How is the Third Law of Geography different?

A-Xing Zhu\textsuperscript{a,b,c,d} and Matthew Turner\textsuperscript{a,b}

\textsuperscript{a}Department of Geography, University of Wisconsin-Madison, Madison, WI, USA; \textsuperscript{b}School of Geography, Nanjing Normal University, Nanjing, China; \textsuperscript{c}State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; \textsuperscript{d}Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing, China

\textbf{ABSTRACT}

Three overarching principles governing patterns of geographic phenomena have been proposed that have been referred to by some as ‘laws of geography’. The first and the second principles address the spatial proximity and spatial heterogeneity of geographic phenomena. These principles, while powerful, fail to resonate with much geographical inquiry. The more recently proposed third principle concerns geographic similarity. The differences of it from the first two can be perceived in three basic aspects: principle expressed, form of expression and role of geographic examples (samples). The third principle emphasizes the geographic context of geographic variables in the form of geographic configuration, compared to a single spatial dimension that are emphasized in the first two principles. The third principle focuses on the comparative nature in the geographic configuration in terms of similarity, that is, in the form of ‘similar to’, as opposed to the relationships ‘between’ that are key to the first and second principles. The third principle emphasizes the individual representation of geographic examples, as opposed to the global representation of geographic examples. These differences not only distinguish the third principle as an important addition to the other two, but also provide a potentially transformative way to address the rigid requirements on samples in geographic analysis, particularly during this age when the collection and provision of geographic data are crowd-sourced and VGI-based. These differences also point to the potential of the third principle opening up a space of inquiry that would resonate more successfully with place-based approaches in human geography.

\textbf{1. Introduction}

There have been many debates and discussions about whether or not geography is a field of nomothetic or idiographic study (Perera 1970; Guelke 1977; Phillips 2004; Cresswell 2013). From a more nomothetic perspective, three geographic principles have been named as laws of geography. Tobler (1970) stated that spatial variation of geographic features/attributes follows a general principle of ‘near things are more related than distant things’ and named this principle (spatial autocorrelation) as the First Law of Geography. Goodchild (2004) stated that the spatial variation of geographic features/attributes also exhibits ‘uncontrolled variance’ and suggested that this principle (spatial heterogeneity) might be taken as the Second Law of Geography. These two principles together provide useful insights into the patterning of geographical phenomena in relation to spatial distance. When exploring the guiding principles in spatial prediction, Zhu et al. (2018) found that the similarity in geographic configuration between two locations could help address the theoretical challenges faced by the existing guiding principles in spatial prediction and named this geographic similarity principle as the Third Law of Geography. These three ‘laws’ together might capture the dynamics of a broader set of basic characteristics of geographic features/phenomena and in so doing, contribute to continued advancements in geographic analysis.

There has been a healthy debate about whether these and other general principles in geography should or should not be named as laws (Sui 2004; Barnes 2004; Goodchild 2004; Tobler 2004). In this paper, we deliberately avoid this debate, referring to these as principles not laws, and instead focus on what can be learned by a comparison of them. An explicit examination of the connections and differences among these three principles would help practitioners understand the essence of these geographic principles and how they can help in geographic analysis. Goodchild (2004) has ably discussed the connection and difference between the first
two principles (spatial autocorrelation and spatial heterogeneity). They can be taken as the two sides of spatial variation of geographic variables. The third principle has been only recently proposed, and its differences with the first two offer opportunities to think about geographic analysis in ways that broaden the intellectual space created by the first and second principles.

2. The three laws of geography

2.1 The First Law of Geography

Tobler (1970), in his paper, ‘A Computer Movie Simulating Urban Growth in the Detroit Region’, suggested that the population growth at place A depends not only the previous population at place A but also on the population of all other places, which is a geographic interpretation of the premise ‘everything is related to everything else’, emphasizing the connection (relationship) among the values of a geographic variable over space. He further added ‘but near things are more related than distant things’ to emphasize that the nature of such connections is ‘parochial, and ignores most of the world’. He referred to this principle of geographic variation, ‘everything is related to everything else, but near things are more related than distant things’, as the First Law of Geography, which has been generally referred to as such since (Longley et al. 2005, 49, 65). From his discussion and others’ subsequent use of the concept (Longley et al. 2005, 106), the First Law of Geography focuses on the connection between the attribute values of the same geographic variable over spatial distance.

