Modelling Business Rates in England with Big Spatial Data

Shanaka Perera
Warwick Institute for the Science of Cities
University of Warwick
Coventry, UK
S.Perera@warwick.ac.uk

Theo Damoulas
Department of Computer Science and Statistics
University of Warwick
Coventry, UK
T.Damoulas@warwick.ac.uk

Paul Davis
Nimbus Property System Limited
Warwick, UK
Paul.Davis@nimbusproperty.co.uk

Stephen Jarvis
Department of Computer Science
University of Warwick
Coventry, UK
S.A.Jarvis@warwick.ac.uk

ABSTRACT

The continuing growth of the e-commerce industry has increasingly put pressure on traditional retail businesses. Additionally, the traditional retail industry carries a higher tax burden, as they require prime locations to attract a larger customer base, which reflects in higher business rates (non-domestic property tax). Current business rate revaluation has been critisised for misrepresenting true market prices. A better approach to model rateable values is hence needed. We introduce a large-scale, geospatial data set of UK non-domestic rateable values at the most granular level. We propose a state-of-the art Fixed Rank Kriging model to cope with high-dimensionality and learn rateable values from spatial context and property characteristics. By accounting for spatial effects, our model improves on current business rates valuation practice and helps with making the process more fair and transparent.

CCS CONCEPTS

• **Information systems → Data mining**; **Geographic information systems**; **Spatial-temporal systems**;

KEYWORDS

Fixed rank kriging; Spatial big data; Non-domestic rateable values; Spatial prediction

ACM Reference format:
Shanaka Perera, Theo Damoulas, Paul Davis, and Stephen Jarvis. 2019. Modelling Business Rates in England with Big Spatial Data. In *Proceedings of SIGKDD '19: International Workshop on Urban Computing, Alaska, USA, August 04–08, 2019 (SIGKDD '19),* 6 pages.
https://doi.org/10.1145/1122445.1122456

SIGKDD '19, August 04–08, 2019, Alaska, USA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

For the first time in history, more than 50% of the world 's population lives in cities and this is expected to rise to 70% by 2050. The UK is particularly urbanised with over 90% of the population living in towns or cities. The real estate market comprises of domestic and non-domestic properties. Non-domestic properties in the UK are more heterogeneous compared to domestic properties. They contain, amongst others, commercial offices, shops, warehouses, industrial units, hotels, entertainment as well as government offices, educational institutions (schools, colleges, and universities), defence establishments and health sector premises. These provide the physical platform for almost all industries and enterprises along with places for people to work, shop and enjoy leisure activities. Commercial properties account for 13% of the UK built environment and their value represents 10% of the countries wealth. All non-domestic properties are assigned a rateable value by the Valuation Office Agency (VOA). This figure broadly represents the annual rentable value for which the property could be let. Rateable values are used to calculate business rates, charged as tax on property used for business purposes from the occupier of the property by their respective local authority.

There have been many media reports claiming that the latest changes in the rateable values are a threat to many high street businesses [17]. Business rates are a significant cost for firms and can often be the biggest contributed tax paid by the firms. Traditional businesses urge for an even playing field between the digital world of e-commerce and traditional retail store based models, since online retailers minimise their cost by locating facilities out of town. While authorities claim that revaluation is done to reflect up-to-date property values and the strength of the local economy. We believe that spatial modelling of rateable values can substantially improve current revaluation practice and make the process more comprehensible, transparent and fair. However, modelling rateable values at a large scale can be computationally challenging.

Traditional spatial statistical models involve inversion of $n \times n$ variance-covariance matrices, which requires $O(n^3)$ when there are n observations in the data set. This becomes computationally infeasible when n is in the tens of thousands and above [9]. Dimension reduction methods are common in statistics for modelling large

data. Chrissie and Johannesson [4] introduced this in a geostatistical method, Fixed Rank Kriging (FRK) using a flexible class of geostatistical models called spatial random effects (SRE).

The main contributions of our study are: (1) We combine open and proprietary data and introduce a geospatial commercial property data set of unprecedented granularity. (2) This study represents, to the best of our knowledge, first application of Kriging to expose spatial variation in rateable values across different categories of non-domestic properties. (3) Our results can inform both public authority practice and retail business decision making.

