Beyond Pairs: Generalizing the Geo-dipole for Quantifying Spatial Patterns in Geographic Fields

Rui Zhu, Phaedon C. Kyriakidis, Krzysztof Janowicz

Abstract With their increasing availability and quantity, remote sensing images have become an invaluable data source for geographic research and beyond. The detection and analysis of spatial patterns from such images and other kinds of geographic fields, constitute a core aspect of Geographic Information Science. Per-cell analyses, where one cell's characteristics are considered (geo-atom), and interaction-based analysis, where pairwise spatial relationships are considered (geo-dipole), have been widely applied to discover patterns. However, both can only characterize simple spatial patterns, such as global (overall) statistics, e.g., attribute average, variance, or pairwise auto-correlation. Such statistics alone cannot capture the full complexity of urban or natural structures embedded in geographic fields. For example, empirical (sample) correlation functions established from visually different patterns may have similar shapes, sills, and ranges. Higher-order analyses are therefore required to address this shortcoming. This work investigates the necessity and feasibility of extending the geo-dipole to a new construct, the geo-multipole, in which attribute values at multiple (more than two) locations are simultaneously considered for uncovering spatial patterns that cannot be extracted otherwise. We present experiments to illustrate the advantage of the geo-multipole over the geo-dipole in terms of quantifying spatial patterns in geographic fields. In addition, we highlight cases where two-point measures of spatial association alone are not sufficient to describe complex spatial patterns; for such cases, the geo-multipole and multiple-point (geo)statistics provide a richer analytical framework.

Rui Zhu
STKO Lab, Department of Geography, University of California, Santa Barbara, USA
e-mail: ruizhu@geog.ucsb.edu

Phaedon C. Kyriakidis
Department of Civil Engineering and Geomatics, Cyprus University of Technology, Cyprus
e-mail: phaedon.kyriakidis@cut.ac.cy

Krzysztof Janowicz
STKO Lab, Department of Geography, University of California, Santa Barbara, USA
e-mail: jano@geog.ucsb.edu

1 Introduction and Motivation

The geo-atom, defined as $\langle x, Z, z(x) \rangle$, plays an important role in Geographic Information Science as the core representation of spatial information (Goodchild et al., 1999, 2007). The geo-atom associates a spatial location x with an attribute feature Z via the functional mapping $z(x)$. In terms of analysis, the geo-atom is applied in the computation of classical statistics used to describe aspects of spatial pattern in geographic fields. Examples of such statistics include the mean, variance, proportion of specific attribute values, and so on. Cell-based analysis of remotely sensed images with multiple attributes being available at each cell, form another common example of the usage of the geo-atom. Per-cell classification of low spatial resolution multispectral images is an example of such analytical operation. Finally, the geo-atom representation also applies to the object-driven perspective of spatial analysis, whereby each atom refers to an object rather than a cell.

The geo-atom considers each location x independently from other locations. This independence, however, ignores any interaction between locations, a critical aspect of geographic pattern (Goodchild et al., 2007). To address this shortcoming, Goodchild et al. (2007) introduced the concept of a geo-dipole, $\langle x, x', Z, z(x, x') \rangle$, whereby the interaction of variables between two locations x and x' is described via the two-point function $z(x, x')$. Such interaction function often involves measures of similarity of attribute values at location pairs, along with their geographical (or other) distance. Statistics relying on the geo-dipole for exploring spatial patterns include the distance to nearest neighbor, Ripley's K, Moran's I, the correlogram, the semivariogram, and so forth. The same can be argued for interpolation, e.g., the interpolation of a temperature surface based on data obtained at monitoring stations, as well as for geographic contextual classification (Atkinson and Lewis, 2000; Lu and Weng, 2007; Congalton, 1991), e.g., the improved classification of land use categories accounting for image texture information. Two-point statistics, such as the variogram that quantifies spatial auto-correlation, have been used several decades before the term geo-dipole was introduced; it was Goodchild et al. (2007), however, that first brought this notion into a more conceptual and theoretical level, which is the main focus of our work as well.

The geo-dipole considers a particular type of spatial interaction; namely, pairwise interactions. Those pairwise or two-point interactions are often (linearly) combined to arrive at interactions characterizing multiple locations, as, is done, for example, in spatial interpolation. In Kriging interpolation, in particular, the semivariogram model is first used to link each sample data location with a single interpolation location, and then such elementary two-point relations are combined through the Kriging system to arrive at interpolation weights. The entire procedure is based on explicit prior probabilistic models, such as the classic multivariate Gaussian model which is fully determined by its first-order statistics – the *mean* component – and its second-order statistics – the pairwise *covariance* function (Remy et al., 2009). However, these two-point models can only capture relatively simple spatial interactions, such as regularity, randomness, or clustering in attribute values. The identification and analysis of more complex spatial interactions, like those associated with curvilinear or other types of geometric structures, call for higher-order or multi-point statistics.

In this work, we propose the *geo-multipole* as a new conceptual model in which the interactions among multiple locations are simultaneously quantified. To model this kind of multiple-point interactions, we employ higher-order statistics, namely multiple-point (geo)statsitics together with their estimating approaches. In order to illustrate the necessity and feasibility of the geo-multipole, we compare it against the geo-dipole and classical two-point statistics for the recognition of urban spatial patterns. Although the geo-multipole concept could be employed to both object- and field-based representations of geographic information, this work focuses on methods and applications to geographic fields only.

The remainder of this paper is structured as follows: Section 2 briefly summarizes related work on analyzing and predicting geographic field patterns. In Section 3, the geo-atom and geo-dipole are approached from a probabilistic perspective in the context of geographic field analysis and then generalized to arrive at the notion of a geo-multipole. To motivate the need for the geo-multipole, Section 4 presents five contrasting spatial patterns extracted from remotely sensed images and compares them using two-point statistics and multiple-point statistics under the geo-dipole and the geo-multiple frameworks, respectively. Finally, Section 5 summarizes our results and highlights future research directions.

