Spatial bootstrapped microeconometrics:

forecasting for out-of-sample geo-locations in big data

Katarzyna Kopczewska

Faculty of Economic Sciences, University of Warsaw, Poland

ORCID: 0000-0003-1065-1790, kkopczewska@wne.uw.edu.pl

Abstract: Spatial econometric models estimated on the big geo-located point data have at least two problems: limited computational capabilities and inefficient forecasting for the new out-of-sample geo-points. This is because of spatial weights matrix W defined for in-sample observations only and the computational complexity. Machine learning models suffer the same when using kriging for predictions; thus this problem still remains unsolved. The paper presents a novel methodology for estimating spatial models on big data and predicting in new locations. The approach uses bootstrap and tessellation to calibrate both model and space. The best bootstrapped model is selected with the PAM (Partitioning Around Medoids) algorithm by classifying the regression coefficients jointly in a non-independent manner. Voronoi polygons for the geo-points used in the best model allow for a representative space division. New out-of-sample points are assigned to tessellation tiles and linked to the spatial weights matrix as a replacement for an original point what makes feasible usage of calibrated spatial models as a forecasting tool for new locations. There is no trade-off between forecast quality and computational efficiency in this approach. An empirical example illustrates a model for business locations and firms' profitability.

Keywords: *tessellation; Voronoi polygons; spatial point-pattern; bootstrapping; spatial weights matrix; spatial big data; predictions out-of-sample*

Plain Language Summary: The paper proposes a highly novel forecasting methodology with the spatial econometric model in the case of point geo-located data. It links statistics, econometrics and machine learning, never combined together as here. It is dedicated to big spatial data as it deals with computational complexity by limiting dataset size. It solves unsolved: how to borrow spatial information from the neighbourhood to predict in new out-of-sample locations. It uses a few smart tricks, which do not exist until now in the scientific literature: bootstrap, which is here resampling of subsets, to find most representative observations in many trials; PAM (Partitioning Around Medoids) algorithm to find the middle (most representative) model and to refer to mutual relations between model's coefficients; Voronoi polygons for data used in the best model to create catchment areas for new points – we use them to substitute old point with new out-of-sample location to run the prediction which uses information from the basic model. It proves that by using only a part of the dataset for

estimation, one gets a very good model, while estimation saves time.

1. Introduction

The key point in the geo-located data is the similarity in the neighbourhood expressed as spatial autocorrelation and captured by spatial weights matrix W (more in Appendix). Classical spatial econometrics uses W matrix with contiguity criterion in modelling regional areal data and k nearest neighbours W matrix for point data. This W matrix, being a core of spatial econometrics, simultaneously is its main problem – due to being intractable in big data and defined only for the observations used for model fitting. This generates two severe problems for spatial econometrics on point data: 1) usage in the case of big data and 2) predictions for new locations. As for now, alternatives for W in big data are not satisfactory – parallel computing increases the speed but does not solve the critical issue of memory; spatial machine learning is still discussed (Kopczewska, 2021) as W is not easy to substitute; models on spatially aggregated point data and polygon-based W erase the information from the local neighbourhood and cut the local variability; kriging, which may deal with out-of-sample smoothing of the target variable, needs inverting distance-dependent large and dense covariance matrices, which limits its applications in big data (Perdikaris et al., 2015); *Matrix Exponential Spatial Specification* can work only with a single neighbour in classical spatial models (LeSage & Pace, 2007; Arbia, 2014). Neglecting spatial autocorrelation in geospatial data is also dangerous as it mostly results in bias, inconsistency, overfitting of the non-spatial model and its false predictions (e.g. Ibrahim & Bennett, 2014). Recent advances include for big spatial data, primarily the storage solutions, and for out-of-sample predictions, mostly the discussion on challenges (Jiang, 2018).

Spatial big data starts at the current edge of the computational capabilities of W, which lies around 70.000 observations (Arbia et al., 2019), and for bigger datasets, one needs new methods. The real challenge is to estimate the spatial econometric models using W and information from the neighbourhood for big data - datasets of thousands or millions of geo-points, e.g. for real estate valuation or business location. Currently, only super-computers can do this task. And even if the current computational progress enables estimation on larger datasets than before, one can easily imagine increasingly bigger tasks that already may appear and will be too complicated again.

The feasibility of geo-data micro-econometric spatial models on the standard machines can be achieved with a bootstrapping technique. Till now, bootstrap was considered mainly as a tool supporting the small-sample data (e.g. Hall, 2013; Hesterberg, 2015), and the bootstrapping replications were expected to discover the hidden population statistical properties. In general, bootstrap is used mainly for smoothing (e.g. Davison et al. 2003), testing (e.g. Manly, 2006), confidence intervals (e.g. DiCiccio & Efron, 1996), internal validation of models (Tran & Tran, 2016). In spatial analyses, bootstrap is used in the uncertainty bands in functional kriging (Franco-Villoria & Ignaccolo, 2017), testing the spatial non-stationarity in the Geographically Weighted Regression coeffi-

cients (Harris et al., 2017), correlation functions (Loh, 2008), sampling under known joint distribution (García-Soidán et al., 2014), testing Moran's I (Jin & Lee, 2015), discovering uncertainty of parameters (Uboldi et al., 2014, Dalposso et al., 2019, Castillo-Páez et al., 2020). However, bootstrapping can be used oppositely, shrinking big data size while preserving its statistical properties. Predominantly, data are reduced by sampling in univariate statistical analyses, and for the econometric purposes in the multivariate analyses by Lasso (Tibshirani, 1996), bootstrap of the candidate parameters matrices (Ye & Weiss, 2003), bolasso (Bach, 2008) or PCA (e.g. Rosipal et al., 2001) when cutting redundant variables. However, bootstrapping can be an efficient tool in limiting the number of observations, alternative to the sampling. Within this developing stream, one can find a proposal by Barbian and Assunção (2017) of spatial subsemble for partial estimations in spatially structured subsets and aggregating the results in the spatial analysis.

Bootstrapped regression, based on the sampling of smaller subsamples and replicating the model estimations, is present for 30 years (since Freedman, 1981 and Wu, 1986, and now as Hesterberg, 2015; Harris et al. 2017). It enables operating on a much narrower scale while obtaining consistent, efficient and non-biased estimates (e.g. Davison et al., 2003; Efron & Tibshirani, 1997). However, there are still two challenging issues because of the spatiality of data. Firstly, when sampling the observations, one samples also the location. Thus, there is no unique spatial weights matrix W as each estimated model is based on its individual W because of the different spatial composition of points. Secondly, the forecasting possibilities for a new geo-point are limited as the new point is not represented in W.

This paper proposes the solution to all three issues discussed above: *i)* computational problems in big data, *ii)* lack of unique W when sampling, and *iii)* difficult forecasting for out-of-sample data. The subsequent sections present the method and its justification. The logic is as follows: the paper starts with true estimates of coefficients and its errors on population (full dataset), which becomes the reference for other solutions and enables the quality comparisons (sec.2). Straightforward reduction of the dataset (sampling) as an alternative would reduce computation time and lower the model quality but does not solve the problem of out-of-sample prediction and unique W (sec.3). Bootstrapped regression improves computational aspects and keeps the characteristics of full sample estimates but requires deciding how many replications and observations to use (sec.4). Having bootstrapped distributions of beta coefficients in multivariate regression, one cannot simply choose an average of each beta coefficient individually due to interrelated values – a solution is to choose the most central set of coefficients in this multi-dimensional setting. This is equivalent to finding the most representative model by searching for the medoid model using Partitioning Around Medoids (PAM) clustering (sec.5). Locations of observations used in the most central model are treated as the best approximation of point-pattern and used to derive a representative spatial weights matrix W. Flexible prediction for any new point within analysed territory requires linking in pairs

the existing *in-sample* points with new *out-of-sample* points. This is possible by partitioning the area into disjoint subregions using tessellation and pairing points that belong to the same tessellation tile (sec.6). Forecasting which inputs to the model one new point and all-but-one old points gives well-fitted predictions (sec.7). The proposed method is discussed (sec.8) and summarised (sec.9). All stages are presented in Fig.1

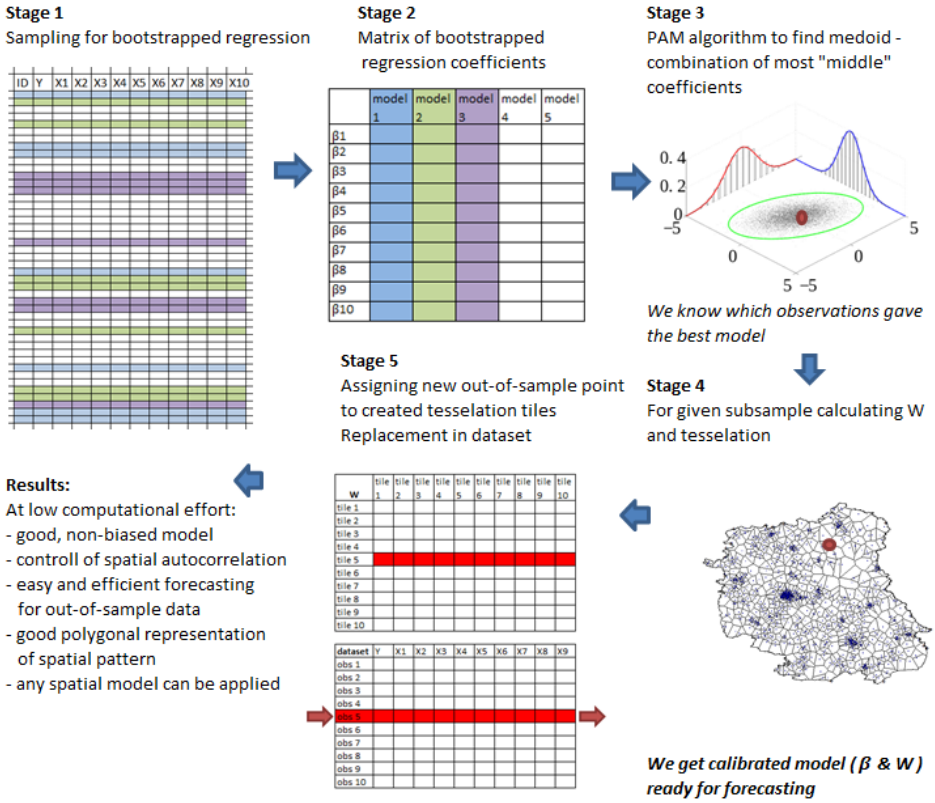


Figure 1: Design of study

The paper's novelty is in: a) using bootstrap to shrink the dataset size, which is different from typical applications of bootstrap to small samples, b) selecting the best bootstrap multivariate model with Partitioning Around Medoids algorithm, c) substituting train data with test data in W for predictions based on pairing design, d) pairing in-sample and out-of-sample locations by overlaying new points on the tessellated surface with old points and assigning to underlying tiles. All those innovations build complex approach to estimate spatial models on big data and run predictions for out-of-sample data. This methodology is novel in spatial econometrics, and it gives very stable and efficient results at the low level of computational effort.