The remainder of the paper is structured as follows: Section 2 reviews previous research in the area and presents the data sets from multiple sources in Section 3. Section 4 describes the scientific methodology. Section 5 details the results of the modelling process and Section 6 provides concluding remarks with possible extensions for this study.

2 BACKGROUND RESEARCH

In the recent years, there has been an exponential growth in spatial data with the emergence of social media, location sensing technologies as well as the progression in public sector towards an open data culture. The big data has already become an essential part of business success and scientific discoveries. One of the major challenges in applying spatial queries to large data sets is the high computational complexity [1]. This has emphasised the need to develop new methods to analyse big data [15].

Kriging is a popular spatial prediction method which incorporates spatial variability through variance-covariance functions (variograms) [11]. Universal kriging is commonly applied in real estate valuation, a prominent area in spatial analysis, on relatively smaller data sets [5, 10] . However, computational cost of applying this approach to large data sets has been a long-standing challenge and is discussed extensively in recent research. Local prediction methods with local covariance functions for moving windows [7], local Kriging neighbourhoods [3] and geoadditive models (merging kriging and additive models) [8] are some of the methods developed for spatial predictions in big data context. The fixed ranked kriging method provides a different approach to speed up spatial predictions, using a flexible family of non-stationary covariance functions [4]. This is defined using a set of basis functions, that is fixed in number and leads to computational simplifications when data set is large. This model has been used to analyse a number of large spatial data sets such as Total Column Ozone (TCO) satellites data set (n = 173,405) and global data set of CO_2 measurements (n = 52,128) [9]. In this study, we apply the FRK model to gain inference on business rates data set with over 250,000 observations.

The business rates collected in UK account for a total of £24.2 billion, which represents around 4.5% of the UK tax revenue in 2016-2017 whereas 30% is collected from London. The most recent revaluation of rateable values was conducted in April 2015, seven years after the previous revaluation and came into effect in April 2017. This implies that new values have to reflect changes of seven years in the property market. Business rate tax is both, one of the main income sources for the UK government, and a major expense for businesses. This, along with various criticism of the latest rateable values [6], highlights the importance for the measure to be as accurate as possible. Data on rateable values has been publicly available for purchase on paper from 1968 and moved onto DVD by 2010. In 2016 the data set was published online and free of charge for the public. The consumer location choice is modelled with retail properties data for central London [12], and it is proven to have positive relationship between accessibility of the retail properties with their rateable values by using a small urban area [16]. There have only been very limited studies analysing spatial patterns in non-domestic rateable values in UK, in part caused by the financial cost attached to getting access to other data sources to create a complete spatial data set. This study is conducted in partnership with a leading prop-tech company, Nimbus Property System Limited, who had given us the access to a comprehensive property database for England.

3 DATA

3.1 Business Rates

The Valuation Office Agency (VOA) maintains rateable values, also known as business rates, of around 2 million non-domestic properties in England and Wales. The latest rating list was compiled in April 2017 and the next publication is due in 2022. The rateable value of business properties are usually adjusted every 5 years to reflect changes in the property market. The most common valuation method is the open market annual rental value of the property. Each local billing authority is responsible to compile and maintain the local rating list. The majority of the properties rateable value are supported by the regular site and building survey. The local councils multiply the rateable value with the multiplier set by the VOA to calculate the business rates of non-domestic properties.

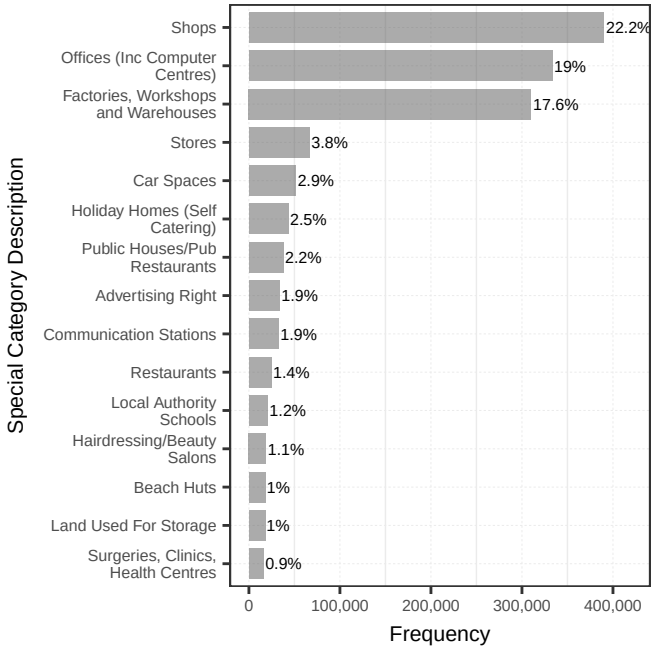

Figure 1: Frequency distribution of non-domestic properties in England.