2 Related Work

In this section we review related work on geographic information representation and analysis required for the understanding of the proposed geo-multipole, as well as background material on multi-point (geo)statistics.

2.1 Geographic Conceptualization

The conceptualization of geographic information has been discussed in GIScience since its emergence (Goodchild, 1992b,a; Couclelis, 2010). The core challenge is how to model (and distinguish) field-based and object-based views on geographic occurrences. Corresponding work includes the geographic field (G-Field) and object (G-Object), field object, object field, general field, and so on (Goodchild et al., 2007; Liu et al., 2008; Cova and Goodchild, 2002; Voudouris, 2010). To unify the multitude of concepts, Goodchild et al. (1999) introduced the *geo-atom* by which the former concepts can be generalized. In a later work, Goodchild et al. (2007) argued that these concepts are designed for describing the static distribution of features and attributes on the Earth surface, whereas dynamic processes of geographic phenomena require different conceptual models, i.e. interaction models. Goodchild et al. (2007) went further to propose the *geo-dipole*, in which the interaction between two locations is modeled. The authors demonstrated that the geo-dipole is capable of representing many analytical interaction models, such as object fields, metamaps, object pairs, and association classes. One common characteristic of these analytical models, however, is the property of pairwise interactions, which is also the conceptual foundation of the geo-dipole. For more complex, but nonetheless very frequent spatial patterns, e.g., those emerging in urban environments, the geo-dipole might not suffice to adequately model the complexity of spatial interactions, as more than two locations may be involved simultaneously in defining a pattern. For example, it makes sense to study a central market place located in a dense residential area with many individual private units, but it would be very limiting to observe only one pair, i.e, a private unit and the market center. Considering the interaction between many private units and the market center simultaneously is different from considering pairwise interactions between each private unit and the market center, e.g., when the task is to uncover a star-shaped pattern formed by the market and the incoming streets with their residential units. To the best of our knowledge, multiple-point interaction has been seldom formalized in conceptual models in GIScience, an exception being the concept of Markov (random) fields where spatial interaction is defined using higher-order cliques encompassing groups (triplets, quadruplets, and so forth) of pixels. In terms of applications, however, such higher-order interactions are rarely quantified, and inference in such fields amounts to considering pair-wise (two-point clique) interactions only.

2.2 Geographic Field Analysis

As discussed in Section. 2.1, the *field* is one core concept of geographic information science (Kuhn and Frank, 1991; Kuhn, 2012). The detection and analysis of patterns from geographic fields, constitutes a critical task not only in geography, but also in related sciences, such as geology, environmental sciences, ecology, oceanography, and so on. Whether spatial context is explicitly considered or not distinguishes analytical approaches into non-contextual analyses and contextual analyses. Non-contextual analysis only focuses on individual cells and no interactions with neighbors are taken into account (Settle and Briggs, 1987; Rollet et al., 1998; Fisher, 1997). This type of approach is commonly used in the classification of hyper-spectral remote sensing images or the spatial prediction of many other multivariate geographic fields (Lu and Weng, 2007). In contrast, contextual analysis introduces spatial patterns into the process of prediction and is frequently applied to high spatial resolution geographic fields (Li et al., 2014), including remotely sensed images. Depending on the way of incorporating spatial information, the analysis of fields can be categorized into distance-based and object-based approaches (Li et al., 2014).

Distance-based Analysis

In this approach, spatial patterns are described by pairwise dissimilarities between attribute values measured at locations separated by specific *distance* lags (Cressie, 1993); examples include the variogram, the correlogram or transition probability diagrams. Such distance-based or two-point statistics are widely employed for incorporating spatial auto-correlation into interpolation and classification of field information (Atkinson and Lewis, 2000; Remy et al., 2009). For classification purposes, in particular, distance-based spatial interaction pertaining to multiple attributes (reflectance values recorded in different spectral bands at each cell or pixel) has been used as a model of field (image) texture, and incorporated into the classification procedure via: (1) local (within a neighborhood template) sample or modeled variograms used as additional entries of the feature vector at each cell (Carr, 1996; Carr and De Miranda, 1998; Ramstein and Raffy, 1989); and (2) multivariate variograms altering the weights originally attributed to entries of the feature vector, had classification been performed without accounting for spatial information (Oliver and Webster, 1989; Bourgault et al., 1992). Variogram-based analysis of geographic fields, however, constitutes a two-point representation of spatial interactions, and typically invokes the rather limiting assumption of second-order stationarity (Remy et al., 2009).

Object-based Analysis

In object-based image analysis (OBIA) the field is first segmented into homogeneous areas, regarded as objects, and then the predictions about the cells contained within these objects are assumed to be the same (Blaschke, 2010; Blaschke et al., 2014; Li et al., 2014). In OBIA, spatial information is considered in the process of segmentation, e.g., for Markovian methods (Jackson and Landgrebe, 2002) and watershed methods (Salembier et al., 1998). Object-based analysis is commonly used in the classification and simulation of remotely sensed images and related work demonstrated the improvement over cell-based analysis (Blaschke, 2010; Ceccarelli et al., 2013). However, OBIA is limited in terms of the assumption of homogeneous objects, the sensibility to segmentation algorithms, as well as the difficulty of using a large amount of conditioning data when it comes to generating patterns in a simulation setting (Remy et al., 2009).

