The importance of spatial sampling methods for their analysis results – selected issues

Jadwiga Suchecka[1]

Abstract

A space or an area is characterised by the specific data and the uncertainties inherent in it. Because of the different goals of spatial analyses, one needs to particularise a set of information which will represent the population. The results of analyses based on spatial data are obtained by applying the relevant methods depending on the types of input information. This dependence results mainly from the location of data relating to objects or processes in the geographic space (topological, geometric or geographic properties). Changes to the location of the objects or processes not only affect the final analysis results, but also the spatial distribution of the analysed variable in particular time intervals. This means that the interdependence and heterogeneity of the localised spatial observations results from their nature. These properties are connected with a relevant location of the measurement points, i.e. spatial autocorrelation and also indicate the existence of directly unobservable factors, which differentiate the occurrence of the spatial phenomena. The aim of this paper is to present selected aspects of two theories: (1) design-based sampling, which assumes that the characteristics of a population in a given region are unknown, but permanent; and (2) model-based sampling, where the set of observed values for the entire region is obtained through the observation of a single implementation of a stochastic volatility model in that system.

Keywords: *spatial sampling, design-based sampling, model-based sampling*
JEL Classification: C46, C180, R120

1. Introduction

Spatial statistics methods originate from spatial analyses covering a set of procedures, which results depend on the location of objects or processes in a 2-, 3- or *d*-dimensional geographical space. The type of spatial data used decides on the way of extracting knowledge from it and on the graphical presentation methods and quantitative analysis techniques applied. Thus, the final result of the conducted analyses mainly depends on the location of the analysed objects or processes. Any change in the location of the examined processes or objects results not only in a change of the final result of the analysis, but also in a change of the analysed variable in the spatial distribution in particular units of time. This means that the nature of the localised spatial observations has an impact on their interdependence and heterogeneity, as well as, on the uncertainty inherent in spatial data itself (Suchecka, 2014).

[1] University of Lodz, Faculty of Economy and Sociology, Department of Spatial Econometrics, Rewolucji 1905 37, 90-214 Lódź, Poland, e-mail: suchecka@uni.lodz.pl.

These properties are related to the relative location of the measurement, the so-called spatial autocorrelation and simultaneously indicate the existence of directly unobservable factors causing a spatial variation of the phenomena/processes. A lack of accurate information on the location of a given feature results in the loss of a significant part of the information.

The specific properties of spatial data and a large amount of information on localised data restrict the application of classical statistical measures for describing the structure of the analysed populations, correlation dependencies and statistical inference based on a spatial sample (Gruijter and Braak, 1990), the basis of which may be the sources or types of information. In the case of the first group, one can distinguish natural (geographic) and socio-economic information, as well as, the combination of these two types of information – infrastructural data. The second group may include information obtained from various types of topographic measurements, satellite images, scanning and software within the scope of the Geographic Information System, Territorial Information System, Global Positioning System or Remote Sensing (Kumar, 2007). Accepting the type of spatial information as a criterion, while taking into account the nature of the variation and the measurement value of the localised variable, one can distinguish surface data, lattice data (also known as area data) and point pattern data. Each of these types of information necessitates the application of different spatial analysis methods. Thus, spatial data obtained directly in the field on the basis of spatial assessments and information may be arranged regularly or irregularly that generally relates to natural resources or refers to phenomena and objects on a continuous surface. One applies surface interpolation methods for this type of data.

Lattice data are derived from the particular geographic area of the analysed objects and describes demographic, economic and social phenomena/processes. Its distinguishing feature is discrete (irregular) variation. The separated fragments of the surface subject to the observation are polygons with defined borders (e.g. the edges of a neighbourhood), and often occur as the aggregation of individual data expressing features, attributes and values of localised variables as a simple or complex location. In the case of the simple location, analysing lattice data can determine what can be found in a given place, while the complex location allows determining where the objects of specific relationships are located.

Information reflecting objects located in specific points in space (point pattern data) refers to the implemented phenomena. The values of these variables tend to form aggregations. Therefore, it is necessary to test these types of phenomenon.