1. **Reference full-sample spatial econometric model**

Let's assume the dataset with more than *n=37,000* geo-located point observations, which are static real business locations within a region (Fig.2). The geo-referenced point-locations (*x,y*) are supplemented with a real business characteristic (*z*) as an industry branch (agriculture *agri*, production *prod*, construction *constr*, service *serv*), employment size (*empl*), Euclidean distance to Lublin core city (*dist*). The analysed dataset is the representative 10% sample of real REGON register for the Lubelskie region in Poland in 2014. The profitability (*roa*) (understood as Return on Assets) was generated assuming the premium of 1,5% in the core city and normal distribution with given parameters within a sector - N(2%, $0.045\%^2$) in agriculture, N(3,5%, $0.045\%^2$) in production, N(5%, $0.045\%^2$) in construction and N(8%, $0.045\%^2$) in service.

The research goal is to estimate and calibrate the model explaining the profitability of a given firm (*j*) with its location and characteristics, including all possible spatial information, and to forecast the profitability (*z**) of a new entry firm in a given new location (*x*,y**) for given industry, employment and distance.

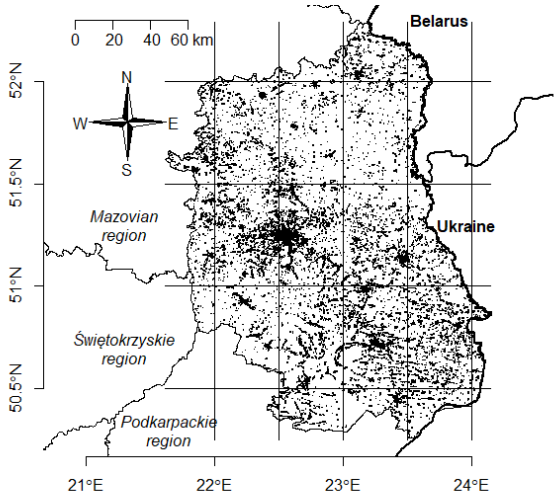


Figure 2: Locations of observations used in modelling

The paper considers four specifications of the same model. The first is a standard linear model specification (estimated with Ordinary Least Squares, OLS):

$$\text{roa}_j = \beta_0 + \beta_1 \bullet \text{empl}_j + \beta_2 \bullet \text{prod}_j + \beta_3 \bullet \text{constr}_j + \beta_4 \bullet \text{serv}_j + \beta_5 \bullet \text{dist}_j + \varepsilon_j \quad (1)$$

where *j=1,2,…,n* are observations in the dataset, roa_j and $dist_j$ are continuous variables on the profitability and distance to the core city, $empl_j$ is a numeric variable specifying the middle of the employment size class, $prod_j$, $constr_j$ and $serv_j$ are the dummy variables differentiating the business sector from the agriculture (base level), while ε_j is error term.

The second, third and fourth specifications are spatial models: spatial error model (SEM) (2), spatial autoregressive model (SAR) (3) and spatial Durbin model (SDM) (4):

$\text{roa}_j = \beta_0 + \beta_1 \text{empl}_j + \beta_2 \text{prod}_j + \beta_3 \text{constr}_j + \beta_4 \text{serv}_j + \beta_5 \text{dist}_j + \varepsilon_j$ and $\varepsilon_j = \lambda W \varepsilon_j + u_j$	(2)
$\text{roa}_j = \beta_0 + \text{Wroa}_j + \beta_1 \text{empl}_j + \beta_2 \text{prod}_j + \beta_3 \text{constr}_j + \beta_4 \text{serv}_j + \beta_5 \text{dist}_j + u_j$	(3)
$\text{roa}_j = \beta_0 + \text{Wroa}_j + \beta_1 \text{empl}_j + \beta_2 \text{prod}_j + \beta_3 \text{constr}_j + \beta_4 \text{serv}_j + \beta_5 \text{dist}_j + + \theta_1 W\text{empl}_j + \theta_2 W\text{prod}_j + \theta_3 W\text{constr}_j + \theta_4 W\text{serv}_j + \theta_5 W\text{dist}_j + u_j$	(4)

where ε_j is a spatially auto-correlated error term, decomposed to its spatial lag $\text{W}\varepsilon_j$ component and the *iid* random term u_j, Wroa_j is autoregressive component and WX_j are spatial lags of dependent variables X, and W is *37,000 x 37,000* spatial weights matrix using *k=5* nearest neighbours (*knn=5*). Estimation of models was run in R software (more in Appendix).

The estimation results show that in OLS model (eq.1) all variables are significant (*** for p-value<0.001) (Tab.1), goodness-of-fit is acceptable (R^2=0.98, AIC=43602). At the same time, the spatial autocorrelation of the error term exists (Moran's I standard deviate=316.91 using the *knn=5* in W). Spatial models: SEM (eq.2), SAR (eq.3) and SDM (eq.4) with *knn=5* W matrix were much better fitted (with AIC_{SEM}=-3030.51, AIC_{SAR}=30491 and AIC_{SDM}=-3780.9 respectively) than OLS and obtained significant spatial coefficients (*rho* ρ, *lambda* λ, and *theta*). There exists an upward bias in OLS coefficients of ca. 5% compared to spatial models, while standard errors of OLS are almost doubled than in SEM, SAR and SDM. Coefficients for *prod*, *constr* and *serv* are the premium of profitability over the agriculture sector, taken as a base. The average ROA for a basic sector - agriculture ($\mu^{\text{ROA}}_{\text{agri}}$=2%) added up with sectoral coefficients ($\beta^{\text{SEM}}_{\text{prod}} \approx 1.5$, $\beta^{\text{SEM}}_{\text{constr}} \approx 3.0$, $\beta^{\text{SEM}}_{\text{serv}} \approx 6.0$) sum up to the assumed average ROA ($\mu^{\text{ROA}}_{\text{prod}}$=3.5%, $\mu^{\text{ROA}}_{\text{constr}}$=5%, $\mu^{\text{ROA}}_{\text{serv}}$=8%). *Emp* variable revealed instability which has no importance for methods presented in a paper. For structural interpretation of a given factor's impact on a dependent variable, one should use direct and indirect impacts, estimated for the best-selected model.

Table 1: Estimation results of OLS, SEM, SAR and SDM for a full sample (37,000 obs)

models → variables ↓	OLS coefficient		SEM coefficient		SAR coefficient	
intercept	2.577***	0.0051	2.55***	0.005	2.56***	0.0095
empl	0.0003***	0.000096	0.000031*	0.00011	0.00031***	0.00011
prod	1.67***	0.0117	1.67***	0.0118	1.673***	0.0118
constr	3.15***	0.01013	3.15***	0.0102	3.149***	0.0102
serv	6.24***	0.0049	6.25***	0.0050	6.246***	0.0050
dist	-0.008***	0.000069	-0.00957***	0.00007	-0.0084***	0.00070
lambda	---	---	0.85391***	---	---	---
rho	---	---	---	---	0.126***	---
AIC	41532	---	-30301.51	---	30491.1	---

There are a few remarks on the sample size Firstly, big data is not unequivocal in terms of the size of the dataset, and it can be defined as "*data which exceed(s) the capacity or capability of current or conventional methods and systems*" (Ward & Barker, 2013). Gandomi and Heider (2015) write about big data volume as "*multiple terabytes and petabytes*". However, in spatial estimation, most of the routines stop at ca. *n<70,000* spatial units when applied on standard computers (see Arbia et al. 2019). Secondly, when testing the statistical solutions to proxy the full sample, one needs an operationalisable dataset to compare the subsample and full sample results. In the light of recent works by Arbia et al. (2019), ca. 37,000 observations is a safe and substantial spatial dataset. Thirdly, dealing with big data requires progress in the computational power of computers and smart statistical solutions. With the flood of mass spatial data (as mobile data, business locations, housing transactions, selling points etc.), one can easily imagine the more extensive dataset, which is bigger than the biggest operationalisable even for super-computers.

1. **Simple sampling and computation time**

Estimating the model on a full sample can be costly with regard to computational time, required computer memory and necessary effort for dataset collection. Sampling solutions understood as single estimation of a model on a subset can give acceptable approximations of full-sample results at a much lower cost. Subsample regression coefficients are expected to hold the full-sample estimates' values, while the main difference appears in their standard errors (SE). The big sample property of estimators saying that when expanding the size of the sample, the accuracy of the estimation rises, and especially when doubling the sample,

its variance decreases $\sqrt{2}$ times, is well-proven in the literature (e.g. Lenth, 2001). This rule can be expressed as equivalent to $10 \bullet n^{-0.5}$ for n observations. The efficiency of the variance estimator for the expanding window estimation (or the inverted *jack-knife*), $Eff_{exp.window}$, with step s, for the observations from 1 to h s taken from the total sample of n observations is:

$$\mathrm{Eff}_{\mathrm{exp.window}} = \frac{\mathrm{var}_{1:(h \cdot s)}}{\mathrm{var}_{1:n}} \quad (5)$$

where h is the number of steps, s is a length of the step, where h s=n, $\mathrm{var}_{1:(h \cdot s)}$ and $\mathrm{var}_{1:n}$ are the variances of the estimator in a subsample and the full sample, respectively. This allows for predicting the SE in a full sample, using a subsample estimate of SE. It shows how much one can cut the full sample to obtain reasonable subsample results. The expanding window procedure, with the step of the length of *s=100* observations and *h=1,2,3,..., 370* steps was applied on the randomly resorted data in a full dataset. In this analysis model specification was estimated 370 times on increasing dataset (100, 200, 300, ..., 1000, ..., 10000, ..., 37000 obs.).

Fig.3 illustrates beta coefficient and its SE for *prod* variable in OLS, SEM, SAR and SDM. Beta coefficients for subsamples hold their central tendency towards the full-sample estimate in all models, even if an over-bias of OLS compared with spatial models is persistent. SE of coefficients were drawn twice: as empirical SE - values of beta SE in specified models, and as theoretical SE - recalculated beta SE using "$\sqrt{2}$ rule" based on beta SE for 100 observations. Empirical and theoretical values follow the same pattern what confirms the existence of the"$\sqrt{2}$ rule". The efficiency of SE estimates $Eff_{exp.window}$ at 10,000, 20,000, and 30,000 observations equals ca. 1.98, 1.36 and 1.12 respectively in all models (what means that SE in a sample of 10.000 is ca. 2 times bigger than in full sample). Thus, increasing the sample size may approximate the subsample SE to full-sample SE, and only the subsample over *30,000* observations gives a relatively slight increase in this efficiency.

