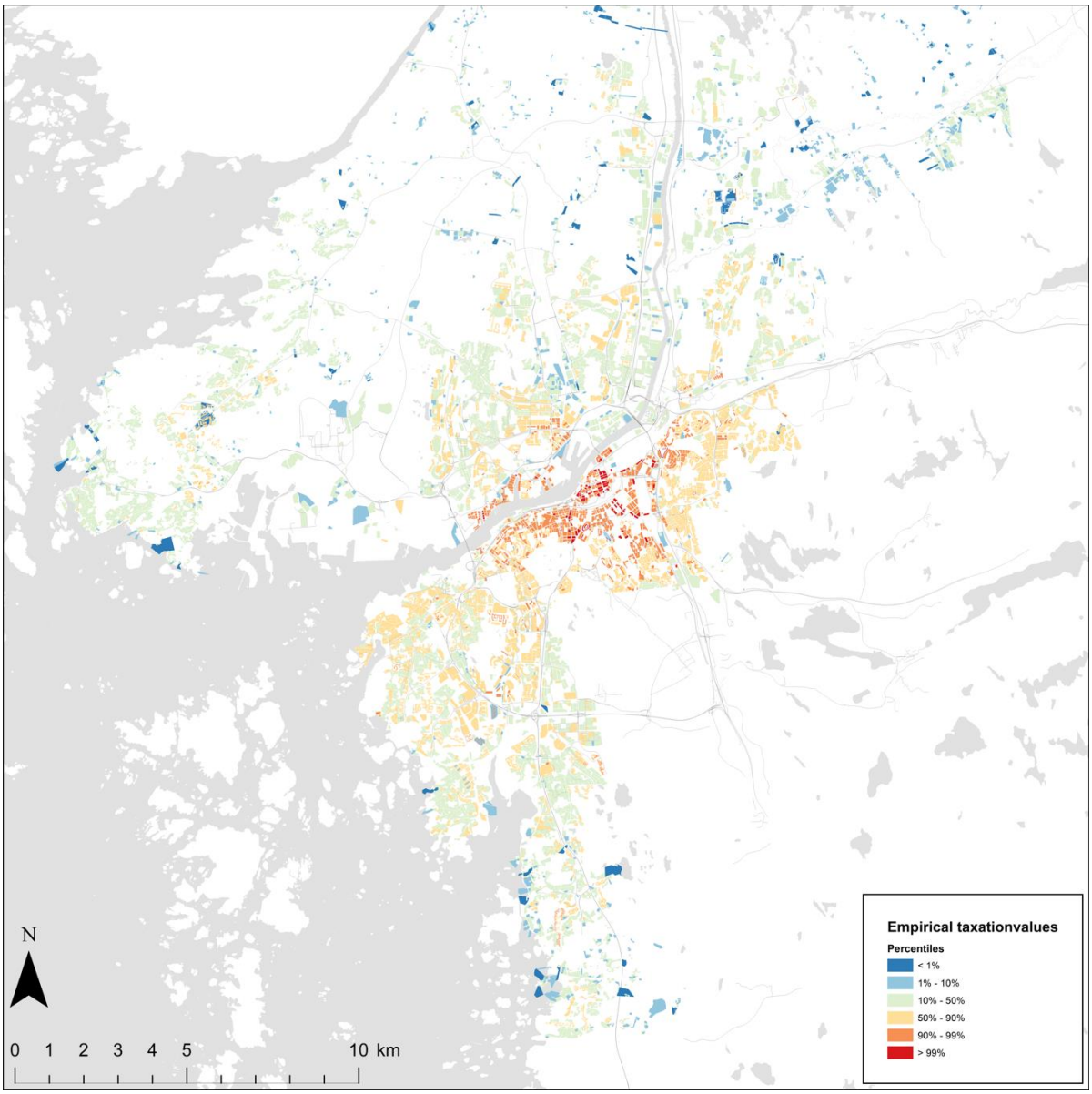


Figure 8. Empirical taxation values used for model validation.