2.3 *Multiple-point (Geo)statistics*

Multiple-point (geo)statistics (MPS) were initially proposed to overcome the limitations inherent in variogram-based and object-based analysis for the identification of complex spatial patterns in the subsurface (Guardiano and Srivastava, 1993). The core idea behind MPS is that, since variogram models are commonly estimated from data pertaining to analog deposits or outcrops or even expert-drawn images due to data limitations regarding the subsurface, why not directly borrow entire (conceptual) images as depositories of spatial patterns (Remy et al., 2009). It is these images that domain experts use to visually detect spatial patterns from and, thereby, estimate variogram model parameters. In addition, variograms being two-point statistics cannot capture spatial patterns resulting from complex earth processes. Implementing this idea, MPS abandons any explicit statistical model, but regards the training image as one realization of non-analytically defined random field pertaining to the actual (target) region being studied. The key assumption under MPS is that the training image contain adequate (in terms of complexity and number) replicates over the patterns deemed to occur at the target region (Strbelle, 2002; Journel and Zhang, 2006). Multiple-point statistics, e.g., the probability of three or more grid cells having simultaneously a particular lithological class, are then directly learned from the training image.

So far, most applications of multiple-point (geo)statistics are limited to the domain of geology, in which subsurface heterogeneities, such as those found in porous media and reservoirs, are modeled and simulated (Strebelle et al., 2001). Several MPS algorithms have been implemented for applications in geology. Examples include simple normal equation sampling (Strébelle and Journel, 2000), filter-based simulation (Zhang et al., 2006), and direct sampling (Mariethoz et al., 2010).

In recent years, two threads of applications of multiple-point (geo)statistics can be distinguished for classifying geographic features, such as roads, buildings, vegetation, and open water-bodies, using remotely sensed images. Tang et al. (2016) incorporated MPS as new weights into K-nearest neighbor (KNN) classification and illustrated the improved performance compared to other supervised learning models such as Bayesian classifiers and Support Vector Machines. Others (Ge et al., 2008; Ge and Bai, 2010, 2011; Ge, 2013) introduced the Classification by Combining Spectral Information with Spatial Information in Multiple-point Simulation (CCSSM), in which MPS-based spatial classification is combined with pixel-based spectral classification using fusion techniques, such as consensus-based and probability-based fusion. The performance of the CCSSM approach compared favorably to traditional classification approaches, such as Maximum Likelihood Classification.

While these studies aim at improving the classification performance for remotely sensed images by applying MPS, our work focuses on investigating the necessity and value of applying multiple-point interactions in analyzing geographic information, particularly geographic fields. We do so by generalizing the geo-dipole to stay within the conceptual framework proposed by Goodchild and others. Using our approach, multiple-point (geo)statistics are not limited to classifications problems, but can also be used for interpolation, simulation, and so forth. Going beyond the recent practice of using MPS in remote sensing, multiple-point (geo)statistics could also be extended to other types of fields such as model outputs and irregular tessellations. Lastly, by introducing the geo-multipole, we hope to foster the development of GIScience-specific MPS algorithms that suit the needs and application areas of our community, e.g., for studying urban environments.

3 Introducing the Geo-multipole

In this section we introduce the geo-multipole as a conceptual generalization of the geo-dipole and also provide a probabilistic perspective on the geo-atom.

3.1 Conceptual Models

Capitalizing on the previously established conceptual models of the geo-atom $\langle x, Z, z(x)\rangle$ and the geo-dipole $\langle x, x', Z, z(x, x')\rangle$, we define the geo-multipole as follows:

Geo-multipole: $\langle x, t_N, Z, z(x, t_N)\rangle$
where $t_N = x_1, .., x_N$ are the N neighbors of x.

Here, we categorize conceptual models into three groups: (1) single-point data models, namely the geo-atom where no interactions between locations are considered; (2) two-point data models, namely the geo-dipole where pairwise interactions are considered; and (3) multiple-point data models, namely the proposed geo-multipole, which can be regarded as a generalized conceptualization of spatial interactions as defined by the geo-dipole. With respect to the geo-multipole, a neighborhood t_N, with N locations, is defined for each target location x. Then the interaction between x and its neighborhood t_N in terms of variable Z is defined as $z(x, t_N)$. The key difference between the geo-dipole and the geo-multipole is the fact that locations $x_1, .., x_N$ in t_N are *simultaneously* considered (along with the corresponding attribute values) when modeling their interactions with x. In contrast, interactions are considered in *pairs* under the conceptualization of the geo-dipole despite that multiple pairwise interactions could be combined in *sequence*. It is important to note that simultaneously modeling interactions between the target and all its neighbors is mathematically different from simply combining pairwise interactions between each neighbor and that target. Namely, $z(x, t_n = \{x_1, ..., x_N\})$ does not imply $f(z(x, x_1), ..., z(x, x_N))$.

3.2 Probabilistic Perspective

Geographic fields are frequently assumed to be generated from stochastic processes, and are thus regarded as realizations of a random field. Along the same lines, this work approaches the three conceptual models from a probabilistic perspective. Therefore, we discuss their descriptive statistics, as well as relevant estimation approaches in what follows.

3.2.1 Geo-atom

To summarize geographic fields in terms of the geo-atom, single-point statistics could be employed. The *mean* and *standard deviation* are the most commonly used examples. They are capable of describing the average magnitude, as well as the spread, of values of the attribute of interest across the domain. Other common statistics include *quantiles*, the *number* of cells whose attribute values satisfy a particular query, and the *probability density function* (PDF) of the attribute values:

$$f(z,x) = prob(Z(x) = z \pm \varepsilon)$$

where ε denotes an infinitesimally small value. It should be noted that, in this work, we use the term PDF also for the case of a categorical attribute Z, instead of the more correct notion of a probability mass function (PMF) for the sake of simplicity. For the same reason, we drop the $\pm\varepsilon$ notation from the PDF in what follows.