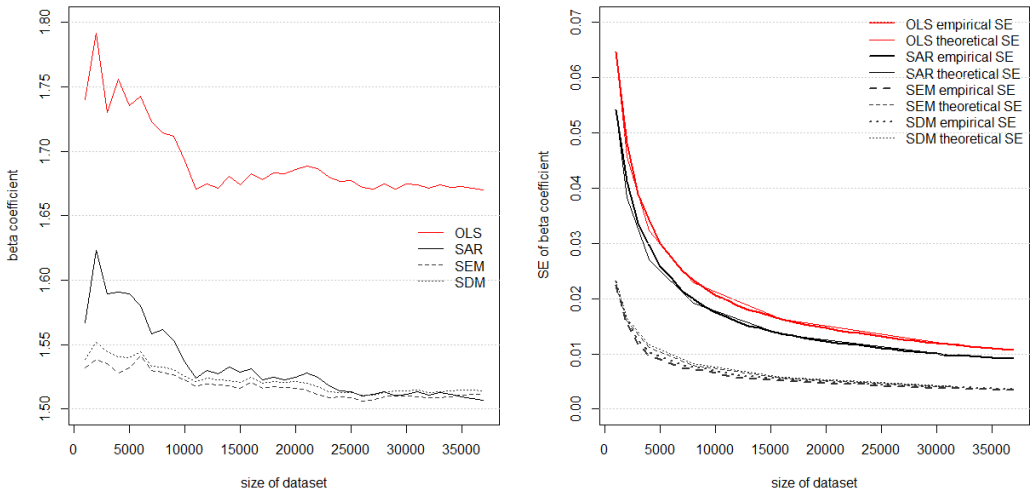


Figure 3: OLS, SEM, SAR and SDM estimation in the expanding window: a) beta coefficient, b) empirical and theoretical standard errors (SE)

Comparing the time cost of the calculations (Fig.4) shows that OLS estimation takes from 0.002 sec for 100 obs. to 0.01 sec for 37,000 obs. and irregularities in computation time result from machine RAM management. Spatial estimations last much longer, from 3.8 sec for 1000 obs. to 40 sec for 37,000 obs. Computation time increases linearly in OLS, while its expansion in spatial models is multinomial quadratic. Spatial calculations last from 3,100 times (SAR) to 4,100 times (SDM) longer than OLS.

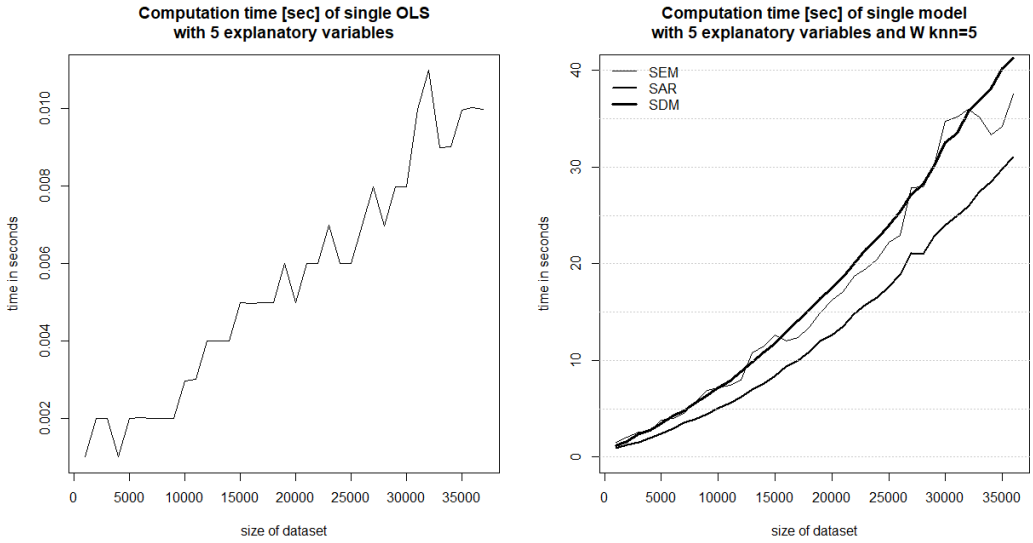


Figure 4: Computation time [in sec] in expanding data window: a) for OLS, b) for SEM, SAR and SDM

The above analysis shows that spatial estimation on thousands of geo-located observations is still very time consuming, if feasible at all, even when using optimised routines. Computer scientists offer technical solutions: super-computers, cloud service, parallel computations etc., to solve the technical problem of estimation. However, the massive inflow of individual spatial data causes bigger data sets than currently served to become available; thus, technical problems may remain unsolved. As in this paper, a statistical approach is to overcome, in general, the obstacles of dealing with big spatial data. This is to underline that simple sampling (in fact cutting the dataset) may not give satisfactory results because of representativeness issue and have a limited possibility of using the calibrated spatial model in forecasting for new geo-point, as it would result in a new spatial weights matrix W, which could destroy the fine-tuned calibration. There is a need for a smart solution to deal with computational problems in big data, lack of unique W when sampling and difficulty of forecasting for out-of-sample data to benefit from the information available. Thus the paper develops a new methodological solution based on resampling, which is step-by-step tested throughout the paper.

1. **Challenges of bootstrapped spatial regression**

The bootstrap's original purpose was to support the estimation on a small sample when the limited availability of data was the main obstacle in getting reasonable results. In big data, the availability of mass data redesigns bootstrap motivations, which may become the way of limiting, not multiplying the data. Before running the resampling procedure, which is to run many regressions on different subsamples, one should understand how this design works to answer

fundamental questions: how many iterations to use, what should be the size of subsample, does the location of selected observations matter, are the estimates of good quality. This Section clarifies those issues.

Following Fox (2015), the coefficient's standard bootstrap error is the standard deviation of the bootstrap coefficient replicates. Thus one can define the efficiency Eff_{boot} of the bootstrapped estimation as:

$$\text{Eff}_{\text{boot}} = \frac{var(\beta_{i,\ bootstrap})}{\text{var}_{\text{full sample}}} = f(s,i) \quad (6)$$

where *s=1,2,…,S* is the size of subsample in a bootstrap (number of observations drawn from a full sample), *i=1,2,…R* is a number of the iterations (number of replications of drawings), $var(\beta_{i,\ bootstrap})$ is a variance of estimate in an *i*-iterated set and $\text{var}_{\text{full sample}}$ is the known variance of estimator derived for a full sample or population.

There are at least three issues on the spatial sampling: i) the effective size of the spatial sample; ii) the sampling design; iii) the number of replications and the subsample's size in the bootstrap. Even though the literature seems to be rich, there is no clear answer to those issues.

The sample's effective size was studied mostly from a non-bootstrap perspective (e.g. Griffith & Zhang, 1999). Griffith (2008) builds a rule of thumb that ca. *n=400* observations may be sufficient in a spatial regression with a single covariate. Chernick and LaBudde (2014), following Hall (1985) indicate, that bootstrap can work for *n=20* observations if there are *i=2,000* replications.

The sampling design meets many recommendations. The literature usually applies parametric bootstrapping, residuals bootstrapping or observations bootstrapping (e.g. Tran & Tran, 2016) as they differ in assumptions, advantages, and disadvantages (Davison & Hinkley, 1997; Moulton & Zeger, 1991). For spatial data, the spatial dimension is of great importance. Franco-Villoria and Ignaccolo (2017) follow Lahiri (2003), that "*a bootstrap procedure needs to mimic the data generating mechanism to reproduce the spatial dependence structure in the bootstrap samples*", thus they recommend non-uniform sampling schemes. This is contrary to Davison et al. (2003), who claim that "*The ordinary nonparametric bootstrap uses uniform resampling from a data sample to mimic the mechanism that originally produced that sample.*" Alternatively, Griffith (2005) proves that for the single-draw sampling designs in the case of spatial autocorrelation, the hexagonal tessellation stratified sampling design performs the best. Since Hall (1985) suggested the blocked bootstrap for spatial dependent data, there is a discussion on the size of blocks (e.g. Hall et al.,1995; Nordman et al., 2007), shape of the blocks (e.g. Lahiri, 2003, Roberts et al., 2017), if they should overlap (Kunsch, 1989, Carlstein, 1986) or comparisons are presented (Radovanov & Marcikić, 2014). For bias or variance estimation Hall et al.

(1995) recommend $n^{1/3}$ blocks. Chernick and LaBudde (2014, p.148), following Lahiri (2003), indicate that "*bootstrap estimates under irregularly space grids are consistent*". One popular method of selecting irregular non-overlapping partitions is *k-means* clustering of spatial coordinates (e.g. Russ & Brenning, 2010, Schratz et al., 2019), which divides spatial points into spatially homogenous k clusters and runs spatially stratified sampling from those clusters. Irregular windows in bootstrap are also used by Kraamwinkel et al. (2018). This paper applies k-means clusters for a stratified sampling of the subsample (Fig.5), keeping a number of clusters between 50 and 100.

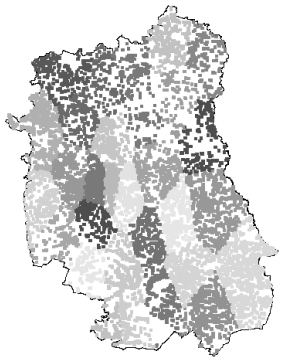


Figure 5: K-means non-overlapping clustering of points for stratified sampling

Bootstrap replications and its subsample size are usually recommended as arbitrary. Harris et al. (2017) note that usually increasing the number of replications is easier than raising the sample size. They indicate that usually there are *i=999* replications, while in their study, because of the computational burden, they use *i=99*, as the sampling design is simple. Escanciano and Jacho-Chávez (2012), in bootstrapped regression, assume 300 replications, Hall et al. (1995) take 200 replications, Efron and Tibshirani (1997) only 50 replications based on the internal variance calculations, while Fox (2015) and Tran and Tran (2016) apply 2,000 replications. Hesterberg (2015, p.380) insists on "*1000 bootstrap samples for rough approximations, or 10^4 or more for better accuracy*".

In this paper we test the bootstrapped OLS regression efficiency, depending on subsample size s and number of replications *i,* in a simulation on 37,000 observations. OLS models for a continuous dependent variable with *m=5* explanatory variables were estimated on the expanding by 1,000 number of observations, starting with 1,000 (*s=1,000, 2,000, 3,000, …, 37,000*). For each s, *i=2,000* replications were performed, sampled with replacement from 37,000 full sample.

The distributions of the coefficients and the standard errors were derived by sampling within 2,000 models a subsample i, which is equivalent to a number of replications (*i=100,200, 300, …, 2,000*) (See Fig.6).

In bootstrapping the regression coefficients, $_{m}$ are well-converging towards the population parameter $_{m}^{*}$, with the efficiency 1 for most of the scenarios (Fig.5d). The variance of the estimators is almost independent of a number of iterations ($\boldsymbol{i}$ on the x-axis), and strongly dependents on a size of a subsample (the consequent lines and $\boldsymbol{s}$ on the y-axis) (Fig.5b). The regression coefficients' bootstrap variance is lower than the expected (theoretical) one from a single estimation as in section 2 (Fig.5a) and reaches the efficiency ~ 1 at ca. 22,000 obs. For a big sample (ca.30,000) it is lower by 90% than the theoretical one. The estimated coefficients' accuracy depends on sample size and symmetrically converges towards the full-sample parameter (Fig.6d).

An efficient resampling design allows for avoiding the bias in the bootstrapped OLS coefficients. The skewness of β_j distributions was *Skew=0±0.15*, which confirms no asymmetry in β_j distributions, and consequently no bias. Also, a symmetry of the upper and lower part of the boxplot (Fig.6c), as well as the symmetric and centred distributions of betas for different sample sizes (Fig.6d), confirm no bias in the estimation process.