This includes properties or land that are not solely used for residential. The complete rateable list is downloaded as CSV from the VOA website. Each property is classified into over 300 categories (schools, pubs, hotels, food stores, nightclubs etc.).80% of the non-domestic properties in England are represented by only 4% of the categories. The frequency distribution of these 15 categories are shown in the figure 1.

3.2 Geographic data

The Ordnance Survey Addressbase premium is the most comprehensive address data set for UK, containing approximately 40 million addresses. Each property has a Unique Property Reference Number (UPRN) and is classified as either commercial or residential, and further classified into over 500 categories. The data set provides the spatial point coordinates for each property. This is a commercial proprietary product from Ordnance survey. The spatial database is queried using PostGIS. The cross reference between the VOA and Addressbase is used in this study to develop a comprehensive spatial point data set with rateable values.

3.3 London Business rates

Our focus in this study is on the rateable values of non-domestic properties in London, the capital and largest city in England. The spatial intersect between the property data set and statistical GIS boundary for London was used to subset the data set. Business rates are charged from 277,906 non-domestic properties in London. 90% of the non-domestic properties in London are represented by 16 of the categories and the frequency distribution is shown in figure 2.

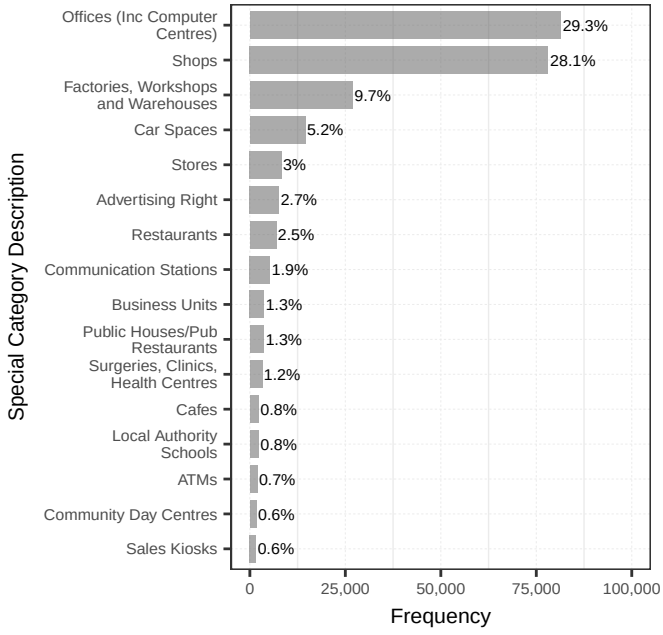

Figure 2: Frequency distribution of Non-domestic properties in London.

The rateable value ranges between £41 and £212.4 million with an average value of £63,461. The log transformation is used on rateable values to make the data less skewed. Figure 3. shows the variability in distributions for log of rateable value for each individual category used in this study for London.

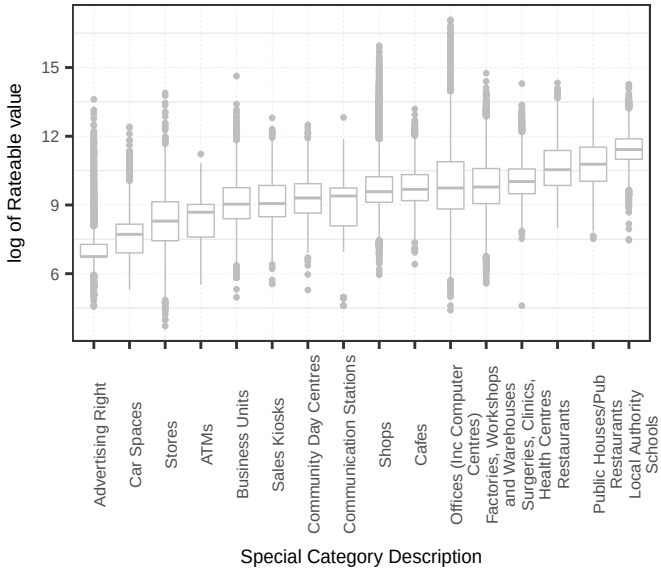

Figure 3: Frequency distribution of Rateable values in London.