Optimal prediction at each location x, requires knowledge of the PDF $f(z,x)$. In the univariate case, the estimation of $f(z,x)$ only depends on the location x itself and no other variables at this location are provided. In addition, there is no interaction between the attribute value at this location and other locations. Therefore, unless the probability density function $f(z,x)$ is estimated by domain experts using physical models or experience considering a limited number of sample data $z(x_s; s = 1,\ldots,S)$, it is challenging, if not impossible, to estimate the function from a probabilistic perspective.

In the multivariate case where the target variable Z is co-located with other variables Z', the relation between $Z(x)$ and $Z'(x)$ could be modeled through sample data; hence, the multivariate version of $f(z,x)$, i.e.$f(z,z',x) = prob(Z(x) = z|Z'(x))$, could be estimated. This second case is common in GIScience. For example, if Z^{temp} is a unknown temperature field and we observe elevation Z^{elev} and solar radiation Z^{solar} as known fields, by using the sample data $\{Z^{temp}(x_s), Z^{elev}(x_s), Z^{solar}(x_s); s = 1,....S\}$, the relation between Z^{temp} and $\{Z^{elev}, Z^{solar}\}$ can be modeled through either linear or non-linear models. Then, the conditional PDF of the random variable Z^{temp} can be estimated by substituting Z^{elev} and Z^{solar} in the trained model. In remote sensing applications, per-cell classification is another example of this case, whereby reflectance values at different spectral bands form multiple feature variables and the class code at each cell forms the target categorical field.

3.2.2 Geo-dipole

Since the interaction between two points is now considered, concepts such as distance and neighborhood are key components of the geo-dipole. Statistics that could be used for describing spatial patterns via geo-dipoles are spatial autocorrelation measures, such as *Moran's I* and *Geary's C*, or their multiple lag-distance analogs, the correlogram and the semivariogram, for continuous data, and *transition proba-*

bilities for categorical data. In addition, the *conditional PDF* in this case is modeled as:

$$f(z,x,Z(x')) = prob(Z(x) = z|Z(x'))$$

To predict $f(z,x,Z(x'))$ within the geo-dipole framework, the key is to model the interaction $Z(x,x')$. Geostatistics provides approaches to model such interaction or association based on distance. For instance, under first- and second-ordered stationarity, $Z(x,x')$ could be characterized through semivariogram models whose parameters are estimated by sample data. Note here, that interaction among data of two different attributes can be also defined via cross-semivariogram models (Goovaerts, 1997). Given the interaction model $Z(x,x')$, the conditional PDF of the random variable $Z(x)$ could be estimated through an observed variable as in the univariate case, or multiple observed variables as in the multivariate case. One example for the univariate case is interpolation, whereby a, say, temperature field can be interpolated using limited sample data. Interpolation methods, such as inverse distance weighting and Kriging, account for pairwise interactions between $Z^{temp}(x)$ and $Z^{temp}(x_s; s = 1, \ldots, S)$. Land use classification using multi-spectral remote sensing images is an example of the multivariate case. In contrast to incorporating data pertaining to only one spectral band, multiple bands, together with their modeled cross-interactions, are used to arrive at land use classifications.

3.2.3 Geo-multipole

In contrast to the geo-dipole, the geo-multipole takes the N neighbors of x into account *simultaneously*. Rather than two-point statistics, higher-order statistics are thus required to model such a multiple-point interaction $Z(x,t_N)$. Similar to the geo-dipole, such multiple-point statistics could be obtained through sample data. However, the size of sample data sets is typically relatively small for such a multiple-point inference endeavor; this might result into biased estimates. A more promising approach is to use training images, which are assumed to contain spatial patterns deemed representative of the actual field under study. Multiple point interactions $Z(x,t_N)$ are then directly learned from the training image without building any parametric model. Specific algorithms to accomplish this are discussed in Section 3.3. The *conditional PDF* of the random variable $Z(x)$ at a target location x can then be built, from which an optimal prediction can be derived for the attribute Z at that location:

$$f(z,x,Z(t_N)) = prob(Z(x) = z|Z(t_N))$$

The geo-multipole is appropriate for analyzing geographic fields that have rather complex spatial patterns. Examples include categorical fields that pertain to urban structures, such as roads that exhibit curvilinearity patterns or rooftops that have polygonal shapes. The geo-multipole could also be used for spatial interpolation, e.g., for air pollution patterns, and spatial simulation, e.g., of urban growth. Concrete examples of utilizing the geo-multipole, together with a comparison to the geo-dipole are given in Section 4.2.

3.3 Higher-order Statistics

The geo-multipole concept employs higher-order statistics with respect to the geo-multipole, whereby two-point statistics are considered. Therefore, the question of how to efficiently compute or model such higher-order or multiple-point statistics becomes a key challenge. Although different algorithms exist for implementing multiple-point (geo)statistics, the core idea is to use the training image as an analog for learning higher-order spatial patterns. Basic elements of MPS algorithms are (Mariethoz and Caers, 2014):

- **Training images (TI)** that contain spatial patterns; see Fig. 2
- **Template (T)** for scanning training images; see column 1 of Fig. 4
- **Data events ($dev(x)$)**, which are simultaneous (joint) combinations of attribute values at template pixels; see column 2 of Fig. 4

Since templates are used to detect spatial patterns, attribute values at more than two points are simultaneously considered in MPS. After obtaining data events from training images, multiple point statistics, i.e. $prob(x = z|t_N)$, can be calculated (Honarkhah and Caers, 2010). Together with actual or directly sampled data, e.g., land cover classes verified at particular cells from ground surveys, these learned conditional probability values can subsequently be applied to estimate, or simulate, attribute values at non-sampled locations.