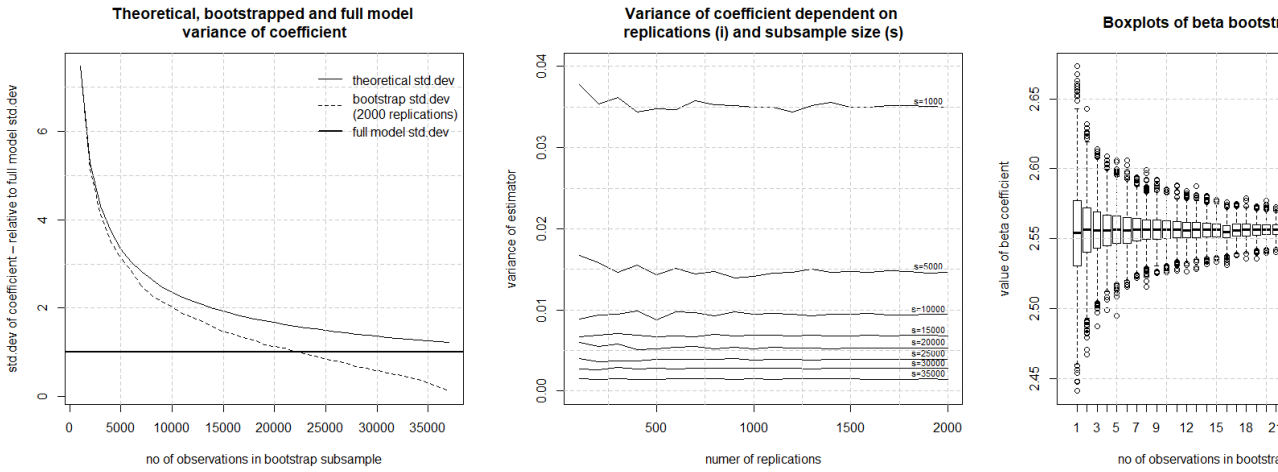


Figure 6: Efficiency of the bootstrapped OLS regression

The above general analysis for OLS confirms that the bootstrap is an efficient and convenient way to obtain the full-sample-like estimates at a low estimation cost. The bootstrap models on a moderate-size sub-sample are achievable with standard software and hardware and provide stable and high-quality results. This analysis confirms that bootstrapped spatial models behave as non-spatial ones (OLS). The comparison of the spatial and non-spatial models was run on a subsamples of *s=1,000, s=2,000, s=4,000* and *s=8,000* observations boot-

strapped *i=500* times (Fig.7). Beta coefficient (Fig.7a) and spatial *rho* from SDM (Fig.6c) are more precisely estimated with an increase of sample size, and the standard error decreases by $\sqrt{2}$ when doubling the sample size (Fig.7b). AIC, because of its dependence on sample size, cannot be used for setting proper sample size, but its variance diminishes with the increase of sample size. This confirms that the behaviour of bootstrapped coefficients and variances are similar in spatial and a-spatial models. One can confirm the above OLS conclusions for spatial models that bootstrapping appears as an attractive way of approximating the estimation on a full sample. At the same time, data characteristics are decisive for spatial and a-spatial specifications. The above shows that one can consciously choose bootstrap parameters – the final model will use 500 iterations (resamplings) of 8000 observations drawn randomly from the full sample.

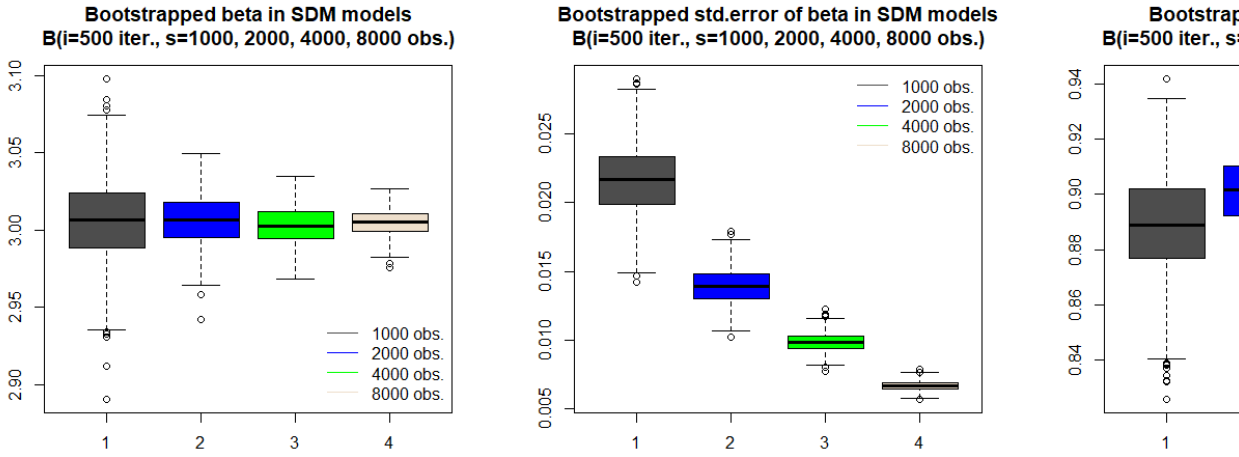


Figure 7: Parameters of bootstrapped SDM: a) beta, b) standard error, c) rho, d) AIC

1. **Selection of the best model and best data representation**

The bootstrapped regression estimation procedure with the parameters $B(i,s,)$ generates the i sets of the regression coefficients. Presented example runs i=500 bootstrapped models on s=8,000 observations. A model with a constant term, estimated with *m=5* explanatory variables in *i=500* replications, gives a matrix of size $[i \times (m+1+1)]$=[*500x7*], while after inclusion of spatial parameters in SDM $[i \times (m+m+2)]$=[*500x12*]. All i scenarios (iterations, resamplings) are estimated on the s-item subsamples of the randomly selected (but stored) observations. Selection of the best representation requires a multi-dimensional methodology as the distributions of individual $_m$ coefficients are not independent of each other. Also, information on the location of observations that gave the best model is necessary, as they will help to build W. This Section discusses how to choose the best model and check its quality.

This paper proposes a novel approach to choose the best model from many candidate bootstrap models – it uses *Partitioning Around Medoids* (PAM) algorithm used in unsupervised learning. In general, the clustering methods as PAM are designed to split the dataset into the most homogenous clusters with regard to many variables analysed. The partitioning procedure assumes finding the best combination(s) among the available ones to maximise the homogeneity within the group and to maximise the heterogeneity between the groups. Analysed points in PAM are multi-dimensional, and the applied distance metrics (e.g. Euclidean or Manhattan) defines the joint relations between them. For a set of bootstrap regression coefficients, a point is a set of coefficients from a single model, while the calculated distances between the points compare the pairs of models. To avoid the overflow of a single variable, the input data, and consequently the coefficients, should be standardised. The quality of partitioning (for *c>1*) is measured with silhouette or dissimilarity measures, computed for a given distance metric. The result of partitioning is twofold: firstly, one gets the *id* of the iteration set, which is the *medoid* of coefficients – the most typical set of coefficients; and secondly, all iterations are assigned to a given cluster. More details on PAM and silhouette in Appendix.

There are few possible number of clusters *c*. With *c=1* medoids, all possible observations are within the same cluster, while the best points representation is single. With *c=2* or more medoids, a double, triple or bigger set of "the best coefficients" is obtained. This can be interpreted as modelling in groups, and then the differentiating criteria are to be found. However, it needs a solid analysis of the significant differences among the groups. This method can also sort out the outlier scenarios, especially those located on the edge of clusters. The efficient estimation will support the *c=1* partitioning, which simplifies the procedure of finding the best (most representative) sampling combination f_y^*. In fact, for *c=1*, the medoid model is the one with the least sum of (Euclidean) distances to other models.

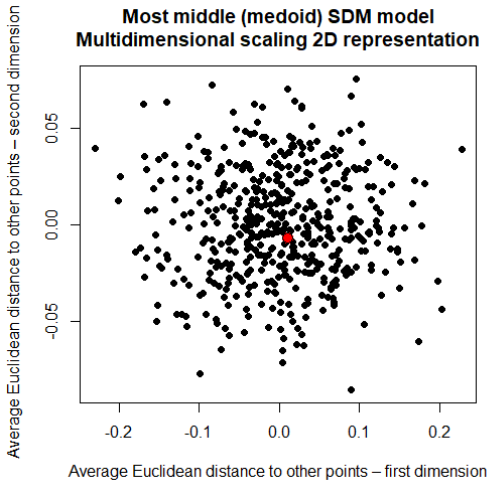
When assuming a single cluster only, there is a need to check the data's clustering tendency into more partitions. The *Hopkins statistics* tests if the data is clusterable (H1) when h~0 (h<0.5). It compares the total distances between the closest neighbours real pairs of points (w_j^d) and closest neighbours real point and uniformly randomly distributed points u_j^d:

$$H = \frac{\sum_{j=1}^{n} u_j^d}{\sum_{j=1}^{n} u_j^d + \sum_{j=1}^{n} w_j^d} \quad (7)$$

where $\sum_{j=1}^{n} u_j^d$ is the average distance to nearest neighbour between real point and uniformly generated random point (with the same variance as the real data) and $\sum_{j=1}^{n} w_j^d$ is the average distance to the nearest neighbour between the real data. As the statistic is based on the randomly generated data, thus the values of the statistics in iterations may differ. In the analysed example partitioning

into one single cluster is the optimal division, which is confirmed with Hopkins H statistics: H_{OLS}=0.21, H_{SEM}=*0.18*, H_{SAR}=0.19 and H_{SDM}=0.17.

For visualisation purposes, to see the medoid model's selection, one needs two-dimensional output, even when models are compared in i-dimensional space, where i is the number of iterations, and the result of comparing i models is $[i \times i]$ matrix). Multi-dimensional scaling (MDS) allows for dimension reduction from $[i \times i]$ to $[i \times 2]$. It keeps the scale of original values - here, the multi-dimensional Euclidean distance between pairs of the models (in fact, model coefficients) (see Fig.8). The medoid scenario, located centrally in clusters on Fig.8, yield the centred model coefficients.



Each point of scatterplot represents one (out of 500) model, medoid point (red) is the most representative model. 2D representation proxies distances between full set of regression coefficients in bootstrapped models.

Figure 8: 2D visualisation of multi-dimensional scaling of Euclidean distance between beta coefficients in bootstrapped SDM model

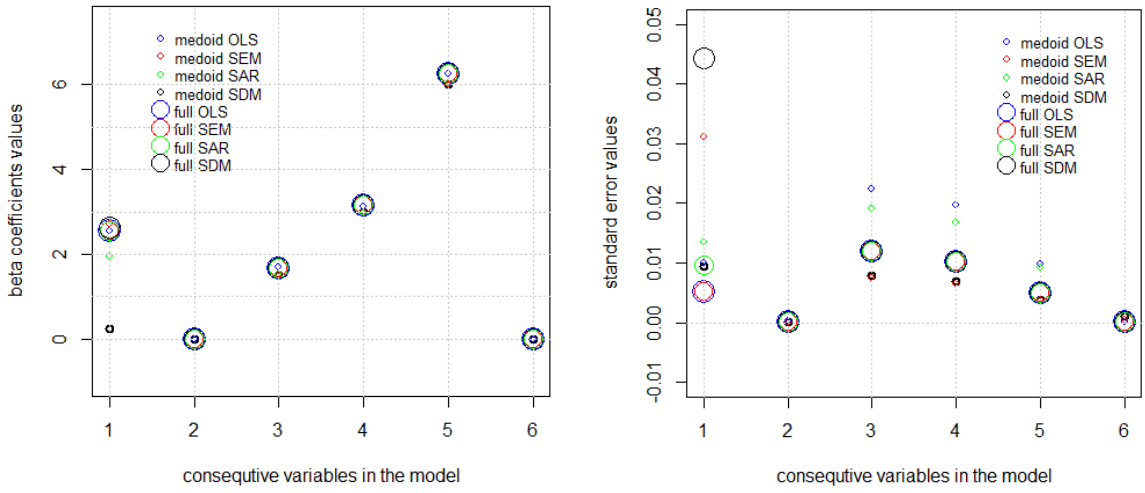
One should note that the model with the lowest possible AIC or BIC is not a representative model but an attractive outlier model. As the AIC and BIC are sensitive to outliers and increase with the bootstrap sample size, they cannot be used for deciding which sample size to use.