Tobler was not the first person to explicitly relate the difference in an attribute value of a geographic variable to spatial distance. Krige (1951), later Matheron (1963), were among the first to formalize such a relationship and explicitly related the difference in the form of semi-variance in attribute value between two locations and the distance separating these two locations for spatial interpolation. Although these studies predate the introduction of the First Law by Tobler (1970), many practitioners in spatial interpolation currently treat the First Law of Geography as the theoretical basis for spatial interpolation (Longley et al. 2005, 333).

The First Law of Geography has often been equated to spatial autocorrelation (Longley et al. 2005, 87, 89; Getis 2010) and has been used to guide many geographic analysis (Longley et al. 2005, 45, 92, 104, 333, 336; Getis 2010). Even though it also contains situations where near things are more different (negative spatial autocorrelation) (Griffith 2019), which is almost the opposite to what captured in the First Law of Geography, spatial autocorrelation nevertheless expresses how attribute values of a given geographic variable are related to each other over spatial distance.

From the above discussion, one can conclude that the First Law of Geography, we will refer to it as the First Principle of Geography hereon, focuses on attribute variation along the dimension of spatial distance and expresses the connection between attribute values as a function of spatial distance.

2.2 The Second Law of Geography

In discussing the role of laws in geographic information science, Goodchild (2004) stated that ‘spatial heterogeneity, or nonstationarity in the statistical meaning of that term, implies that geographic variables exhibit uncontrolled variance’ and suggested that the principle of spatial heterogeneity might be taken as the Second Law of Geography. He elaborated ‘uncontrolled variance in space’ as the result of any analysis changes when one ‘move the study area’. From a statistical perspective, spatial heterogeneity is ‘structural instability as expressed by changing functional forms or varying parameters’ (Anselin 1988) from one spatial location to another or the ‘uniqueness of geographic entities’ that cannot be represented ‘by the mean or global representation (average)’ (Jiang 2015). In these senses, the second principle states that the outcome of a geographic variable varies by location or rather is location-specific, which can be taken as the ‘first-order effect, concerning the places taken one at a time’ (Goodchild 2004).

Spatial heterogeneity can also be applied to the ‘second-order effect of places’ when the outcomes of a geographic variable are compared between two locations. This can be extended to what is expressed by the First Principle of Geography, to mean that the connection between attribute values and the distances of places (spatial autocorrelation) is unavoidably varying among pairs of locations as well as from one orientation (direction) to another orientation of the pairs. For example, the spatial variation of temperature over a flat area would be different from that over a mountainous area. For another example, if we have pollutants dispersing from upstream of a river, the variation of the concentration of pollutants along the direction of flow is different from the direction perpendicular to the flow. This anisotropic nature of the geographic variation is another excellent example of spatial heterogeneity.

In sum, Second Law henceforth referred to as the Second Principle, is about the variability of geographic variables over space. Such variability is by nature ‘uncontrolled’ and cannot be captured or expressed using
statistical means. This uncontrolled variability applies either as a first-order effect of places (a function of place (location)) or as a second-order effect when referring to the uncontrolled variance of the connection (relationship) between attribute value and distance (spatial dependence, spatial autocorrelation).

2.3 The Third Law of Geography

Zhu and his colleagues (Zhu et al. 2018, 2020) examined the guiding principles used in the spatial prediction and found that if the environmental configurations (combination and arrangement of relevant geographic conditions around a location) related to a given geographic variable at a point is similar to that of a sample, then the value of the given geographic variable at that point is also similar to the value of that variable at the sample point. They summarized this finding as 'The more similar geographic configurations of two points (areas), the more similar the values (processes) of the target variable at these two points (areas) and named it as the similarity principle or the Third Law of Geography. For simplicity, the Third Law is simplified here as ‘the more similar in the geographic configuration, the more similar in the value of the target geographic variable'. What is expressed by the Third Law can be found in other studies. For example, in soil science, it has been well recognized that if soil formation conditions (a set of environmental conditions related to soil formation such as temperature, precipitation, vegetation, parent materials and topography) are similar between two locations, then the soil formation processes would be similar, leading to similar soil properties at the two sites (Dokuchaev 1883; Jenny 1994). The niche concept in ecology, used to identify locations suitable for different individuals of a particular tree or animal species, is another example of this third principle (Polechová and Storch 2008; Whittaker 1975).