In our work, we implement one of the many MPS algorithms available, namely simple normal equation simulation (SNESIM), to estimate the required higher-order statistics. We then employ simulation to generate synthetic images of fields, in order to visually explicate the patterns learned by MPS. Several steps are involved in SNESIM (Remy et al., 2009): (1) a search template T_j is first defined; (2) a search tree specific to template T_j is then constructed; (3) the conditioning data are located on the field (this step can be skipped for unconditional simulation); (4) a random path that visits all locations to be simulated is established; then for each location x along the path: (5) the conditioning data event $dev_j(x)$ defined by template T_j is selected; (6) the corresponding conditional probability from the search tree is retrieved; (7) and finally the simulated value from the conditional probability is generated and added to the conditioning data set. In this work, we make use of the Matlab library mGstat[1] to run SNESIM; illustrative examples are given in Section 4.2.

Note that higher-order statistics are different from classic map algebra or image processing operations, e.g., focal or zonal operations, or kernel filters. Higher-order statistics consider neighboring interactions simultaneously rather than splitting them

[1] http://mgstat.sourceforge.net/

into (weighted) linear combinations of pairwise interactions. In classical map algebra operations, neighbors are considered in a first-order (linear combination of neighboring attribute values) or at most second-order (neighboring attribute values weighted as function of pairwise distances) manner.

4 Case Study

In this section we demonstrate the utility of the geo-multipole concept in describing spatial patterns. We do so by means of employing multi-point statistics to highlighted use cases, and comparing the results against those obtained by solely relying on the geo-dipole and therefore two-point statistics such as variograms.

4.1 Sample Patterns

To illustrate the benefits of introducing the geo-multipole, as well as the feasibility of applying multiple-point (geo)statistics, we extracted several spatial patterns (shown in Fig. 2) in the form of binary maps from remotely sensed images (shown in Fig. 1). The binary maps are derived from remotely sensed images by threshold-based brightness segmentation to distill target patterns. The proportions of black pixels in those maps are quite similar (Pattern 1: 0.2697, Pattern 2: 0.258, Pattern 3: 0.257, Pattern 4: 0.267, Pattern 5: 0.269). The spatial patterns in the five binary maps, however, are rather different. Pattern 1 is extracted from streams, thus showing curvilinear patterns; patterns 2 and 3 are extracted from vegetation of a park and a golf court, respectively, and thus show circular patterns; pattern 4 is extracted from a residential area with rectangular patterns; and finally pattern 5 is extracted from the public garden of a mission, showing bounded patterns of different simple shapes.

4.2 Experimental Results and Discussion

The geo-dipole and the geo-multipole are compared in this section using the five patterns described in Section 4.1. Specifically, variogram-based and MPS-based approaches are applied for quantifying the selected patterns. Two sets of experiments are conducted in both approaches to highlight their differences: (1) a statistical description of the pattern, and (2) a simulation expression of the pattern, visualizing the information contents conveyed by this description.

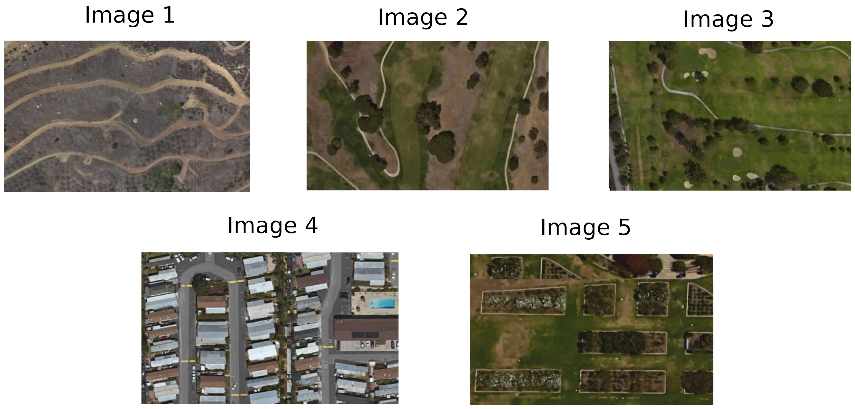


Fig. 1 Five remotely sensed images at 1 meter spatial resolution.

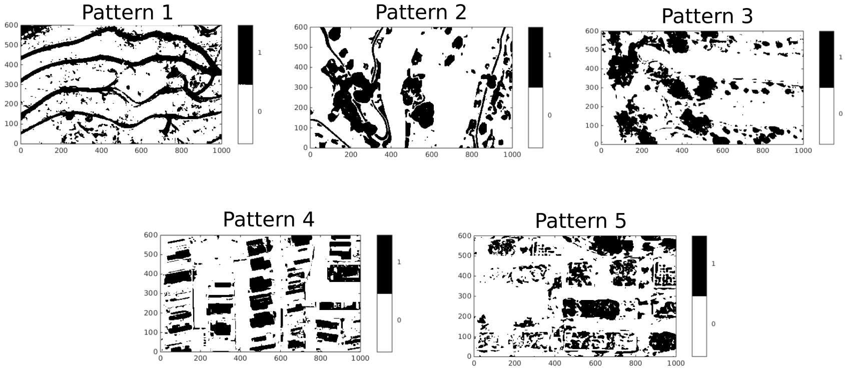


Fig. 2 Binary maps of five spatial patterns (600 × 1000).

4.2.1 Description of the Pattern

Directional semivariograms and conditional multiple-point probabilities are calculated to show their ability to characterize the selected spatial pattern. As the five examples show distinctive spatial patterns visually, the more different the results of the employed statistics are, the more successful the methods are in detecting distinct complex patterns.