The quality of the bootstrapped model is measured in a typical way. Following Escanciano and Jacho-Chávez (2012), the performance of the bootstrapping estimators can be assessed with RAMSE (Root Average Mean Squared Error), denoted as:

$$RAMSE[\widehat{f_y}] = \sqrt{\tfrac{1}{i}} \sum_{i=1}^{R} AMSE[\widehat{f_y^{(i)}}], \text{ where AMSE} \left[\widehat{f_y^{(i)}}\right] = \frac{\sum_{j=1}^{n} [\widehat{f_y^i}(y_j) - y_j]^2}{n} \quad (8)$$

where i is the number of replications i=*1,2,...R*, $[\widehat{f_y}]$ is the overall prediction of the model, $[\widehat{f_y^{(i)}}]$ is the prediction of *i-th* replication, in particular $\widehat{f_y^i}(y_j)$ is the prediction of *i-th* replication for *j-th* observation, while y_j is the value of the dependent variable for *j-th* observation. In the analysed example, RAMSE of the medoid model was significantly lower in spatial SDM and SEM models ($\text{RAMSE}_{\text{SEM}}$=$\text{RAMSE}_{\text{SDM}}$=0.12) than in a-spatial models ($\text{RAMSE}_{\text{OLS}}$=0.31) and in spatial SAR model ($\text{RAMSE}_{\text{SAR}}$=0.27) (see Tab.3). This indicates the substantially higher quality and better fit of the spatial models.

Tab.2&3 and Fig.9 show the details of estimation and comparison of medoid models with average and full sample coefficients. Comparing the estimated beta coefficients (Fig.9a) clearly shows the upward bias of 5%-10% in OLS and relatively slight differences in betas between the full and medoid SEM, SDM and SAR models. Bootstrapped (average and medoid) coefficients replicate well the full-sample estimates. The coefficients and the standard errors of betas, differ mainly for the constant term (*variable 1*), but not for the rest of the variables (Fig.8a). Standard errors of estimates are as expected: the lowest in case of medoid SDM and SEM, then all full models, and later for medoid OLS and SAR (Fig.9b). All standard errors behave predictably and enable approximating the full-sample error with the "double size, $\sqrt{2}$ decrease in error" rule, while the most representative middle models' coefficients are as good as from the full models. The medoid models' efficiency is higher than the averaged models (columns 8-9 in Tab.2). Moran tests for the residuals' spatial autocorrelation proved that residuals from spatial models are random (Moran's I=-0.085, *p-value*=1).



Values on X-axis represent consecutive variables, with 1 being the constant term.

Figure 9: The comparison of the beta estimates in the OLS, SEM, SAR and SDM models: a) beta coefficients, b) standard errors of beta coefficients

The comparison of the medoid PAM-selected model with the full-sample and the average results prove that the PAM algorithm selects the most typical medoid representation, very similar to the average and full-sample results. The *z-test* for equality of coefficients from a full-sample and PAM selected medoid bootstrap regressions does not reject H_0 about the equality, which means no bias (columns 1-3 in Tab.2). The value-added of the PAM-selected medoid model over the averaged coefficients is that a) with PAM, one can replicate the subsample observations that support this typical pattern and enable the further analysis (tessellation), and b) that coefficients are considered jointly, not independently, as in case of the average.

Those comparisons suggest that the bootstrapped spatial models, especially SDM, replicate well the expected full-sample estimation using the limited subsample only. This indicates that bootstrapping together with PAM procedures are efficient tools to support the big data spatial econometrics. The bootstrap estimates are as they were obtained from a full sample; their SE can be rescaled by constant factor resulting from "$\sqrt{2}$ rule", while the computational effort is incomparably lower and the spatial models are feasible. The obtained medoid coefficients can be used to calibrate the econometric model and spatial weights matrix.

1. **Tessellation as a method of space calibration**

The medoid combination of the regression coefficients selected with the PAM algorithm above needs to be extended for a spatial dimension. The selected observations, which were used in a medoid bootstrap model, are treated as the best representation of the full sample, both regarding the values (z) as well as in terms of a location (x,y). The issue here is about the representation of spatial point distribution. By analogy to the econometric model's calibration by finding the best point estimates of regression coefficients, the calibration of space is needed to obtain one and universal W necessary for a spatial estimation.

In the approach where the point data are aggregated within the polygons along the administrative borders, the continuous polygonal representation of spatial pattern exists but follows the aggregation and the MAUP problems. It also lowers the accuracy of spatial information. The sample representation of continuous spatial data was developed well in geostatistics as a point pattern. There are a few available methods as kriging or thin-spline etc. (e.g. Chun & Griffith, 2013). However, they are primarily single-dimensional and far from mimicking the spatial weights matrix W.

This paper proposes using the Voronoi polygons (also called *Dirichlet tessellation* or *Thiessen Polygons*) as a method of a discrete polygonal representation of continuous spatial data based on a sample point data. In general, the tessellation constructs the polygons around the points by delimiting the points

in half of the distance between the points. The irregular shape tiles cover the whole area of a region in a continuous and non-overlapping manner. Each tile contains one point, and each point is assigned to one tile. For real geo-locations, a list of the tessellation tiles may be shorter than a number of observations in case of the overlapping locations. If locations of two points are the same, they are assigned to a single and unique tile. Spatial weights matrix W in defining the neighbourhood can accept the same neighbourhood information doubled in rows for the overlapping points. Some solutions can be jittering of locations by shifting location by a small value of epsilon.

The geo-locations of the observations that were used in the estimation of the best Medoid model serve as the best representation of the spatial point pattern. The tessellation transforms the point pattern into the continuous polygons set (Fig.10). This natural tessellation replicates well the point pattern of underlying location data. This approach to space delimitation and point data aggregation does not suffer from the MAUP. Many studies from the last 40 years (e.g. Sibson, 1980) confirm that tessellation is an attractive method for data analysis because of its flexibility. Ahuja (1982) indicates the great potential of the Voronoi polygons, which "*possess intuitively appealing characteristics, as would be expected from the neighborhood of a point*", while Halls et al. (2001) confirm the tessellation can be used efficiently "*in determining polygonal neighbourhood relationships between point locations*". There is exhaustive literature on the properties of the Dirichlet tessellation (e.g. Hinde & Miles, 1980; Du et al., 1999), which generally sees mostly the favourable features of this method.

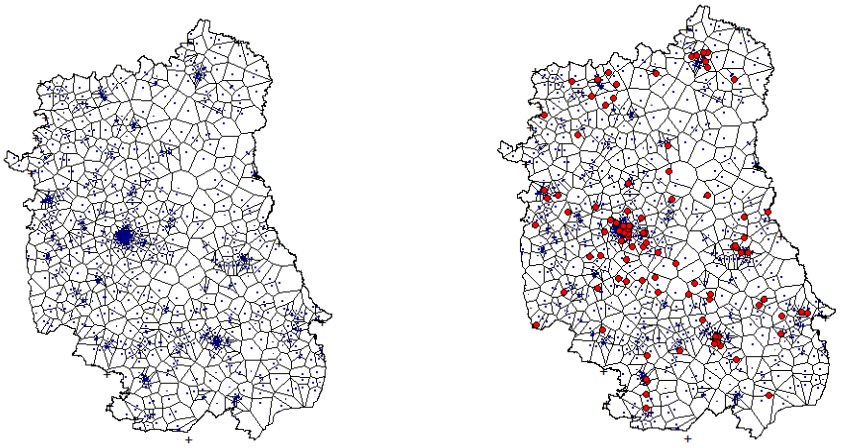


Figure 10: Voronoi tessellation of space for data underlying the medoid combination of coefficients: a) for s=2,000 observations, b) with new geo-locations (in red)

Recent developments started to link the tessellation to the bootstrap. Sec-

chi et al. (2013) use the bootstrap and the tessellation to reduce the original dataset and find the neighbourhood's local representatives. Their approach is statistically and computationally efficient, even if they use it in the context of clustering the functional data, not spatial regression. More on tessellation in bootstrapping in Appendix.

1. **Forecasting for out-of-sample geo-located points**

Spatial forecasting seems to have more challenges than available solutions. For in-sample predictions (the same training and testing dataset) as well as for previsions (the same spatial units, different values of variables), one can use the well-known trend–signal–noise predictor (Cressie, 1993), which is suitable for regional data only. For out-of-sample data (new locations, new values of variables) Goulard et al. (2017) overviews the methods and propose the algorithms based on kriging. Jiang (2018) similarly revises existing spatial predictions methods for regional data and indicates challenges for other data types. Popular kriging used for predictions in point patterns is not suitable for big data, and its forecast, even in regression models, are based on smoothing of the surface of the target variable. Zhu et al. (2018) propose using the Third Law of Geography, which assumes that the similarity of characteristics of two locations is reflected in a similarity of target variable at analysed points. The similarity is used as a weight in the prediction of the target variable in a given point. When using the spatial econometrics approach, the problem in forecasting with the spatial micro-econometric model for the new location of the out-of-sample point *(x*,y*)*, is that it is not a part of the estimated W in a final model, while it must be linked to this W. Simple recalculation of W in case of a new point added, would decalibrate the model and potentially disturb the result, as the medoid combination would lose its properties of the best representation. Kriging, which is often used, serves as a good approximating method, while it also has profound limitations with regard to the sample size (van Stein, 2015). Thus, the alternative solution proposed here is to use the tessellation (Voronoi polygons) to calibrate the space, assign new points to tiles via the "overlying" procedure and, consequently, link them to the already calibrated W (see Fig.10b).

An idea lies in the controlled imputation. In the analysed example, the final W is *8,000 x 8,000* and tessellation tiles t were enumerated as observations. Let's assume that a new geo-point $(x_{new},\ y_{new})=(x^*,y^*)$ was assigned to tile no.17 (t_{17} out of $t=8{,}000$). As the new observation also contains the values of the explanatory variables (z^*), it can be imputed to the dataset in place of observation no. 17, which automatically defines both a neighbourhood and a place in W. With this replacement, the calibrated model can be easily used to forecast the value of the dependent variable of a new point. The most important feature of the spatial model – the characteristics of the neighbourhood – stay unchanged. Thus the information necessary for computing the spatial lags is available.

There are few methods of internal validation of models: simple validation by splitting the population into learning and testing datasets; cross-validation (as

the K-fold validation), where the dataset is divided into K subsets and in K trials k-th subset is treated as a testing sample, while the other remaining as learning samples; and bootstrap resampling, where the resampling parameters of the model are used in building their confidence intervals. Tran & Tran (2016), following Kuhn and Johnson (2013), as well as Molinaro et al. (2005) and Steyerberg et al. (2001, 2003), indicate that the bootstrap resampling is more efficient than simple and cross-validation. This paper applies a complex approach - validated bootstrap resampling: the regression coefficients selected with PAM from a series of the bootstrapped models are in fact the central values in confidence intervals of the parameters. The observations which were used in the final medoid model are treated as a learning sample, while the others constitute "out-of-sample" training sample, from which the test locations are drawn.