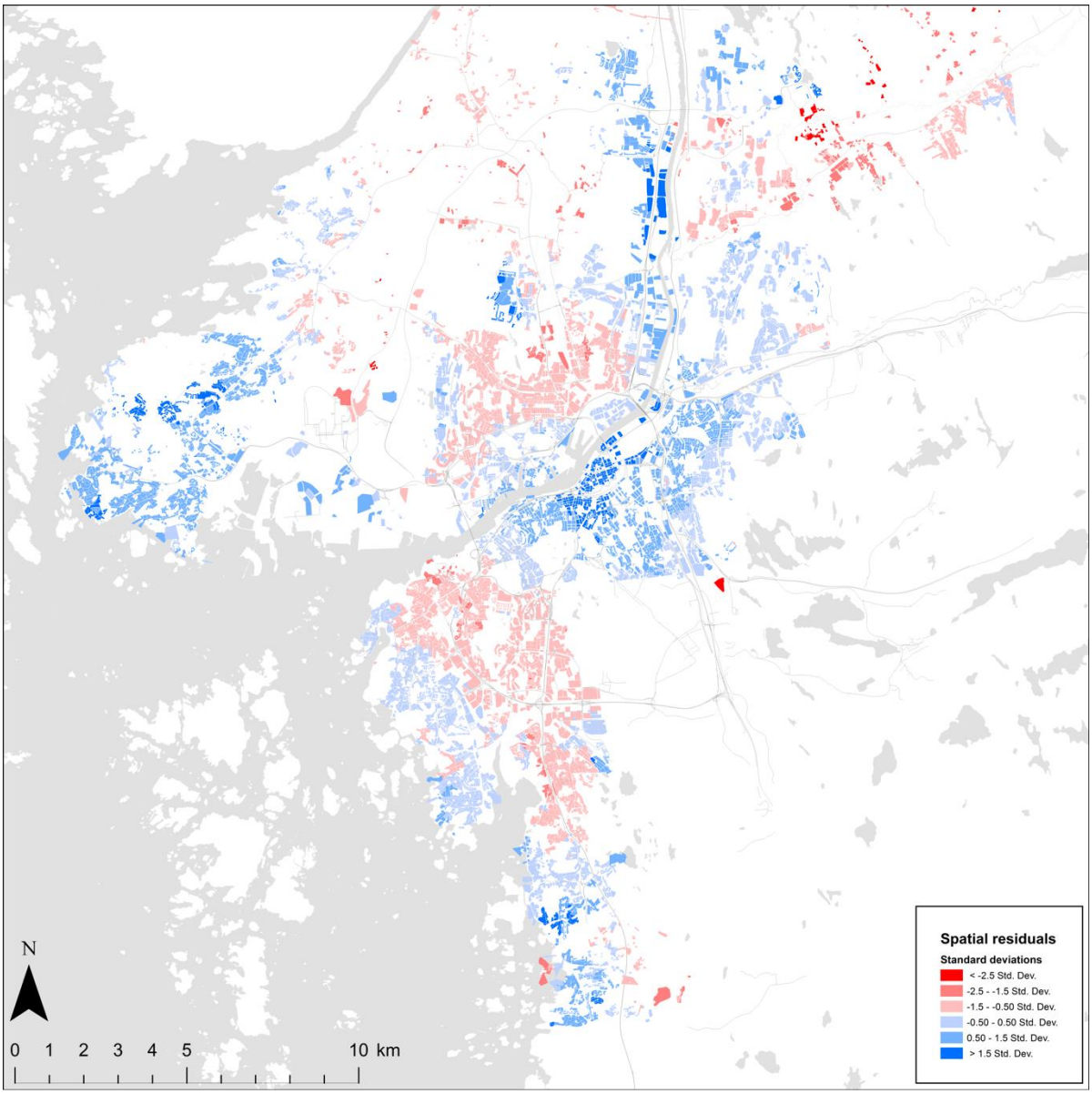


Figure 9. The spatially autocorrelated errors for the preferential model, whom are separated from the "ordinary residuals", by the spatial error model.