Variogram-based Analysis

The two-directional semivariograms (i.e., West-East and North-South) for the five examples are illustrated in Fig. 3. As can be seen, despite the visually different patterns, their semivariograms for the two directions are generally similar, with a dramatic increase from distance lag 0 to about 50. The semivariograms also remain flat after the distance lag 60. The only salient characteristic is the bump at the distance lag 50 of the North-South semivariogram for Pattern 1; this is due to the repetition of multiple elongated (along West-East) features of relatively regular width (pseudo-periodicity). This observation indicates that two-point (geo)statitsics are barely enough to capture the complex spatial patterns embedded in these (urban) structures.

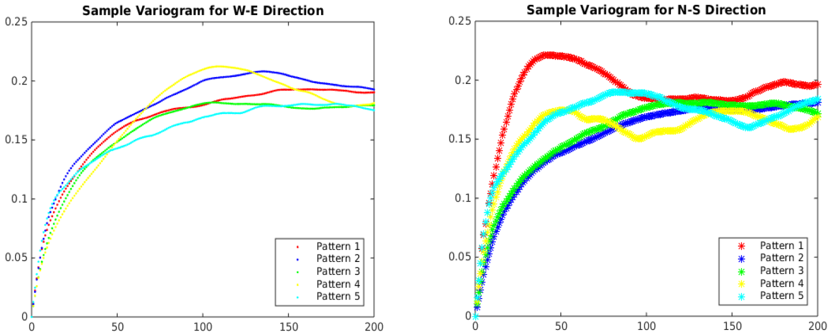


Fig. 3 Directional semivariograms for the five examples (Left: West-East; Right: North-South).

MPS-based Analysis

In multiple-point (geo)statistics, one computes the conditional probability of class occurrence given nearby classes in the template directly from the training image. The order of the statistics employed is determined by the size and geometry of the template. The lager the template, the more neighboring locations will be simultaneously considered. To show the capability of MPS in detecting different spatial patterns, a simplified template (see the first column of Fig. 4) was used for pattern 1 and pattern 4. To determine the class, i.e., black (1) or white (0), at the central pixel, its 8 neighbors are simultaneously considered as data events shown in column 2. The class of the central pixel will be assigned to the one that has the highest conditional probability. A data event's conditional probability is calculated as the frequency of occurrence, for example:

$$prob(z(x)=0|t_N) = \frac{\#(z(x)=0|t_N)}{\#(z(x)=0|t_N)+\#(z(x)=1|t_N)}$$

There are 2^8 possibilities for such a neighborhood configuration; in this work we sampled 6 of them for illustration purposes. From Fig. 4, we can see that the conditional probabilities using the 3×3 template are different between pattern 1 and pattern 4. Note that only a relatively simple template is tested here; had a more complicated template, such as a 80×80 square template, been used, the conditional probabilities would be even more different. Such an observation indicates the capability of MPS for learning complex patterns compared to the simple semivariogram-based analysis.

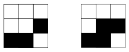
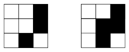
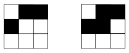
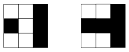
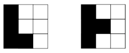
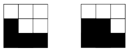
Templates	**Data Events**	**$P(x\mid n_1, \dots, n_8)$**			
		Pattern 1		Pattern 4	
		$P(x=0\mid n_1, \dots, n_8)$	$P(x=1\mid n_1, \dots, n_8)$	$P(x=0\mid n_1, \dots, n_8)$	$P(x=1\mid n_1, \dots, n_8)$
$n_1 n_2 n_3$ n_4 × n_5 $n_6 n_7 n_8$		0.000	**1.000**	**0.600**	0.400
		**0.600**	0.400	0.300	**0.700**
		0.500	0.500	**0.625**	0.375
		0.500	0.500	0.000	**1.000**
		**0.530**	0.470	0.458	**0.542**
		**0.610**	0.390	0.487	**0.513**

Fig. 4 Conditional multiple-point probabilities for patterns 1 and 4 (only 6 out of $2^8 = 256$ possibilities are shown).

4.2.2 Simulation of Pattern

To visualize the information content of variograms and multipl-point statistics, unconditional simulations are conducted using modeled variograms and multiple-point (geo)statistics, respectively. Our rationale here is that the more similar the simulated patterns are to the original examples, the more feasible the approach is in terms of learning spatial patterns.

Variogram-based Simulation

Unconditional moving average simulation via the Fast Fourier Transform (FFT) was used in this work to simulate realizations of 2-D multivariate Gaussian fields given the semivariogram model (i.e., exponential model); the resulting continuous images were then thresholded using suitable cutoff values so as to reproduce the same proportion of black pixels as the corresponding original binary images. From the Fig. 5, we have mainly two observations: (1) although the original examples 1 and 4 (in Fig.

2) show two different spatial patterns, their variogram-based simulations demonstrate similar spatial patterns; (2) the spatial patterns of both simulations in Fig. 5 are not consistent with the original patterns (i.e., the curvilinear pattern of pattern 1 and polygonal pattern of pattern 4). These two observations showcase the limitations of using variograms for simulating (and thus analyzing) spatial patterns; one should not expect a two-point variogram function to capture complex higher-order spatial patterns as those corresponding to elongated features or other curvilinear or geometric shapes.

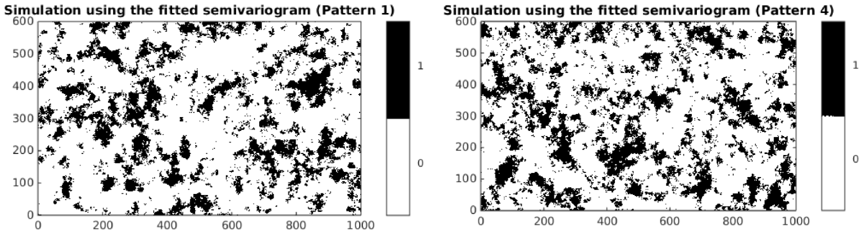


Fig. 5 Variogram-based simulations (pattern 1 and pattern 4).