The forecast quality can be measured with RAMSE (*Root average mean squared error*), introduced in (6) as a measure of model quality. For out-of-sample j observations from training data (x^*,y^*,z^*), modified RAMSE compares the forecasted $\hat{y}$ with the observed y^* as follows:

$$RAMSE[\widehat{f_y^*}] = \sqrt{\frac{\sum_{j=1}^{n} [\widehat{f_y^*}(y_j) - y_j^*]^2}{j}} \quad (9)$$

where f_y^* is the calibrated model - best model selected with PAM. In this approach, RAMSE includes only out-of-sample observations and their forecast, while the observations used in calibration were omitted in calculating RAMSE.

For the analysed dataset, the quality check of four models calibrated on *s=8,000* obs. was performed on out-of-sample *j=100* obs. The new observations were introduced to the calibrated model step-wise, one observation per check, to keep as much original data in the model as possible. For each j, the respective tessellation tail of point location was indicated, and original data for this tile were removed. Thus, the theoretical values $\hat{y}_j$ were calculated with *n-1* original observations from *in-sample* dataset (x,y,z) and single *out-of-sample* observation (x^*,y^*,z^*).

The goodness-of-fit of the models itself on training data is significantly better in SDM and SEM models ($\text{RAMSE}_{\text{SEM}}$=$\text{RAMSE}_{\text{SDM}}$=0.12) and a few times lower than in OLS ($\text{RAMSE}_{\text{OLS}}$=0.31) and in SAR ($\text{RAMSE}_{\text{SAR}}$=0.27). In the forecasts for out-of-sample data, $\text{RAMSE}_{\text{SAR}}$=0.508 and $\text{RAMSE}_{\text{SDM}}$=0.506 outperform $\text{RAMSE}_{\text{OLS}}$=0.527 and $\text{RAMSE}_{\text{SEM}}$=0.542. For the dependent variable *roa* ranged [1,10], $\text{RAMSE}_{\text{SAR}}$ of ca.0.5 obtained in the bootstrap models is an attractive result - on average 5% error of forecast.

This suggests that bootstrapped SDM model performed the best in terms of RAMSE for model and forecasts, with all variables significant and lowest standard errors of estimates. Bootstrapped SDM is also far better than boot-

strapped OLS and slightly better than other spatial bootstrapped models. Full sample SDM gives similar RAMSE as bootstrapped SDM, both for a model ($\text{RAMSE}_{\text{SDM full}}$=0.137) as well as for *j=100* forecasts ($\text{RAMSE}_{\text{SDM full}}$=0.555).

This indicates two critical facts. Firstly, spatial estimation matters for the quality of fit and spatial information reflected in the spatial weights matrix. The value-added of using spatial models lies in the additional information on spillovers that one can get, controlling for spatial autocorrelation and using the specific neighbourhood information. Secondly, the proposed methodology of calibrated tessellation and using a medoid model from bootstrapped sampled alternatives is efficient and reliable for forecasting out-of-sample data at a relatively low technical cost. Tessellation can be efficiently calculated for thousands of points (Fig.10b).

1. **Discussion of results**

This novel multistep estimation procedure, designed for dealing with forecasting for out-of-sample with spatial big-data models, is a mixture of traditional econometrics, bootstrapping and machine learning. All its elements are well-positioned in the literature. However, this combination was never proposed. Treating 37,000 observations as big data has an only illustrative purpose: to compare procedure and full-sample results and reliably asses the proposed solution. All tests presented above confirm the power of the bootstrap approach in improving estimates. Presented OLS, SAR, SEM and SDM models were selected arbitrarily as a case study, and this procedure works for any model.

The paper outlines the idea of the approach and shows its high quality. When applying the solution to other datasets, the researcher has to choose bootstrapping parameters: the number of observations taken as a subsample, the number of bootstrap iterations, and the number of k-means clusters to sample data from irregular shapes. These decisions are the function of available computing capacity and time. The analysis above showed that doubling the dataset doubles the computation time (Fig.4b). Using standard PC with Windows and R, simulation of a single spatial model on 5,000 observations, with five variables and *knn=5* W matrix iterated 500 times takes ca. 500(times)*2.5(sec)=20 minutes. The recommended solution in this paper is to choose at least ca.5,000 observations and 500 replications. The bigger the subset, the more precise results, but obtained in a longer time. Clustering procedures with *k*-means or PAM, or CLARA may seem efficient for big data. However, they are sensitive to a high number of clusters. The recommendation is to apply not more than 50-100 clusters.

Thus, the user of this multistep procedure is to: i) prepare the data, ii) decide about the bootstrapping parameters (min. s=5000 observations and i=500 iterations), iii) divide sample into training (e.g. 90%) and test (e.g. 10%) data, iv) cluster with k-means (eg. k=100) the geo-coordinates to run sampling of locations from irregular shapes, v) estimate in a loop i times the desired model (for a given variable specification and type of model), vi) select with PAM the

best medoid model out of a set of i models, vii) tessellate the space with observations from best PAM-selected model, viii) overlay the new out-of-sample points on the tessellated surface to assign points to tiles, ix) run the controlled imputation in test dataset by replacing single observation data with new point data, x) forecast value with *n-1* old records and *1* new record. The estimates obtained in this way are non-biased, efficient and guarantee low RAMSE.

This study is within the framework of spatial autocorrelation solutions, which in taxonomy by Jiang (2018) is next to spatial heterogeneity, limited ground truth, and multiple spatial scales and resolutions. As Jiang (2018) shown, there is a customary attitude to use spatial econometrics to areal data and geostatistics (as kriging) or GWR to point data. The trend of using spatial econometrics to point data is increasing (Arbia et al., 2021; Abruzzo et al., 2021; Piacentino et al., 2021; Santi et al., 2021), but due to discussed in paper limitations, not exploited. The presented solution may push those studies forward. An important aspect are the alternative solutions to the proposed one. For problems of spatial heterogeneity, limited ground truth, and multiple spatial scales and resolutions innovations slowly appear – as modifications of kriging and GWR, decomposition-based ensemble, multi-task learning, semi-supervised learning, active learning etc. However, as shown in Jiang (2018) there are more challenges than ready-to-use solutions, especially for spatio-temporal models, for anisotriopic spatial dependency, and for big data. Spatial predictions need deepened interest and bringing new concepts which enable bypassing the obstacles identified until now.

1. **Conclusions**

Paper offers few methodological novelties and interesting solutions to econometric problems. Firstly, it develops complex approach in spatial microeconometric modelling, solving computational efficiency problems in case of big spatial data, lack of unique spatial weights matrix W when sampling, and difficulty of forecasting for out-of-sample data. Secondly, it introduces a new method for forecasting in the spatial models for the out-of-sample geo-located points by linking in pairs train and test location to substitute in W. Pairing uses tessellated surface to represent the generalised point pattern with Voronoi polygons and assigns new points by overlaying them on tessellation tiles. Third, it uses the bootstrap technique to shrink the dataset's size and find the most representative combination of sub-sample observations. Fourth, it introduces the double calibration concept in spatial models, calibrating both the regression coefficients in SAR, SEM and SDM, and the spatial weights matrix W. Fifth, it applies the Partitioning Around Medoids (PAM) algorithm in a joint analysis of the bootstrap regression coefficients to select the jointly most middle scenario and underlying observations – treated as the best representation of coefficients and spatial sampling.

The proposed solution provides a stable, efficient and feasible approach to estimate and calibrate typical (SAR, SEM and SDM) spatial econometric models on geo-referenced big data and to forecast for the new out-of-sample locations.

It significantly increases the computational efficiency of modelling. Bootstrap estimations yield accurate estimates with highly efficient errors and precise forecasts. This method can find its applications in all micro-econometric problems, especially in models for individual business locations and in real estate valuation models for datasets of hundreds of thousands or more of observations. In the case of large data sets or big data, the estimation on the whole dataset, even when feasible, is usually technically very demanding. Thus the bootstrap technique facilitates and simplifies the calculations without losing the accuracy.

Supplementary material: Information on: A) Spatial weights matrix, B) The R software in estimation of spatial models, C) Clustering with Partitioning Around Medoids (PAM) algorithm, D) Tesselation (Voronoi polygons), references

Funding: This is a part of a project on "*Spatial econometric models with fixed and changing structure of neighbourhood. Applications to real estate valuation and business location*" financed by National Science Center Poland (Krakow, Poland) [OPUS 12 call, grant number UMO-2016/23/B/HS4/02363].

Data Availability: The dataset and R codes to replicate the whole analysis are available at https://github.com/kkopczewska/bootstrapping.

References

Abbruzzo, A., Ferrante, M., & De Cantis, S. (2021). A pre-processing and network analysis of GPS tracking data. Spatial Economic Analysis, 16(2), 217–240.

Ahuja, N. (1982). Dot pattern processing using Voronoi neighborhoods. IEEE Transactions on Pattern Analysis and Machine Intelligence, (3), 336-343.

Arbia, G. (2014). A primer for spatial econometrics: with applications in R. Springer.

Arbia, G., Ghiringhelli, C., & Mira, A. (2019). Estimation of spatial econometric linear models with large datasets: How big can spatial Big Data be?. Regional Science and Urban Economics.

Arbia, G., Espa, G., & Giuliani, D. (2021). Spatial microeconometrics. Routledge.

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In Proceedings of the 25th international conference on Machine learning (pp. 33-40). ACM.

Barbian, M. H., Assunção, R. M. (2017). Spatial subsemble estimator for large geostatistical data. Spatial Statistics, 22, 68-88.

Carlstein, E. (1986). The use of subseries methods for estimating the variance of a general statistic from a stationary time series. Ann. Statist. 14, 1171-9.

Castillo-Páez, S., Fernández-Casal, R., & García-Soidán, P. (2020). Nonparametric bootstrap approach for unconditional risk mapping under heteroscedasticity. Spatial Statistics, 40, 100389.

Chernick, M. R., LaBudde, R. A. (2014). An introduction to bootstrap methods with applications to R. John Wiley & Sons.

Chun, Y., Griffith, D. A. (2013). Spatial statistics and geostatistics: theory and applications for geographic information science and technology. Sage.

Cressie, N. A. C. (1993). Statistics for spatial data, Wiley, New York

Dalposso, G. H., Uribe-Opazo, M. A., Johann, J. A., Bastiani, F. D., & Galea, M. (2019). Geostatistical modeling of soybean yield and soil chemical attributes using spatial bootstrap. Engenharia Agrícola, 39, 350-357.

Davison, A.C., Hinkley, D.V., (1997). Bootstrap Methods and Their Application. Cambridge University Press.

Davison, A. C., Hinkley, D. V., & Young, G. A. (2003). Recent developments in bootstrap methodology. Statistical Science, 141-157.

DiCiccio, T. J., Efron, B. (1996). Bootstrap confidence intervals. Statistical science, 189-212.

Dowd, B. E., Greene, W. H., Norton, E. C. (2014). Computation of standard errors. Health services research, 49(2), 731-750.