4 METHODOLOGY

4.1 Fixed Rank Kriging (FRK)

Kriging infers a best linear unbiased prediction (BLUP) and is hence a popular geostatistical prediction method [2]. However, its covariance matrix calculations are computationally expensive, as outlined above. While powerful with small data sets, the growth in spatial big data poses a growing challenge. Chrissie and Johannesson [4] introduced the FRK model to analyse very large data sets, reducing computational cost to $O(n)$ from $O(n)^3$.

This study is interested in making inference on a hidden spatial process $\{Y(\mathbf{s}) : \mathbf{s} \in D_s\}$ on the spatial domain of London. Following [4] and considering the process Z(·) of actual and potential observations

$$Z(\mathbf{s}) \equiv Y(\mathbf{s}) + \varepsilon(\mathbf{s}) \tag{1}$$

where $\{\varepsilon(\mathbf{s}) : \mathbf{s} \in D\}$ is a spatial white noise process distributed as $\varepsilon(\mathbf{s}) \sim N(0, \sigma^2 v(\mathbf{s}))$ and $v(\mathbf{s})$ is known. The vector of available data at spatial locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$

$$\mathbf{Z} \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))' \tag{2}$$

The process $Y(\cdot)$ assumed to have a linear mean structure,

$$Y(\mathbf{s}) = x(\mathbf{s})'\alpha + v(\mathbf{s}) \tag{3}$$

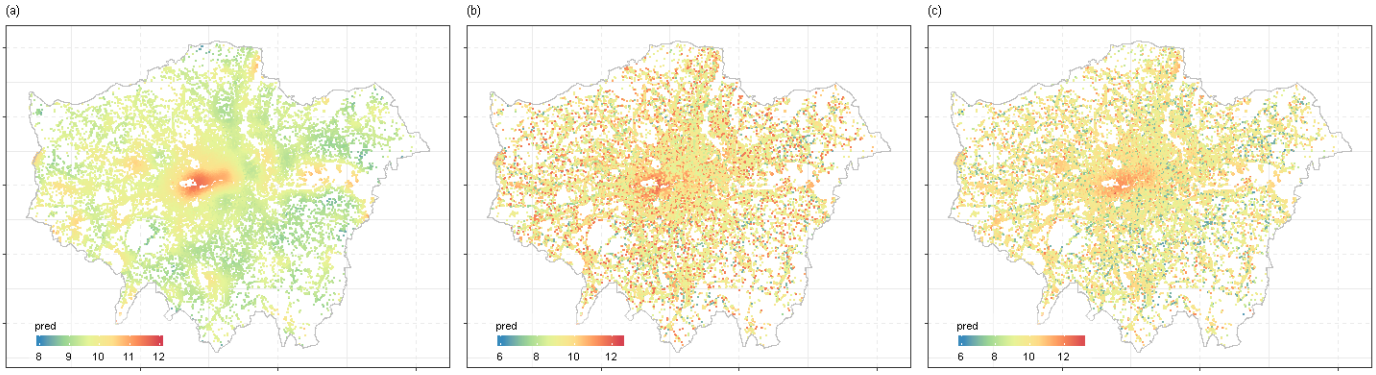

Figure 4: Prediction of log Rateable value obtained from FRK: (a)Model 1; (b)Model 2; (c)Model 3.

where $x(\cdot)$ represents a vector process of known covariates and coefficient α are unknown. $\nu(\cdot)$ has zero mean, $0 < var\{\nu(\mathbf{s})\} < \infty, \forall \mathbf{s} \in D$ and generally a non stationary spatial covariance function,

$$cov\{\nu(\mathbf{u}), \nu(\mathbf{v})\} \equiv C(\mathbf{u}, \mathbf{v}) \tag{4}$$