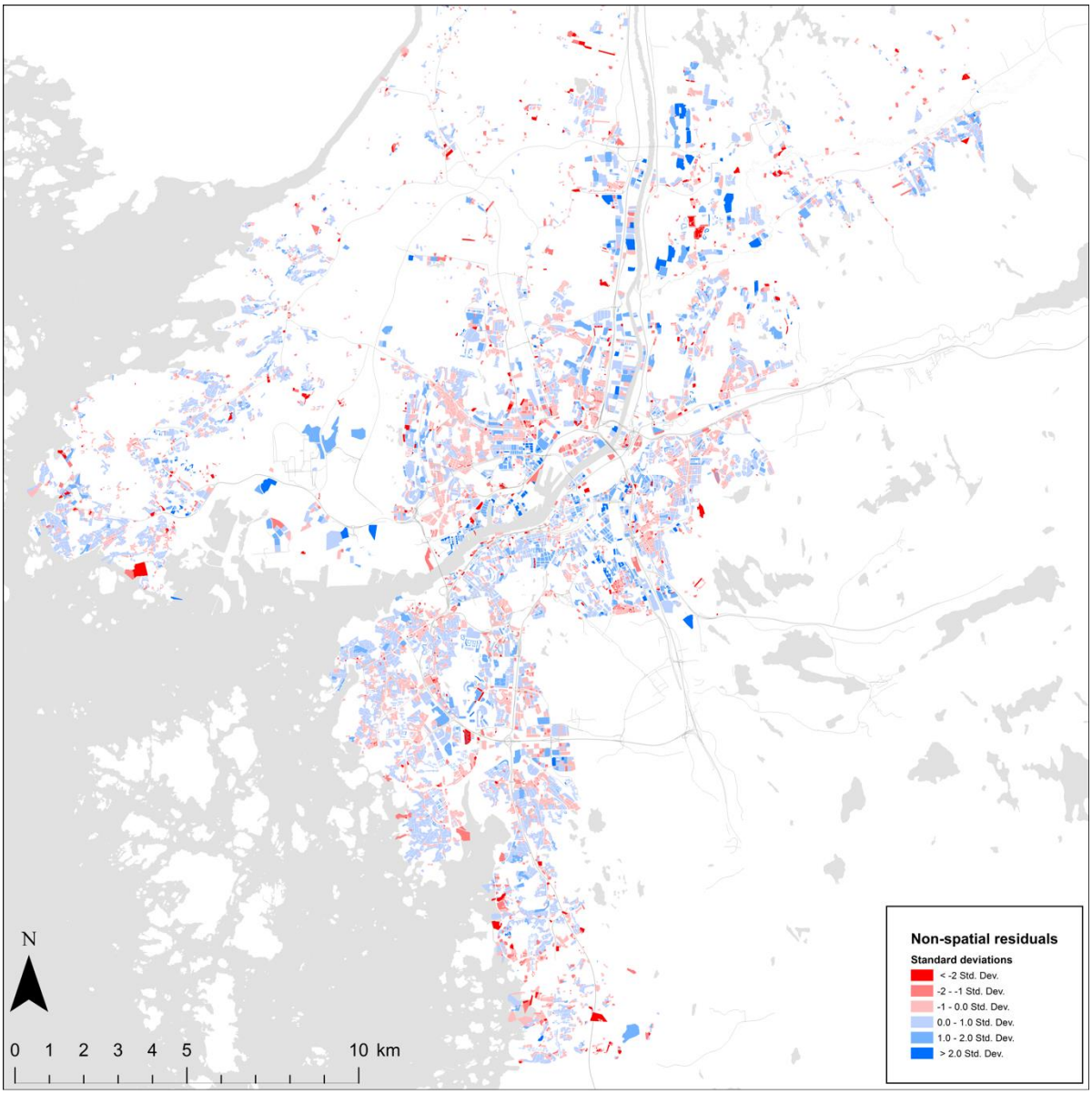


Figure 10. The non-spatial residuals for the preferential model, as usual randomly distributed with mean=0 and standard deviation=1.

Figure 11. A closer look at the spatial structure and the spatial entities used in the model; zones and roads. Close-up view of central parts of Gothenburg.