Du, Q., Faber, V., Gunzburger, M. (1999). Centroidal Voronoi tessellations: Applications and algorithms. SIAM review, 41(4), 637-676.

Efron, B., Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. Journal of the American Statistical Association, 92(438), 548-560.

Elhorst, J. P. (2010). Applied spatial econometrics: raising the bar. *Spatial Economic Analysis*, *5*(1), 9-28.

Escanciano, J. C., Jacho-Chávez, D. T. (2012). $\sqrt{n}$-uniformly consistent density estimation in nonparametric regression models. *Journal of Econometrics*, *167*(2), 305-316.

Fox, J. (2015). Applied regression analysis and generalised linear models. Sage Publications. Chapter 21 Bootstrapping Regression Models

Franco-Villoria, M., Ignaccolo, R. (2017). Bootstrap based uncertainty bands for prediction in functional kriging. Spatial Statistics, 21, 130-148.

Freedman, D. A. (1981). Bootstrapping regression models. The Annals of Statistics, 9(6), 1218-1228.

Gandomi, A., Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

García-Soidán, P., Menezes, R., & Rubiños, Ó. (2014). Bootstrap approaches for spatial data. Stochastic environmental research and risk assessment, 28(5), 1207-1219.

Goulard, M., Laurent, T., Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost optimal strategies. Spatial Economic Analysis, 12(2-3), 304-325.

Griffith, D. A. (2005). Effective geographic sample size in the presence of spatial autocorrelation. Annals of the Association of American Geographers, 95(4), 740-760.

Griffith, D. A. (2008). Geographic sampling of urban soils for contaminant mapping: how many samples and from where. Environmental geochemistry and health, 30(6), 495-509.

Griffith, D. A., Zhang, Z. (1999). Computational simplifications needed for efficient implementation of spatial statistical techniques in a GIS. Geographic Information Sciences, 5(2), 97-105.

Hall, P. (1985). Resampling a coverage process. Stoch. Proces. Applic. 20, 231-46.

Hall, P. (2013). The bootstrap and Edgeworth expansion. Springer Science & Business Media.

Hall, P., Horowitz, J. L., Jing, B. Y. (1995). On blocking rules for the bootstrap with dependent data. Biometrika, 82(3), 561-574.

Harris, P., Brunsdon, C., Lu, B., Nakaya, T., & Charlton, M. (2017). Introducing bootstrap methods to investigate coefficient non-stationarity in spatial regression models. Spatial Statistics, 21, 241-261.

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. The American Statistician, 69(4), 371-386.

Hinde, A. L., Miles, R. E. (1980). Monte Carlo estimates of the distributions of the random polygons of the Voronoi tessellation with respect to a Poisson process. Journal of Statistical Computation and Simulation, 10(3-4), 205-223.

Ibrahim, A. M., Bennett, B. (2014). The assessment of machine learning model performance for predicting alluvial deposits distribution. Procedia Computer Science, 36, 637-642.

Jiang, Z. (2018). A survey on spatial prediction methods. IEEE Transactions on Knowledge and Data Engineering, 31(9), 1645-1664.

Jin, F., & Lee, L. F. (2015). On the bootstrap for Moran's I test for spatial dependence. Journal of Econometrics, 184(2), 295-314.

Kelejian, H., Piras, G. (2017). Spatial econometrics. Academic Press.

Kopczewska, K. (2021). Spatial Machine Learning – New Opportunities for Regional Science, Working Paper at University of Warsaw, Faculty of Economic Sciences

Kraamwinkel, C., Fabris-Rotelli, I., & Stein, A. (2018). Bootstrap testing for first-order stationarity on irregular windows in spatial point patterns. Spatial statistics, 28, 194-215.

Kuhn, M., Johnson, K., (2013). Applied Predictive Modeling. Springer New York, New York, NY.

Kunsch, H.R.(1989). The jackknife and the bootstrap for general stationary observations. Ann.Statist.17, 1217-41.

Lahiri, S., 2003. Resampling Methods for Dependent Data. Springer-Verlag, New York.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. The American Statistician, 55(3), 187-193.

LeSage J.P., Pace, R.K. (2007), A Matrix Exponentials Spatial Specifications, Journal of Econometrics, 140:1.

Loh, J.M.(2008). A valid and fast spatial bootstrap for correlation functions.The Astrophysical Journal,681(1),726.

Manly, B. F. (2006). Randomisation, bootstrap and Monte Carlo methods in biology (Vol. 70). CRC press.

Molinaro, A.M., Simon, R., Pfeiffer, R.M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21 (15), 3301-3307.

Moulton, L.H., Zeger, S.L., (1991). Bootstrapping generalised linear models. Computational Statistics Data Analysis 11, 53-63.

Nordman, D. J., Lahiri, S. N., & Fridley, B. L. (2007). Optimal block size for variance estimation by a spatial block bootstrap method. Sankhyā: The Indian Journal of Statistics, 468-493.

Perdikaris, P., Venturi, D., Royset, J. O., & Karniadakis, G. E. (2015). Multi-fidelity modelling via recursive co-kriging and Gaussian–Markov random fields. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 471(2179), 20150018.

Piacentino, D., Aronica, M., Cracolici, M. F., Giuliani, D., & Mazzitelli, A. (2021). The effect of agglomeration economies and geography on the survival of accommodation businesses in Sicily. Spatial Economic Analysis, 16(2), 176–193. https://doi.org/10.1080/17421772.2020.1836389

Radovanov, B., & Marcikić, A. (2014). A comparison of four different block bootstrap methods. Croatian Operational Research Review, 5(2), 189-202.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... & Warton, D. I. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929.

Rosipal, R., Girolami, M., Trejo, L. J., & Cichocki, A. (2001). Kernel PCA for feature extraction and de-noising in nonlinear regression. Neural Computing & Applications, 10(3), 231-243.

Ruß, G., Brenning, A. (2010). Spatial variable importance assessment for yield prediction in precision agriculture. In International Symposium on Intelligent Data Analysis (pp. 184-195). Springer, Berlin, Heidelberg.

Santi, F., Dickson, M. M., Espa, G., Taufer, E., & Mazzitelli, A. (2021). Handling spatial dependence under unknown unit locations. Spatial Economic Analysis, 16(2), 194–216.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling, 406, 109-120.

Secchi, P., Vantini, S., Vitelli, V. (2013). Bagging Voronoi classifiers for clustering spatial functional data. International Journal of Applied Earth Observation and Geoinformation, 22, 53-64.

Sibson, R. (1980). The Dirichlet tessellation as an aid in data analysis. Scandinavian Journal of Statistics, 14-20.

Steyerberg, E.W., Bleeker, S.W., Moll, H.A., Grobbee, D.E., Moons, K.G., 2003. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J. Clin. Epidemiol. 56, 441-447.

Steyerberg, E.W., Harrell, F.E., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y., Habbema, J.D.F., (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J. Clin. Epidemiol. 54, 774-781.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

Tran, P., Tran, L. (2016). Validating negative binomial lyme disease regression model with bootstrap resampling. Environmental Modelling & Software, 82, 121-127.

van Stein, B., Wang, H., Kowalczyk, W., Bäck, T., Emmerich, M. (2015, October). Optimally weighted cluster kriging for big data regression. In International Symposium on Intelligent Data Analysis (pp. 310-321). Springer, Cham.

Uboldi, F., Sulis, A. N., Lussana, C., Cislaghi, M., & Russo, M. (2014). A spatial bootstrap technique for parameter estimation of rainfall annual maxima distribution. Hydrology and Earth System Sciences, 18(3), 981-995.

Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. arXiv preprint arXiv:1309.5821.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. the Annals of Statistics, 14(4), 1261-1295.

Ye, Z., Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. Journal of the American Statistical Association, 98(464), 968-979.

Zhu, A. X., Lu, G., Liu, J., Qin, C. Z., & Zhou, C. (2018). Spatial prediction based on Third Law of Geography. Annals of GIS, 24(4), 225-240.

Table 2: Results of bootstrapped *B(s=8,000 and i=500)* OLS, SEM, SAR and SDM regressions

B(i=500 iter, s=8,000 obs)	model	Coefficient in the single medoid model – bootstrap ALL SIGNIF	Average coefficient from models – bootstrap ALL SIGNIF	Coefficient in the full-sample model – ALL SIGNIF	Std.error in the medoid single model - bootstrap	Std.dev of beta in bootstrapped models	St.d.error in the full-sample models	efficiency ratio of the single medoid models	efficiency ratio of the average single model	efficiency ratio of the bootstrapped models
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)=(4)/(7)	(9)=(5)/(7)	(10)=(6)/(7)
Intercept	OLS	0.00009	0.0003	0.0003	0.0003	0.0002	0.0003	0.0001	3.0	2.0	2.7
Employment		1.71	1.68	1.67	0.02	0.02	0.026	0.012	1.9	1.9	2.2
		3.12	3.14	3.15	0.02	0.02	0.021	0.010	1.9	1.9	2.1
		6.27	6.25	6.24	0.009	0.009	0.01	0.005	1.9	1.9	2.0
Production		-0.008	-0.008	-0.008	0.0001	0.0001	0.0002	0.00007	1.9	1.9	2.4
Construction											
Service											
Distance											

B(i=500 iter, s=8,000 obs)	model	Coefficient in the single medoid model – bootstrap ALL SIGNIF	Average coefficient from models – bootstrap ALL SIGNIF	Coefficient in the full-sample model	Std. error in the medoid single-model - bootstrap	Std. error in bootstrapped models	Std. dev. of beta in bootstrapped models	Std. error in the full-sample models	efficiency ratio of the single medoid model	efficiency ratio of the average single model	efficiency ratio of the bootstrapped models
Intercept	SEM	0.0002	0.00002	0.0003	0.0001	0.00007	0.00009	0.0001	0.97	0.63	0.84
Employment		1.50	1.51	1.67	0.007	0.007	0.009	0.01	0.63	0.62	0.84
		2.99	3.00	3.15	0.006	0.006	0.008	0.01	0.64	0.62	0.80
		6.00	6.007	6.25	0.003	0.003	0.004	0.005	0.74	0.71	0.86
Production		-0.009	-0.009	-0.008	0.0004	0.0004	0.0003	0.00007	5.75	6.82	5.15
Construction											
Service											
Distance											
Intercept	SAR	0.0001	0.0002	0.0003	0.0003	0.0003	0.0002	0.0001	2.75	3.40	1.97
Employment		1.53	1.52	1.67	0.019	0.019	0.023	0.01	1.60	1.66	1.92
		2.97	2.99	3.15	0.016	0.017	0.018	0.01	1.63	1.65	1.76
		6.02	6.01	6.24	0.009	0.009	0.007	0.005	1.83	1.84	1.42
Production		-0.006	-0.006	-0.008	0.0001	0.0001	0.0002	0.00007	1.71	1.76	2.44
Construction											
Service											
Distance											

B(i=500 iter, s=8,000 obs)	model	Coefficient in the single medoid model – bootstrap ALL SIGNIF	Average coefficient from models – bootstrap ALL SIGNIF	Coefficient in the full-sample model ALL SIGNIF	std.error in the medoid single-model - bootstrap	std.error average in models - bootstrap	Std.dev. of beta in bootstrapped models	Std.error in the full-sample model	efficiency ratio of the single medoid model	efficiency ratio of the average single model	efficiency ratio of the bootstrapped models
Intercept	SDM	0.0001 / -	-0.00004 / -	0.0003 / -	0.0001 / -	0.00012 / -	0.00009 / -	0.0001 / -	2.99 / -	3.39 / -	3.53 / 0.39
		0.0005	1.51	-	0.008 / -	0.027 / -	0.029 / -	0.02 / -	0.85 / -	2.85 / -	3.28 / 0.28
Employment		1.51	/ -	0.0001	0.006 / -	0.026 / -	0.028 / -	0.02 / -	0.56 / -	3.35 / -	3.76 / 0.47
		/ -	1.33	1.67	0.003 / -	0.023 / -	0.024 / -	0.045 / -	0.74 / -	2.71 / -	3.36 / 0.44
Production		1.30	3.00	/	0.001 / -	0.001 / -	0.001 / -	0.0007 / -	1.01 / -	4.5 / -	3.8 / 6.75
		2.99	/ -	0.02							
Construction		/ -	2.69	3.15							
		2.65	6.01	/							
		6.01	/ -	0.06							
Service		/ -	5.38	6.24							
		5.33	-	/							
Distance		-	0.01	0.14							
		0.01	/	-							
		/	0.01	0.008							
		0.01		/-							
				0.0002							

General remark: in SDM models values for / , all variables listed in a table were significant at =0.01.