The research in a certain geographic space is supplemented by analyses using spatio--temporal data relating to objects, phenomena or processes. While performing the selected

spatial analyses, attention should be paid to another important aspect of the fundamental property of spatial information – the uncertainty inherent in the data. This uncertainty can appear in other consecutive analyses, affecting the results. The main source of uncertainty is the way the analysed objects are defined in space. When selecting objects, one can define them well or poorly. The distinction between poorly- and well-defined objects determines the choice of the quantitative analysis method.

A well-defined object can be distinguished from other objects provided that one knows its attributes and spatial boundaries. In contrast, poorly-defined objects are characterised by indefiniteness (this results from a lack of clear boundaries for the object) and heterogeneity (connected with the assignment of an object to a particular class).

The presented considerations show that because of the multitude of spatial information, its specificity and the inherited uncertainty, as well as, the different goals of the conducted spatial analyses, one should precisely define what kind of information set is to represent the wider population, i.e. a so-called universe representing a total set of spatial units. In this case, the organisation of spatial statistical research occurs in similar stages as in the case of "classical" research. Griffith (1988) in his work justifies the usefulness of applying spatial sampling in the analysis of geographic data by noting that: (1) geographical information allows for the creation of a subset of area units according to specific criteria arising from the chosen research goal; (2) area units can be moved to create different samples (permutation) and the values in a set of area units can be randomised; (3) subsets in surface units can be created; (4) geographical data is characterised by basic stochastic processes; (5) the surface can be divided randomly; and (6) the movement of elements in space and their changes over time can be analysed.

The area units defined in the initial stage of the statistical analysis further become observations and their set as a whole constitutes the geographical population. This is possible if the geographical population includes N units, consisting of the sampling frame, and if one can create a subset (sample) consisting of n units in a certain way. Another approach to the creation of spatial statistical samples is required in situations where the target population is unknown, if there are directly unobservable factors causing the spatial variation of a phenomenon (e.g. in the region or time) and if it is necessary to create a super population as the basis for statistical inference. It should be noted that in statistical inferencing, it is also important taking into account the way the surface is divided (randomly or non-randomly) into area units and the designation of areas or boundaries of such areas/regions, as well as, the application of a different selection method for spatial sampling units (Wang et al., 2012). The

solution to this problem is proposed by two-sampling and inference theories, which use design-based sampling and model-based sampling (Shelin, 2012). Therefore, the aim of this paper is to present selected aspects of two theories. Design-based sampling assumes that the characteristics of the population in a region are unknown, but permanent and model-based sampling assumes the set of observed values in the entire region is obtained through the observation of a single implementation directory of the stochastic volatility model in that system.

When applying one of the discussed approaches, one should take into account a different method to introduce an element of randomness into the stochastic structure as a condition for statistical inference. The classical inference assumption, based on partial research, applies to the possession of a series of independent random variables with identical probability distributions coinciding with the distribution of the analysed feature in the population. In spatial research (as in the case of other partial/non-exhaustive statistical research), in order to obtain a relevant spatial sample, one often uses additional knowledge and information from outside the sample in the early stages of the design and implementation of the research (Szreder and Krzykowski, 2005). The application of this information in spatial research leads, on one hand, to the design of non-simple samples, and, on the other hand, to the development of relevant inference methods applied for these samples. This problem is of particular importance in non-exhaustive social and economic studies, in which the object, phenomenon or process's location in terms of place and space plays an important role (Kumar, 2007).

In spatial research dealing with social phenomena or processes, the space is usually identified as a region or territory, as well as, with a search for answers to the question what states for "here and where". In the first case, it should be assumed that the characteristics of the population are unknown, but fixed, and the randomness of the sample is introduced by a set of stochastic models called super populations. In the second case, the assumption concerns the instability of the characteristics of the population in the region of interest (often the population is infinite) and the randomness of the sample, as in the first case, is achieved by a set of stochastic models (super populations). It should be also noted that one usually introduces a super population into spatial research when a single implementation of the stochastic process enabling to obtain parameters for this distribution is the unit of the measurement.