In general, the covariance function is modelled as being stationary, in which case it must be a non-negative-definite function of $\mathbf{u} - \mathbf{v}$. In the FRK model the spatial dependence is captured through a set of basis functions,

$$\mathbf{S}(\mathbf{u}) \equiv (S_1(\mathbf{u}), \ldots, S_r(\mathbf{u}))', \qquad \mathbf{u} \in \mathbb{R}^d \tag{5}$$

and $cov\{\nu(\mathbf{u}), \nu(\mathbf{v})\}$ is modelled as,

$$C(\mathbf{u}, \mathbf{v}) = \mathbf{S}(\mathbf{u})'\mathbf{K}\mathbf{S}(\mathbf{v}), \qquad \mathbf{u}, \mathrm{v} \in \mathbb{R}^d \tag{6}$$

where $\mathbf{K}$ is an unknown $r \times r$ symmetric positive-definite matrix. The expression (6) is a consequence of writing $\nu(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\eta, \mathbf{s} \in D$, where η is a r dimensional vector with $var(\eta) = \mathbf{K}$ and $\nu(\cdot)$ is called a spatial random effects (SRE) model.

The method was implemented using the FRK package [18] in the R statistical programming language. The basis functions are generated as a set of local basis functions in the domain with maximum of 2000 basis functions, and prune in regions of sparse data.

4.1.1 Three models are evaluated in modelling the logarithm of rateable values:

(1) Model with no covariates.
This model only uses the spatial coordinates to fit the model and will not use the information about the category of the properties.
(2) Model for each category with no covariates.
An individual model fits for each category using the spatial coordinates.
(3) Model with category as the covariate.
The category of the non-domestic property and spatial coordinates are used in the model.

4.2 Cross validation

The standard data sampling methods used for cross validation (CV) to evaluate prediction performance assumes the training and testing data are independent of each other. According to the first law of geography "Everything is related to everything else, but near things are more related than distant things" [14]. This causes the standard sampling methods to produce optimistic performance measures for spatial models. Spatial k-fold cross validation (SKCV) is a modification method of the standard CV to remove the spatial auto correlation (SAC) between the training and testing data [13]. This is achieved by removing training data within a pre-determined radius, known as the deadzone, around the test data. There is a trade off between the radius of deadzone and the loss of data in the training sample.

4.2.1 Three data sampling methods used for CV in this study:

(1) Standard k-fold CV
(2) SKCV with 20m deadzone
(3) SKCV with 50m deadzone

Three validation matrices are utilised to evaluate prediction performance:

(1) r^2 to measure predictor's 'goodness of fit'.

$$r^2 = \left(\frac{n\Sigma(y_i\hat{y}_i) - \Sigma(y_i)\Sigma(\hat{y}_i)}{\sqrt{(n\Sigma y_i^2 - (\Sigma y_i)^2)(n\Sigma\hat{y}_i^2 - (\Sigma\hat{y}_i)^2)}} \right)^2 \tag{7}$$

where y_i is the actual log of rateable value and $\hat{y}_i$ is the predicted log of rateable value.

(2) Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

(3) Mean Absolute Error (MAPE) expressed as a percentage

$$MAPE = \frac{100}{n}\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{y_i}\right) \tag{9}$$

Table 1: Results table for three models with the three validation techniques.

	k-fold			Dead zone - 20M			Dead zone - 50M		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
r^2	0.17	**0.54**	0.49	0.16	**0.53**	0.49	0.14	**0.52**	0.47
RMSE	0.096	**0.072**	0.075	0.097	**0.073**	0.078	0.098	**0.074**	0.079
MAPE	9.48	**6.96**	7.27	9.56	**7.04**	7.37	9.7	**7.13**	7.49

5 RESULTS

The three models discussed in section 4.1.1 are fitted for each of the 10 folds using the data sampling methods mentioned in the section 4.2.1. Under the k-fold sampling 90% of the data were used in training at each fold but for SKCV on average, only 68% and 38% of the data were used for 20m and 50m deadzones respectively. Further increase in deadzone radius would result in less data for training, hence 20m and 50m radius were used for cross validation.