Columns: (1) values of coefficients in selected model; (2) the average coefficient of all bootstrapped models; (3) values of coefficients in full-sample model; (4) standard errors of beta coefficients in selected model; (5) the average standard error of all bootstrapped models; (6) the standard deviation of the bootstrap coefficient replicates; (7) standard error at coefficient in full-sample model; (8) std.error in medoid model compared to the standard error of coefficients for the OLS model on a full sample; (9) std.error of average model compared to the standard error of coefficients for the OLS model on a full sample; (10) std.error in bootstrap model compared to the standard error of coefficients for the OLS model on a full sample

Source: Own calculations

Table 3: Parameters of bootstrapped *B(s=8,000 and i=500)* OLS, SEM, SAR and SDM regressions

Parameters of estimation	OLS	SEM	SAR	SDM
AIC in	7893	-7660	5252	-7789
medoid	0.314	0.124	0.274	0.123
model	0.527	0.542	0.508	0.506
Average AIC	---	Lambda=0.915***	Rho=0.127***	Rho=0.899***
RAMSE in	---	Lambda=0.919***	Rho=0.124***	Rho=0.905***
medoid	---	Lambda=0.921***	Rho=0.126***	Rho=0.906***
model	0.0025	2.14	1.41	2.85
RAMSE of	1.28	1072	705	1423
forecast				
(j=100 obs)				
Spatial				
parameter in				
medoid				
model				
Average				
spatial				
parameter in				
bootstrapped				
models				
Spatial				
parameter in				
full sample				
model				
Computation				
time of				
average				
model (sec)				
Computation				
time (all 500				
models) (sec)				

Source: Own calculations

SUPPLEMENTARY MATERIAL

A Spatial weights matrix

Spatial weights matrix W for n observations is n x n matrix. It includes neigh-

bourhood information (1 if two observations are neighbours, and 0 if not) standardised by a total number of neighbours of a given observation. For the contiguity criterion, the neighbourhood is to share a common border, for k nearest neighbours, it is to belong to a group of k observations with the shortest distance. The inverse distance matrix assumes that all other observations are neighbours, and the strength of relations is the inverse distance between observations.

Spatial weights matrix $\boldsymbol{W}$ is the crucial and sensitive element of spatial estimation. The literature is not unequivocal about $\boldsymbol{W}$ selection, with studies being indifferent on $\boldsymbol{W}$ selection (e.g. LeSage & Pace, 2014) as well as insisting on the importance of $\boldsymbol{W}$ because of the potential bias of the coefficients or the direct and indirect impacts (e.g., Lee & Yu, 2012). Aside from the estimation quality, there is an issue of computational feasibility. The operational requirements of big point geo-located data requirements fairly limit the possibility of using the inverse squared distance $\boldsymbol{W}$ for all-to-all units, thus leaving the selection area mostly to k-neighbours $\boldsymbol{W}$. Contiguity $\boldsymbol{W}$ matrix is typically used for regional data and not considered for point data due to no border. This paper in all estimations uses *knn=5* nearest neighbours spatial weights matrix $\boldsymbol{W}$, made symmetric upward. Only one matrix is being used to limit the degree of complexity and concentrate on the paper's central issue. The selection of the matrix has no impact on the conclusions in this paper.

B The R software in estimation of spatial models

All estimations were programmed and run in R software and on standard PC computer, what is to express the availability of the models for average applied researchers. R software was used in estimation with packages *spdep* (Bivand & Piras, 2015), *sp* (Bivand et.al, 2013), *rgdal* (Bivand et al., 2017), *cluster* (Maechler et. al, 2017), *doBy* (Hojsgaard & Halekoh, 2016), *spatstat* (Baddeley et al., 2015). Spatial estimation in R can be speed up by using planar, non-spherical coordinates (option *lonlat=NULL* in *knearneigh()* function) and an alternative sparse matrix decomposition approach (option *method="LU"* in *errorsarlm()* function). More on coding techniques in spatial analysis can be found in a book by Kopczewska (2020).

C Clustering with Partitioning Around Medoids (PAM) algorithm

PAM algorithm clusters data around medoids which are the real points from the dataset. In general, there are two phases of PAM: first one, to select initial medoids and to calculate the cost function (BUILD) and the second one, to change medoids for other and to check the improvements (decrease) in the cost function (SWAP). The goal is to minimise the overall dissimilarity, measured with distance (e.g.Euclidean, Manhattan), between the representatives of each cluster and its members. The number of the clusters is assumed *apriori*. The cost function sums the distances between each point and its closest medoid. For two points $X = (x_1, x_2, x_3, \dots, x_n)$ and $Y = (y_1, y_2, y_3, \dots, y_n)$ the Euclidean distance is calculated as $\sum_{i=1}^{n} (x_i - y_i)^2$ and Manhattan distance is calculated as $\sum_{i=1}^{n} |x_i - y_i|$. PAM is a twin procedure to *k-means*, but the most important

difference is that *k-means* creates the centres of clusters not using real data but simply with any available values which optimise the function, while PAM selects out of existing scenarios. Thus, with k-means one cannot get the underlying set of observations, which is necessary to build spatial weights matrix W and tessellation procedure.

For quality check of clustering, most often, one uses the silhouette statistic. It can validate the consistency within the clusters. The individual silhouette $\boldsymbol{s_i}$ statistics (called also scaled individual separations) is expressed as $s_i = (b_i - a_i)/max(a_i, b_i)$, where a_i is an average distance to all other objects in the cluster, b_i is a minimum of the average distance to other clusters (cluster by cluster). The s_i statistics is limited s [-1,1]. Negative s_i is undesirable, as it means that $a_i > b_i$, so the other clusters are closer than "our" cluster. Oppositely, positive s_i is desirable. The best is $s_i \sim 1$, which appears when $a_i \sim 0$, which means that the distance in "our" cluster is heavily reduced. The global silhouette s is expressed as the arithmetic average of individual silhouette statistics s_i.

There are few potential alternatives for PAM. One is CLARA - big data equivalent of PAM, which applies sampling to increase the computational efficiency (while PAM operates iteratively on the whole dataset). However, the matrix of regression coefficients in this case is not big data; thus, PAM is sufficient. The other alternative is a multi-distribution joint copula function. This approach, however, requires advanced multi-dimensional optimisation to detect a kind of "optimum" for this surface, and the literature overview suggests it still needs development to act efficiently in this kind of application. The general remark is that unsupervised learning is still rarely linked to regression results, even if opposite applications appear (Pan et al., 2013). PAM algorithm, because of the random initial solution in some complex scenarios may fail (Bernábe-Loranca et al., 2014), however here, the homogeneity of input data prevents clustering problems.

D Tesselation (Voronoi polygons)

Voronoi polygons split the surface completely into non-overlapping tiles. As they are based on lines located halfway between points, they naturally follow the point pattern. The agglomeration of the points significantly increases the variance of the areas of tiles by decreasing the size of tiles in dense areas, and oppositely, enlarging tiles in peripheries. Thus, the tiles' size is inversely correlated with the probability of finding the observation in a given location. Also, the sample size selected in the bootstrapping procedure is reflected in the number of tiles. The more observations (and consequently tiles) in the regression model, the shorter the distance to the next neighbourhood and the more spatially integrated tiles.

The presented solution uses tessellation as a method of calibration of space. However, tessellation appears also in literature in connection to bootstrap. Herrera et al. (2013) represent the space with the continuous and non-overlapping blocks for *spatial block bootstrapping* (SBB). However, those blocks are the set

of n observations (not an individual one as in tessellation), formed around the arbitrarily defined fixed points and following the rule that each observation is assigned to the nearest point to form an n-long block. Herrera et al. (2013) use SBB and resampling to test the inter-independence of the spatial processes. They observe that the bootstrapped spatial ordering mismatches compared to the original distributions are negligible, and SBB represents the spatial structure well. This supports this paper's approach that Dirichlet tessellation around the points that were the observations in medoid-coefficients regression can be a robust subsample approximation of a full-sample point pattern. Most recent studies (Righetto et al., 2020) also link mesh (grid, tessellation) with estimation properties, while these studies are in their initial phase.

References of supplementary material

Baddeley A., Rubak, E., Turner R., (2015). Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press, 2015

Bernábe-Loranca, B., Gonzalez-Velázquez, R., Olivares-Benítez, E., Ruiz-Vanoye, J., & Martínez-Flores, J. (2014). Extensions to K-Medoids with Balance Restrictions over the Cardinality of the Partitions. Journal of applied research and technology, 12(3), 396-408.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2013). Applied spatial data analysis with R. Second Edition, New York: Springer.

Bivand, R., Keitt T., Rowlingson B., (2017). rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.2-16.

Bivand, R., Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. American Statistical Association.

Herrera, M., Ruiz, M., Mur, J. (2013). Detecting dependence between spatial processes. *Spatial Economic Analysis*, *8*(4), 469-497.

Hojsgaard S., Halekoh U., (2016). doBy: Groupwise Statistics, LSmeans, Linear Contrasts, Utilities. R package version 4.5-15.

Kopczewska K. (ed) (2020) Applied Spatial Statistics and Econometrics. Data Analysis in R, Routledge Advanced Texts in Economics and Finance

LeSage, J. P., Pace, R. K. (2014). The biggest myth in spatial econometrics. *Econometrics*, *2*(4), 217-249.

Lee, L.-f. and J. Yu (2012). Qml estimation of spatial dynamic panel data models with time varying spatial weights matrices. Spatial Economic Analysis 7 (1), 31–74.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2017). cluster: Cluster Analysis Basics and Extensions. R package version 2.0. 1. 2015.

Pan, W., Shen, X., Liu, B. (2013). Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. The Journal of Machine

Learning Research, 14(1), 1865-1889.

Righetto, A. J., Faes, C., Vandendijck, Y., & Ribeiro Jr, P. J. (2020). On the choice of the mesh for the analysis of geostatistical data using R-INLA. Communications in Statistics-Theory and Methods, 49(1), 203-220.