2. Spatial sampling

In the process of spatial sampling, depending on the purpose of the research, one can propose probabilistic (random) and non-probabilistic (non-random) sampling techniques. The distinction between these two techniques introduces the application of sampling methods (including a combination of sampling techniques) for research and statistical inference. A correct numerical interpretation of the results requires information on: a) the research design and implementation; b) knowledge of the population from which this sample was taken, rather than of the sample itself (Szreder, 2015); c) the application of the sample results for a generalisation to the analysed population. The specificity of the spatial research resulting from the nature of the spatial data also requires the *a priori* knowledge of objects or phenomena/processes existing in a given space, determination of the target numbers of the research sample and methods for replenishing (imputation) the missing information obtained in the early stages of the research.

In spatial research, particularly on social or economic issues, it is necessary or expedient to apply non-random sampling techniques (Szreder, 2010), introducing a different mechanism of selection of objects for the spatial research (other than the probabilistic version).

2.1. Representative samples in spatial research

Spatial analyses using random samples are associated with the validity of the results and generalisation of these results to other, similar populations. These issues are connected with probabilistic sampling and the application of the theory of probability for the generalisation of the results to the population. An additional issue in the selection of the sample is a choice enabling making a distinction between the spatial population and the spatial sample. That is the reason for proposing resampling. The importance of the selection and location of the sample in the analysed space was emphasised in the works of Upton and Fingleton (1985) and Griffith (1988).

A characteristic feature of spatial information is its division into different categories. Using an appropriate sampling method, one should guarantee the inclusion into the sample of representatives of each category (a distinguished group of its representatives) without the need to maintain a proportional representation of all subgroups. In spatial research, the following random samples: simple, stratified, systematic and stratified systematic non-linear are being used.

The simple random sample is used when the population is situated in a wide geographical area and it is necessary to select relevant objects from this area. The issue boils down to

a cluster or area randomisation of the sample. A simple sample is the result of several stages. After determining the area, limits are defined and part of the surface which is subjected to be analysis is determined. This area is further divided into squares of adequate size. These squares become units in the research. Then, the entire area displayed on the map is numerated, thus creating a sampling frame in the form of a grid from which fragments of space are randomly selected according to the predetermined sample size. An advantage of this method is that it provides easy ways of drawing particular units, but a disadvantage is the risk of obtaining an uneven distribution of the squares in certain fragments of the analysed area (occurrence of clusters or lack of them).

Stratified random sampling can be applied when the population is divided into several categories. This sample is obtained as a result of several stages. The first stage consists of the division of the area into zones (squares) and the drawing of the objects (units) takes place within each zone. This method can be applied when the number of squares in particular zones is proportional to their surface. Systematic sampling can be used when the squares that divide the entire area are evenly distributed within the generated grid. This sampling technique consists of the selection of objects at equal intervals from an enumerative set of units according to a criterion adopted by the researcher. The advantage of this technique is the fact that it does not require the determination of the sampling space prior to the selection. Its disadvantage, however, is that the quality of the sample depends on the initial arrangement of the units (objects). Stratified systematic non-linear sampling is a combination of the two techniques discussed previously, with lacking the disadvantages of the systemic sample. Applying this method requires the division of the areas intended for analysis into zones consisting of a few squares. A square is then drawn from each zone.

In discussed spatial sampling drawing techniques, the defined random variable is a weighted sum of two components: information about the location (element of spatial structure) and independent sampling error, the distribution of which is usually determined by a certain probability model. One can also perform analyses by applying the appropriate Monte Carlo methods to repeat the research on the same samples in controlled conditions and using the analyses of spatial autocorrelation.

2.2. Design-based sampling

When analysing various socio-economic aspects in a particular region, the number of inhabitants may be unknown, but fixed. In this case, one should create a spatial sample using design-based sampling. In order to apply the relevant statistical methods using design-based

sampling, two populations should be determined: a single realisation of stochastic field and super populations as a stochastic field. Randomness is an important element of this sampling technique. In first step, a random sample should be used as the source of randomness. Then, the adopted sampling pattern should be repeated to generate a distribution of sample values until one is able to assess the parameters of the analysed population and the uncertainty inherent in the data based on this distribution (Wang et al., 2012). Taking into account specific features of the spatial data, sampling can be performed with consideration of spatial autocorrelation or heterogeneity.