MPS-based Simulation

The simple normal equation simulation (SNESIM) was applied in this work to generate the MPS-based simulation using the original patterns 1 and 4 as training images. The templates for both patterns were set to 80×80 squares. From the results in Fig. 6, we can observe that the two simulations show significantly different patterns, with pattern 1 showing more curvilinearity along the west-east direction and pattern 4 showing more polygonal geometries. Furthermore, comparing the simulated images with the original training images (in Fig. 2), we observe that the illustrated patterns in the simulations are relatively similar to the ones from the original images, although there are still inconsistencies between the two. A viable explanation for such inconsistencies is that the original training images (Fig. 2) are small in size and their patterns are rather complex with many elementary patterns being combined. For example, there are only four curved lines, which is the main pattern visually in the pattern 1, but there are also many small clusters across the domain. Summing up, despite some inconsistencies, the advantage of using MPS for learning spatial patterns is clearly highlighted by these simulations; particularly when compared to variogram-based approaches. Evidently, more work is required (possibly involving testing different MPS-based simulation methods) in order to improve the similarity between simulated and training images. It should be stressed, however, that the generation of spatial patterns with geometrical characteristics is a very improbable outcome when using variogram-based simulation algorithms.

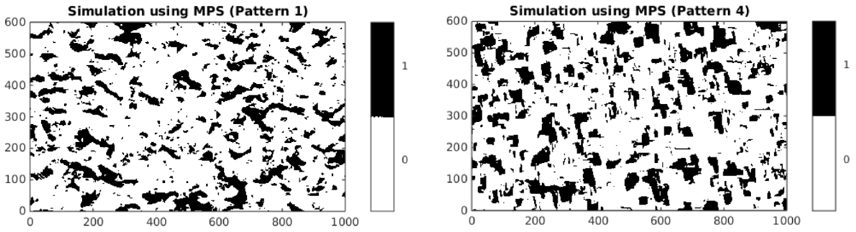


Fig. 6 MPS-based simulations (Pattern 1 and Pattern 4).

5 Conclusions and Future Work

In this work, we generalized the traditional two-point interaction model, the geo-dipole, by introducing the geo-multipole concept whereby multiple-point interactions are simultaneously modeled. Furthermore, a general framework for geographic field analysis was discussed from both geographic and probabilistic perspectives. All three conceptual models, the geo-atom, the geo-dipole and the geo-multipole, are included in the framework, and they represent statistics of different order, i.e. first-order statistics for the geo-atom, second-order for the geo-dipole and higher-order for the geo-multipole. Different descriptive statistics, prediction techniques, and concrete examples were given to demonstrate such a framework.

This work also discussed the application of multiple-point (geo)statistics as one potential approach for estimating higher-order statistics for geographic fields. In MPS, the training image is regarded as an explicit (non-parametric or better multi-parametric) model that replaces the role of implicit statistical models. The only assumption in using MPS is that the training image contains a representative collection of the spatial patterns expected at the target site; thus, the target field characteristics can be learned using approximate replicates contained in the training image. It should be noted, however, that since MPS places extreme ”faith” in the training image, there is a risk of over-parameterization; thus, more attention should be placed on selecting appropriate training images, possibly considering more than one such images (Mariethoz and Caers, 2014).

A series of experiments were conducted to illustrate the necessity and value of using the geo-multipole in quantifying patterns in field data. In short, we showed that spatial patterns extracted from multiple-point (i.e., MPS-based) interaction models are more realistic (better reproduce the complexity of patterns) compared to the ones extracted from two-point (i.e., variogram-based) interaction models.

There are several potential research directions for future work. First, the application details of multiple-point (geo)statistics for quantifying spatial patterns in geographic phenomena should be further explored. For example, the sensitivity of template geometry and size, the impact of the training image size and pattern richness, as well as the feasibility of using other algorithms, should be studied in more depth. Second, in addition to using MPS for contextual classification, MPS could also be applied to spatial simulations. For example, the performance of cellular automata could be improved by incorporating information from training images using MPS. Last but not least, techniques for estimating multiple-point interactions could be extended from applications pertaining to field information to applications involving other types of geographic information as well. For example, higher-order interactions among different places (objects) in gazetteers could be considered to supplement spatial signatures for place types when learning alignments between geo-ontologies, as proposed in Zhu et al. (2016).

References

Atkinson, P. M. and P. Lewis (2000). Geostatistical classification for remote sensing: an introduction. *Computers & Geosciences 26*(4), 361–371.

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS journal of photogrammetry and remote sensing 65*(1), 2–16.

Blaschke, T., G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Q. Feitosa, F. van der Meer, H. van der Werff, F. van Coillie, et al. (2014). Geographic object-based image analysis–towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing 87*, 180–191.

Bourgault, G., D. Marcotte, and P. Legendre (1992). The multivariate (co) variogram as a spatial weighting function in classification methods. *Mathematical Geology 24*(5), 463–478.

Carr, J. R. (1996). Spectral and textural classification of single and multiple band digital images. *Computers & Geosciences 22*(8), 849–865.

Carr, J. R. and F. P. De Miranda (1998). The semivariogram in comparison to the co-occurrence matrix for classification of image texture. *IEEE Transactions on Geoscience and Remote Sensing 36*(6), 1945–1952.

Ceccarelli, T., D. Smiraglia, S. Bajocco, S. Rinaldo, A. De Angelis, L. Salvati, and L. Perini (2013). Land cover data from landsat single-date imagery: an approach integrating pixel-based and object-based classifiers. *European Journal of Remote Sensing 46*, 699–717.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment 37*(1), 35–46.