There are two important foundational elements of the Third Law: the definition of geographic configuration and the comparative nature in seeking similarity. According to Zhu et al. (2018), geographic context is expressed in geographic configuration, which is composed of three aspects: a list of geographic covariates, the hierarchy of the covariates and the spatial structure of covariates. The list of geographic covariates is a list of geographic variables that spatially covary with the target geographic variable. For a different target geographic variable, the list of covariates used to define the geographic configuration will be different. For example, if the target variable is about soil condition (such as soil texture, organic matter content), then the list would include those geographic variables related to the development of soil, such as climate variables, geological variables, topographic variables and time (if possible). If the target variable is about the spread of a pandemic, then the list would include those geographic variables tied to the spread of the pandemic, such as connectivity, among others. Another clarification about the list of covariates is that these covariates not only include those only describing in-situ conditions at the location, but also those quantifying the spatial relations (spatial context) to the location (such as around, connected or distance to).

The hierarchy of the geographic covariates in geographic configuration refers to the importance of each covariate in impacting the processes related to the geographic variable. For example, for target variables related to soil, climatic conditions (such as temperature and precipitation, among others) play a more important role in soil formation, thus are at a higher level in the hierarchy of the configuration than topographic covariates. Hierarchy in the geographic configuration is again target variable dependent.

The spatial structure of covariates in the context of the geographic configuration includes two parts: spatial granularity (footprint or neighbourhood) of the relevant processes and spatial arrangement of covariate conditions. Spatial granularity is about the spatial extent over which the geographic processes related to the target variable manifest themselves. For example, soil formation processes require a certain spatial extent to interact with each other for the soil to develop. A tree requires a sizable area for it to get enough water and nutrients to grow. Clearly, spatial granularity is different for different target variables. This mirrors the work done by human geographers who point to not only the importance of the scale of observation but also of the scaled nature of geographic processes and objects themselves (Sayre 2005; McMaster and Sheppard 2004).

The spatial arrangement of covariate conditions refers to the spatial layout of the values of covariates either in the spatial granularity or around the location. For example, for the soil erosion target variable, the spatial arrangement of slope condition and vegetation condition in the spatial neighbourhood would heavily impact the amount of soil eroded from that neighbourhood. Vegetation at the lower part of the slope in the neighbourhood would block the soil particles from being moved out from the neighbourhood, thus making soil erosion from the neighbourhood more difficult than vegetation at the upper part of the slope in the neighbourhood. Thus, different spatial arrangements of covariate conditions in the neighbourhood or around the neighbourhood would have different impacts on the outcome of the target variable.
The comparative nature of the Third Law referred to as the Third Principle of Geography from hereon, is to seek the similarity in geographic configurations between two locations through comparison and then relate this similarity to the similarity in the attribute value (even in processes) of the target variable between the two points. If one of the two locations is a sample and at which the attribute value, even the process, of the target variable is known, then it is possible to infer the attribute value or the status of the process of the target variable at the other location (the unknown) through the comparison of geographic configurations at the two locations. In this way, the comparison is only applied to two locations, and it does not call for an explicit relationship in some mathematical forms between the geographic configuration and the value of the target variable to be established.

3. How the third principle different from the first and second?

The first difference among the above three major principles is the perspective from which they each examine geographic phenomena. The First Principle focuses on the spatial connection of geographic phenomena through a spatial distance perspective. The Second Principle emphasizes the variability (heterogeneity) of geographic phenomena at different places and the variability of the connections expressed in the First Principle. The Third Principle examines geographic phenomena from the perspective of the geographic context (geographic configuration) in which the given phenomenon resides. The second difference is how the connections among geographic phenomena at different locations are expressed. The Third Principle uses a comparative way to express the similarity of geographic phenomena rather than develop functional relationships over distance or assess the variability of such functional relationships. The third difference is how knowledge in geographic samples is treated. The First and Second Principles focus on the global representativeness of a set of samples in terms of means and variances, while the Third Principle, through paired comparisons, exploits the individual representation of particular, context-specific samples (individual representativeness).