Tobler's first law of geography (near things are more related than distant ones) shows that spatial autocorrelation is an indispensable attribute of spatial populations, contradicting the assumption of independent and identical population distribution. This autocorrelation also has an impact on the effectiveness of the sample as expressed by means of the error variance of the estimator. According to statistical theory, the value of the error variance of the estimator depends on the sampling technique and its size. In turn, the sample size is dependent upon whether the parameters of the population or the super population will be subjected to the estimation. If the inference relates to parameters of the population where the units are independent, the sample has to be larger because the variance error of the estimator is higher due to the existence of spatial autocorrelation (Haining, 1988). The solution suggested in the literature on the subject points to the validity of stratified sampling in the case of two-dimensional space or systematic sampling. The spatial sample will be selected based on a map of layers, taking into account the spatial autocorrelation level and the distance between the locations of the objects. If the spatial autocorrelation level is high and the distances between the objects are large, the redundancy level in the sample will be greater if the distance between the locations of the objects in the sample is smaller (Griffith, 2005). In the previous works on spatial statistics (Ripley, 1981) it was suggested that in the case of decreasing correlation functions small layers should be included in order to minimise the number of objects in a given layers and to account for their diversity (heterogeneity).

2.2.1. Sampling considering spatial heterogeneity

Heterogeneity of attributes in a space with geographically random fields includes a dependence of the second order: global variance and the spatial structure of the variance (Wang et al., 2010). These two components of the geographical variation should be included in the planning and evaluation of the sample, as well as, at the stage of planning the sampling and selection of the estimators (Griffith, 2005). In order to reduce general variance with the

application of spatial stratified sampling, the heterogeneous area should be divided into a few smaller subareas that are more homogeneous compared to the area as a whole (Rodeghiero and Cescatti, 2008). This division can be made to fulfil the following conditions: a) there is *a prior* knowledge of the analysed population; b) the initial sample is available; and c) the relevant auxiliary variables or distributions of other variables are known. The fulfilment of these conditions should enable determining the values of the target variables. Discussed spatial sampling method is implemented in two stages. In the first stage, the analysed space should be divided in a way enabling one to obtain a number of selected locations proportional to the number of locations in the entire space or to the variance. In the second stage, simple random sampling in each separate subarea can be resorted. One should also compare the efficiency of estimators obtained with the application of different random fields with the efficiency of estimators obtained by means of other stratified sampling methods (Wang et al., 2012).

2.3. Model-based sampling

When conducting analyses of regions, one may encounter variables that are random instead of fixed, while the set of values observed in the entire region is represented by a single realisation of the stochastic model of the variation in the universe. One of the proposed solutions to this problem is the application of model-based sampling in the process of spatial sampling (Gruijter and Braak, 1990). It is important to obtain randomness by applying stochastic models called super populations. This method can be applied when the sampling sites are stable, and inference can be performed when the stochastic model is valid. Determining the estimator or predictor depends on the weights of the sample data. These weights are determined by a covariance between the observations reflected in the model as a function of the co-ordinates of the sampling localisations (Groenigen and Stein, 1998). The application of model-based sampling allows one to achieve the basic objectives: the minimisation of the estimation error variance, equal spatial coverage in the case of irregular polygons and equal coverage in feature space. Another advantage of model-based sampling is that the locations do not have to be selected randomly. In practice, this indicates that the locations in the sample are initially assigned in the analysed region at random and then are included in the target sample using optimisation techniques. The second major challenge for model-based sampling is the achievement of equal spatial coverage. It should be noted that there are many kinds of optimisation methods, the main task of which is taking into account all analysed regions. However, it may happen that sparse data evenly distributed in the

analysed area could be found. In such cases, one should include weights in order to cover the units which might be not included in the sample. The primary criterion for such an inclusion is the minimisation of the average of the shortest distances from the "unsampled" sites to the nearest sampled one. In order to achieve the target sample, the analysed area should be presented in the form of a finite grid. This criterion should be also applied inside each cell of the determined grid. In the literature, one can also find other methods for obtaining a dispersed distribution of points in the sample. In practice, one must often resort to methods such as grid sampling, transect sampling, sequential sampling and nested sampling (Wang et al., 2012).