Couclelis, H. (2010). Ontologies of geographic information. *International Journal of Geographical Information Science 24*(12), 1785–1809.

Cova, T. J. and M. F. Goodchild (2002). Extending geographical representation to include fields of spatial objects. *International Journal of Geographical Information Science 16*(6), 509–532.

Cressie, N. (1993). Statistics for spatial data: Wiley series in probability and statistics. *Wiley-Interscience, New York 15*, 105–209.

Fisher, P. (1997). The pixel: a snare and a delusion. *International Journal of Remote Sensing 18*(3), 679–685.

Ge, Y. (2013). Sub-pixel land-cover mapping with improved fraction images upon multiple-point simulation. *International Journal of Applied Earth Observation and Geoinformation 22*, 115–126.

Ge, Y. and H. Bai (2010). Mps-based information extraction method for remotely sensed imagery: a comparison of fusion methods. *Canadian Journal of Remote Sensing 36*(6), 763–779.

Ge, Y. and H. Bai (2011). Multiple-point simulation-based method for extraction of objects with spatial structure from remotely sensed imagery. *International Journal of Remote Sensing 32*(8), 2311–2335.

Ge, Y., H. X. Bai, and Q. Cheng (2008). New classification method for remotely sensed imagery via multiple-point simulation: experiment and assessment. *Journal of Applied Remote Sensing 2*(1), 023537–023537.

Goodchild, M. F. (1992a). Geographical data modeling. *Computers & Geosciences 18*(4), 401–408.

Goodchild, M. F. (1992b). Geographical information science. *International journal of geographical information systems 6*(1), 31–45.

Goodchild, M. F., M. J. Egenhofer, K. K. Kemp, D. M. Mark, and E. Sheppard (1999). Introduction to the varenius project. *International Journal of Geographical Information Science 13*(8), 731–745.

Goodchild, M. F., M. Yuan, and T. J. Cova (2007). Towards a general theory of geographic representation in gis. *International journal of geographical information science 21*(3), 239–260.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press on Demand.

Guardiano, F. B. and R. M. Srivastava (1993). Multivariate geostatistics: beyond bivariate moments. In *Geostatistics Troia92*, pp. 133–144. Springer.

Honarkhah, M. and J. Caers (2010). Stochastic simulation of patterns using distance-based pattern modeling. *Mathematical Geosciences 42*(5), 487–517.

Jackson, Q. and D. A. Landgrebe (2002). Adaptive bayesian contextual classification based on markov random fields. *IEEE Transactions on Geoscience and Remote Sensing 40*(11), 2454–2463.

Journel, A. and T. Zhang (2006). The necessity of a multiple-point prior model. *Mathematical Geology 38*(5), 591–610.

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science 26*(12), 2267–2276.

Kuhn, W. and A. U. Frank (1991). A formalization of metaphors and image-schemas in user interfaces. In *Cognitive and linguistic aspects of geographic space*, pp. 419–434. Springer.

Li, M., S. Zang, B. Zhang, S. Li, C. Wu, et al. (2014). A review of remote sensing image classification techniques: The role of spatio-contextual information. *European Journal of Remote Sensing 47*, 389–411.

Liu, Y., M. F. Goodchild, Q. Guo, Y. Tian, and L. Wu (2008). Towards a general field model and its order in gis. *International Journal of Geographical Information Science 22*(6), 623–643.

Lu, D. and Q. Weng (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing 28*(5), 823–870.

Mariethoz, G. and J. Caers (2014). *Multiple-point geostatistics: stochastic modeling with training images*. John Wiley & Sons.

Mariethoz, G., P. Renard, and J. Straubhaar (2010). The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources Research 46*(11).

Oliver, M. and R. Webster (1989). A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology 21*(1), 15–35.

Ramstein, G. and M. Raffy (1989). Analysis of the structure of radiometric remotely-sensed images. *International Journal of Remote Sensing 10*(6), 1049–1073.

Remy, N., A. Boucher, and J. Wu (2009). *Applied geostatistics with SGeMS: A user's guide*. Cambridge University Press.

Rollet, R., G. Benie, W. Li, S. Wang, and J. Boucher (1998). Image classification algorithm based on the RBF neural network and k-means. *International Journal of Remote Sensing 19*(15), 3003–3009.

Salembier, P., A. Oliveras, and L. Garrido (1998). Antiextensive connected operators for image and sequence processing. *IEEE Transactions on Image Processing 7*(4), 555–570.

Settle, J. and S. Briggs (1987). Fast maximum likelihood classification of remotely-sensed imagery. *International Journal of Remote Sensing 8*(5), 723–734.

Strbelle, S. (2002). Conditional simulation of complex geological structures using multiple point geostatistics. *Math Geol 34*(1), 122Strbelle.

Strébelle, S. and A. Journel (2000). Sequential simulation drawing structures from training images.

Strebelle, S. B., A. G. Journel, et al. (2001). Reservoir modeling using multiple-point statistics. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Tang, Y., L. Jing, P. M. Atkinson, and H. Li (2016). A multiple-point spatially weighted k-nn classifier for remote sensing. *International Journal of Remote Sensing 37*(18), 4441–4459.

Voudouris, V. (2010). Towards a unifying formalisation of geographic representation: the object–field model with uncertainty and semantics. *International Journal of Geographical Information Science 24*(12), 1811–1828.

Zhang, T., P. Switzer, and A. Journel (2006). Filter-based classification of training image patterns for spatial simulation. *Mathematical Geology 38*(1), 63–80.

Zhu, R., Y. Hu, K. Janowicz, and G. McKenzie (2016). Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS*.