3.1 Geographic context vs spatial dimension alone

As stated above, the First and Second Principles of Geography express the nature of variation of geographic phenomena along the spatial dimension (spatial variation). The First Principle focuses on the spatial continuity of spatial variation and is commonly applied through spatial autocorrelation. Its uniqueness is clearly reflected in ‘spatial’, ‘auto’ and ‘correlation’, indicating that the value of the geographic variable (such as temperature) is self-correlated over space. This correlation is a function of distance. The strictest interpretation of the First Principle implies that the values of a geographic variable at two locations are only related to the distance separating these two points. While other interpretations of the First Principle might exist, many spatial analytical methods based on this particular interpretation have been developed and widely applied, particularly in the field of geographic information science. Some of these methods (such as inverse distance weighting and kriging) do attribute geographic variation purely as a function of distance.

The ‘uncontrolled variance’ perspective of the Second Principle can be interpreted in two different ways: a) geographic characteristics (status) vary from location to location or what can be called conventional spatial variation (first-order), and b) nonstationarity of spatial autocorrelation (second-order). The latter is an important consideration in geographic information science, where quantitative analyses relying on spatial autocorrelation are much more common. The revisions (such as box kriging and directional kriging) to the ordinary kriging in spatial interpolation are great examples of the combination of the First and Second Principles of Geography (Isaaks and Srivastava 1989). Therefore, from the perspectives of GIScience, it is fair to say that the Second Principle not only emphasizes the spatial dimension of geographic phenomena but also focuses on the varying nature of spatial autocorrelation.

With the Third Principle, the characteristics of a geographic variable at a location is examined through the similarity in ‘geographic configuration’ (used interchangeably with ‘geographic environment’) to a known location (a location at which the characteristics of that geographic variable is known, seen as an example or sample). Through its use of ‘geographic configuration’, the Third Principle emphasizes the geographic context that affects the status of a given geographic variable at the location of interest. This context provides the potential to study the characteristics of a target geographic variable at a given location or over an area under the concept of interaction of the geographic factors that have correlations (or inexplicit causation) with this target variable. Geographic configuration is consistent with the ecological concept of ‘niche’ (Polechová and Storch 2008) or the human geographical concept of place that points to not only the interaction of overlapping variables in space but the interaction’s unique emergent properties (e.g. Sack 1997).
3.2 Comparison vs relationship

The First Principle states that the values of a geographic variable at two locations are related to each other as a function of the distance separating these two locations. The application of the First Principle in the field of GIScience, i.e. spatial analysis in a quantitative sense, requires specific forms of functions to be prescribed for characterizing this spatial relationship. For example, in spatial interpolation (a type of spatial analysis), the well-known inverse distance weighting method employs a distance decay function to express this relationship (Equation 1).

\[
\hat{z}_0 = \sum_{j=1}^{n} \frac{1}{\sum_{i=1}^{n} d_{ij}^q} * z_i
\]  

(1)

Where \( \hat{z}_0 = \sum_{j=1}^{n} \frac{1}{\sum_{i=1}^{n} d_{ij}^q} * z_i \) is the attribute value to be predicted at the unvisited-site 0, \( \hat{z}_0 = \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} d_{ij}^q} * z_i \) is the attribute value at sample point \( i \), \( n \) is the total number of sample points involved, \( \hat{z}_0 = \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} d_{ij}^q} * z_i \) is the distance between prediction location 0 and sample point \( i \), \( j \) is the counter referring to other sample points, \( attributevalueatsamplepoint_j \) is the distance decay coefficient describing how distance impacts the relationship between the attribute values at two points. The component \( \frac{1}{\sum_{j=1}^{n} d_{ij}^q} \) in Equation 1 describes this relationship as a function of distance \( \left( \sum_{j=1}^{n} d_{ij}^q \right) \). Clearly, the distance decay coefficient \( q \) controls the nature of this relationship. From this analysis, one can see that the characteristics of geographic phenomena (spatial autocorrelation) captured by the First Principle are expressed as some form of relational function in terms of distance. The different values of \( q \) can be taken as the different forms of spatial autocorrelation. Different forms of functions are also used in other types of spatial analysis (such as point pattern analysis and spatial interaction models) (Baddeley and Rubak 2016; Fotheringham et al., 2000; Fotheringham and O’Kelly 1989; Gatrell et al. 1996). Clearly, different spatial analytical techniques will characterize spatial autocorrelation quite differently, as manifested in the Second Principle of Geography.