We obtained FRK predictions for point locations on the test data set for each fold. Figure 4 shows the outcome of the three models for predictions at all points using SKCV with 20m deadzone. Similar patterns were observed for the other sampling methods. Hot spots of high rateable values are observed in the centre of each map, which represents Central London. 4(a) is more smoother compared to 4(b) and (c). This is likely due to the fact that Model 1 is not using the category of the property in modelling.

A summary of all the performance for validation data are recorded in table 1. The bold font represents the best model for each sampling technique. r^2 double when the model uses information about the category of the property. This emphasises that business rates are influenced significantly by the category of the property in addition to the location. SKCV with a deadzone is utilised to penalise the over bias caused by spatial autocorrelation. The k-fold cross validation is providing optimism due to the overestimation of statistical effects but 50m deadzone removes 60% of the training set, so although it removes the SAC between the training and test set, it also provides pessimism in the fact that it has a smaller training set. However, notably, there is no significant difference in the performance across the sampling techniques.

In order to understand the performance of our model, for each category we have calculated the r-squared of the validation data for the three models and shown in Figure 5. Model 2 performs better for each category compared to model 1 and 3. There is notable difference in r-squared for three models in the *sales kiosks* category which has the least number of properties in the subset used for London (Figure 2). Furthermore, *restaurants* shows the highest r-squared under all the three models despite representing only 2.5% of the data. Overall figure 5 shows that the predictability varies greatly (r^2 between 0.01 to 0.48) with the category of the property and in combination with figure 2 provides no evidence this is driven by the number of observation in each category. Furthermore, *restaurants*, *cafes* and *offices* show similar r-squared for all three models which indicates that these categories are representative of the overall system compared to *Sales Kiosks*, *Pubs*, *Business units* and *Factories* tend to have their own sub systems.

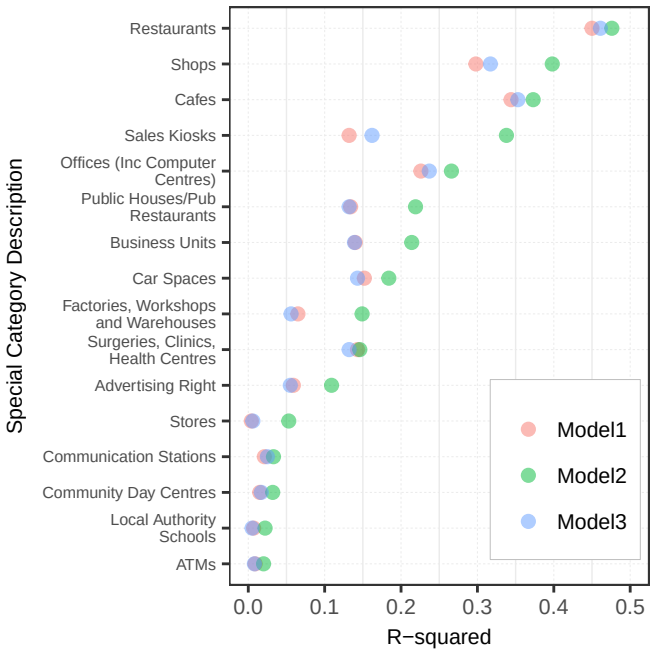

Figure 5: R-squared for each property category in SKCV with 50m deadzone.

6 CONCLUSIONS

In this study a comprehensive spatial big data set for non-domestic rateable values for England is developed. A state-of-the-art kriging method for large data, FRK, is applied to obtain inferences on hidden spatial patterns. We propose three separate models for predicting rateable values: (1) FRK without covariates (purely spatial), (2) FRK for each property category without covariates and (3) FRK with property categories as covariates. We find that the disaggregate model 2 performs best with a r^2 of 0.53. This implies that not all the non-domestic property categories follow the same spatial distribution. The overall system can only partially describe the processes in the sub systems. Our predictions improve for *Shops* and *sales kiosks* substantially when looking at the category level.

However, *Restaurants*, *cafes* and *offices* behave similar to the overall system which reflects in almost identical r^2 values. We also find that *Restaurants*, *shops* and *cafes* show more spatial dependence whereas *ATMs*, *schools* and *day centers* do not follow distinct spatial patterns.