Another purpose of model-based sampling is the equal coverage in featured space. This condition is met when the distribution of units in the sample reflects as closely as possible the distribution of units in the population. If the distribution of units in the population is unknown, such a distribution is usually created on the basis of information about the distribution of units based on experience or distribution of ancillary data. Usually two criteria are used: (1) the Warrick-Myers criterion optimising sample locations to estimate the variograms (Warrick and Myers, 1987) and (2) Latin hypercube sampling as a stratified random sampling procedure (Iman and Conover, 1980).

Conclusion

The objective of the paper was to indicate the importance of spatial research in relation to the issues of choosing the right fragment of an area and the ensuing inference of the course of events/processes in space. Due to the specificity of spatial information it should be suggested to use the two types of spatial samples: design-based sampling and model-based sampling. Design-based sampling allows differing combinations of probabilistic sampling and perform design-based inference. This approach is associated with a search for answers to the question of "how much". On the other hand, applying model-based sampling allows obtaining information related to the question of "where". In conclusion, it should be emphasised that both methods provide a full sampling strategy through the use of three components: random fields, sampling project and estimators.

References

De Gruijter, J. J., & Ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, *22*(4), 407-415.

Griffith, D. A. (1988). *Advanced spatial statistics*. Dordrecht: Kluwer.

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.

Groenigen, J. W., & Stein, A. (1998). Constrained Optimization of Spatial Sampling using Continuous Simulated Annealing. *Journal of Environment Quality, 27*(5), 1078.

Haining, R. (1988). Estimating spatial means with an application to remotely sensed data. *Communications in Statistics – Theory and Methods, 17*(2), 573-597.

Iman, R. L., Conover, W. J., & Campbell, J. E. (1980). *Risk methodology for geologic disposal of radioactive waste: Small sample sensitivity analysis techniques for computer models, with an application to risk assessment*. Washington, D.C.: The Commission.

Kumar, N., (2007). Spatial Sampling Design for a Demographic and Health Survey. *Popul Res Policy Rev Population Research and Policy Review, 26*(5-6), 581-599.

Ripley, B. (1981), *Spatial Statistic*. New York: Wiley.

Rodeghiero, M., & Cescatti, A. (2008). Spatial variability and optimal sampling strategy of soil respiration. *Forest Ecology and Management, 255*(1), 106-112.

Shelin, L. (2012). *Spatial Sampling and prediction*. Umeä: Print and Media.

Suchecka, J. (2014). *Statystyka przestrzenna: Metody analiz struktur przestrzennych*. Warszawa: Wydawnictwo C.H. Beck.

Szreder, M. (2010), Losowe i nielosowe próby w badaniach statystycznych. *Przegląd Statystyczny, R. LVII*(4), 168-174.

Szreder, M. (2015), Zmiany w strukturze całkowitego błędu badania próbkowego. *Wiadomości Statystyczne, 1*, 4-12.

Szreder, M., & Krzykowski, G. (2005), Znaczenie informacji spoza próby w badaniach statystycznych. *Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego: Ekonometryczne modelowanie i prognozowanie wzrostu gospodarczego, 1*, 157-168.

Upton, G. J., & Fingleton, B. (1985). *Spatial data analysis by example*. Chichester: Wiley.

Wang, J., Haining, R., & Cao, Z. (2010). Sample surveying to estimate the mean of a heterogeneous surface: Reducing the error variance through zoning. *International Journal of Geographical Information Science, 24*(4), 523-543.

Wang, J., Stein, A., Gao, B., & Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics, 2*, 1-14.

Warrick, A. W., & Myers, D. E. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research Water Resour. Res., 23*(3), 496-500.