The essential idea expressed in the Third Principle of Geography is that the more similar in geographic configuration, the more similar the value of the target geographic variable. This resonates with human and physical geographic understandings that spatial variation in a variable is not tied solely to spatial proximity to locations but to covariation with other variables to which it has direct or indirect relations (Sayer 1984). Figure 1 illustrates this idea. The condition (status or value) of the target variable at Location B is more similar to that of the target variable at Location A if geographic configurations between the two sites are more similar because this would indicate that the processes leading to the conditions of the target variable at these two sites are similar. Its emphasis is on the similarity of geographic configuration. It does not state a particular functional relationship between the geographic configuration and the value of the target variable (which is what the regression models do), nor define a relationship among the values of the target variables between locations (which is what the First Principle stipulates). However, by comparison, the Third Principle implicitly expresses the similarity between the interaction among the

![Figure 1](https://via.placeholder.com/150)  

Figure 1. The comparative nature of the Third Principle of Geography: the more similar in geographic configuration, the similar in value of the target variable.
geographic configuration and the target variable at the known site (geographic example) and the interaction at the unknown site.

One may argue that the computation of similarity in geographic configuration requires quantitative functions. However, these functions are used to compute the differences in geographic configuration between two points as a measure of their similarity, but not to define a quantitative form relating the attribute values to distances for two points. For example, if elevation is one of the covariates in characterizing the geographic configuration for a site, then the standardized subtraction function (Equation 2), commonly used to compute the difference in elevation as an inverse expression of similarity between two locations based on the elevation variable alone, should not be considered as-a function defining the relationship between the elevations at the two locations, nor the relationships between elevation and the value of the target variable.

\[
d^e_{ij} = \frac{|e_i - e_j|}{\sigma_e}
\]

Where \(d^e_{ij}\) is the difference in elevation (e) between location \(i\) and location \(j\), \(e_i\) and \(e_j\) are the elevation values at location \(i\) and location \(j\), respectively, \(\sigma_e\) is some version of standard deviation in elevation defined by the user to standardize the difference (Zhu et al. 2015a).

One may still argue that functions used to calculate differences to estimate similarity may be specific for different situations. It might be true that these functions may be different for different types of covariates (categorical vs. continuous) (Zhu et al. 2015a), or even different for different covariates, but they should be fairly constant for the same covariate over different places and at different times. We suspect that even with different forms of functions for computing differences to estimate the similarity based on a single covariate, the final effects on the analytical outcome should be very minimal as long as the form is constantly applied across that covariate. In other words, what specific ruler is used to measure the differences in length may not be substantive as long as the ruler is applied consistently. This does not hold true for the functions needed to define spatial autocorrelation.

The use of similarity also allows the interaction of the covariates at the location to be quantitatively expressed. When computing the similarity between two locations, the similarities based on individual covariates mentioned above are integrated to produce a final similarity between the two sites. This integration in the aggregation of similarities based on individual covariates would allow geographers more flexibility to incorporate their understandings about the nature of this interaction by weighting some variables more heavily than others. The mechanism for expressing this type of knowledge is not available with methods based solely on the first two principles.

The above discussion rests on the key difference between similarity in comparison and relationship as a function. With relationship, the emphasis is on the connection between two entities, often two variables, and this emphasis has the connotation for this connection to be applied to locations beyond those where such a relationship is originally defined (such as providing an educated guess as to what is at the location between sample locations). With similarity through comparison, the emphasis is on the resemblance in a comparable aspect, often between two and only two instances of a variable, and is often limited to particular instances, not extended to other instances.