This study contributes to current research in the following ways: we introduce a newly compiled geospatial data set for non-domestic properties across England. We model spatial interdependencies on a large scale using a flexible FRK method and provide first spatial insights into the underlying processes. These findings can help to improve the current rateable values revaluation practices. The industry can benefit from more reliable business value estimates to motivate data-driven decisions making. Future work should incorporate more relevant covariates such as size of building or local demographics. The FRK method could be extended to all of England and other countries and the outcomes across major cities, countries and the urban/ rural domain could be compared.

ACKNOWLEDGMENTS

We would like to thank the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Urban Science (EP/ L016400/ 1). T. Damoulas and S. A. Jarvis are members of the Alan Turing Institute in London, the UK 'new national centre for data science. We would also like to thank Nimbus property system limited for giving access to a comprehensive property database and the colleagues H. Crosby, K. Klemmer for their support.

REFERENCES

[1] Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, and Joel Saltz. 2013. Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases* 6, 11 (2013), 1009–1020. https://doi.org/10.14778/2536222.2536227 arXiv:NIHMS150003

[2] Noel Cressie. 1990. The origins of kriging. *Mathematical Geology* 22, 3 (1990), 239–252. https://doi.org/10.1007/BF00889887

[3] Noel Cressie. 1993. Statistics for spatial data. *Terra Nova* 4, 5 (1993), 131–134.

[4] Noel Cressie and Gardar Johannesson. 2008. Fixed Rank Kriging for Very Large Spatial Data Sets. *Journal of the Royal Statistical Society, Series B* 70, 1 (2008), 209–226.

[5] Henry Crosby, Paul Davis, Theo Damoulas, and Stephen A Jarvis. 2016. A Spatio-Temporal , Gaussian Process Regression , Real-Estate Price Predictor. (2016), 3–6.

[6] Barry Davies. 2017. The going rate. *Land Journal* (2017), 14.

[7] Timothy C. Haas. 1995. Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *J. Amer. Statist. Assoc.* 90, 432 (1995), 1189–1199. https://doi.org/10.1080/01621459.1995.10476625

[8] E. E. Kammann and M. P. Wand. 2003. Geoadditive models. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 52, 1 (2003), 1–18. https://doi.org/10.1111/1467-9876.00385

[9] Matthias Katzfuss and Noel A C Cressie. 2011. Tutorial on Fixed Rank Kriging (FRK) of CO2 Data. *Terra* 858 (2011).

[10] Michael Kuntz and Marco Helbich. 2014. Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. *International Journal of Geographical Information Science* 28, 9 (2014), 1904–1921.

[11] Georges Matheron. 1963. Principles of geostatistics. *Economic geology* 58, 8 (1963), 1246–1266.

[12] Duccio Piovani, Vassilis Zachariadis, and Michael Batty. 2017. Quantifying Retail Agglomeration using Diverse Spatial Data. *Scientific Reports* 7, 1 (2017), 1–8. https://doi.org/10.1038/s41598-017-05304-1 arXiv:1612.06441

[13] Jonne Pohjankukka, Tapio Pahikkala, Paavo Nevalainen, and Jukka Heikkonen. 2017. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science* 31, 10 (2017), 2001–2019. https://doi.org/10.1080/13658816.2017.1346255 arXiv:arXiv:1505.06786v1

[14] Waldo R Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, sup1 (1970), 234–240.

[15] Ranga Raju Vatsavai, Auroop Ganguly, Varun Chandola, Anthony Stefanidis, Scott Klasky, and Shashi Shekhar. 2012. Spatiotemporal data mining in the era of big spatial data. *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data - BigSpatial '12* (2012), 1–10. https://doi.org/10.1145/2447481.2447482

[16] Peter J. Wyatt. 1997. The development of a GIS-based property information system for real estate valuation. *International Journal of Geographical Information Science* 11, 5 (1997), 435–450. https://doi.org/10.1080/136588197242248

[17] Rob Young. 2018. Business rates changes 'are threat' to High Streets - BBC News. (2018). www.bbc.co.uk/news/business-43632000

[18] Andrew Zammit-Mangion. 2018. *FRK: Fixed Rank Kriging*. https://CRAN.R-project.org/package=FRK R package version 0.2.2.