The First Principle focuses on the relationship between two variables (the difference in attribute value between two locations and the distance between the two locations), which is often applied to distances not covered by the samples used. The Third Principle focuses on the resemblance of the instance of geographic configuration at a location to that of geographic configuration at another location, and this resemblance is confined to a specific pair of locations, not applied to other instances of geographic configuration. This feature is more consistent with current understandings in both physical and human geography that reject the nomothetic/ideographic dichotomy and instead seek highly qualified and conditional generalization from empirical study.

### 3.3 Individual vs. average

As indicated in Section 3.2, ‘relationship’ is tied to variables while ‘similarity’ is tied to specific pair of instances. Thus, characterization of spatial autocorrelation is more about finding a function that best fits, in a statistical or mathematical sense, the difference in an attribute value of a geographic variable and the distance between two points. The coupling of differences with distances between specific pairs of points is not as useful for spatial analysis because the interesting (meaningful) tasks are to be able to figure out conditions over locations where such couplings are not observed. For example, in the prediction of spatial variation of pollution concentration under the First Principle, we couple the difference in pollution concentration and the distance between a specific pair of observed locations in order to know what the pollution levels are at the points where observations were not made. This requires us to come up with a function
describing how the difference in concentration of pollution is related to the distance between two points where ‘distance’ can be varied (a variable) so that we can estimate the pollution level at a location where no observation was made, but its distances to some known points (samples) are known.

Clearly, derivation of such a function based on one pair of points is highly susceptible to the specific situation of that pair, and the so-derived function does not represent the coupling of difference with distance at other pairs well, thus would not work well for other locations. Thus, the tendency is to obtain more of these couplings over many pairs, ideally with pairs covering a wide range of distances so that the derived function is more reliable (representative) for other distances that were not covered in pairs observed for developing the function. This calls for more pairs and thus more sample points. Geographic distribution (variation) occurs not only in one dimension but often in two (easting, northing) or three dimensions (easting, northing, height/depth), even in four dimensions (the three-plus the time dimension). This calls for the samples to be distributed not simply along the distance dimension but also in the spatial and temporal dimensions so that the characterized relationship (function) is representative in these other dimensions as well. These are the reasons that spatial analysis based on the First Principle requires the samples to be in sufficient number across a sufficiently wide distribution to represent these dimensions well.

It is inevitable that the derived function based on so many pairs from these different dimensions is an average condition. The stationarity requirement imposed on spatial analysis is to make sure that practitioners know that deviation in the study area from such average function is minimal so that the application of the derived function would not result in substantial errors. Clearly, a consideration of the Second Principle reveals that it is difficult for geographic applications to meet this stationarity requirement. Thus, various alterations were made to limit the extent of averaging or the application of these resultant average relationships. For example, the various versions of kriging in spatial interpolation (such directional kriging, box kriging) were developed to minimize the chance of violating the stationarity assumption (Isaaks and Srivastava 1989). The geographic weighted regression method (Fotheringham et al., 2000) and spatial autoregressive models (Anselin 1988) are other good examples of avoiding the violation of the stationarity assumption of spatial analysis.

Characterization of similarity under the Third Principle is between two specific instances, two geographic configurations with each at a particular location. The utility of similarity is in its comparative nature, not interpolation as with the methods based on the First Principle. For example, in wildlife habitat mapping, if we know a location where a particular type of wildlife (say snub-nosed monkeys) ‘prefers’ to use as their favourable habitat (It is a separate issue from how this ‘favourable habitat’ is defined), then under the Third Principle of Geography we would treat any location that is similar to the current location in geographic configuration (defined as relevant to snub-nosed monkey habitat) would be a good candidate for a good habitat location for snub-nosed monkeys. There are two important aspects in this example that need to be highlighted. The first is that the similarity is defined between the known location and an unknown location in terms of their similarity in geographic configuration. The similarity between this known location and another unknown location is redefined based on their own difference in geographic configuration, not derived from the similarity from any other pairs of points. In other words, the similarity is solely between the specific pair of points and is independent of similarity from any other pairs of points.

The second, the status (the monkey habitat suitability) of the unknown location, is determined in a comparative manner using the similarity in geographic configuration with the known point, the more similar the geographic configuration between the two locations, then the more similar the habitat suitability between the two locations. If the known location (sample) is of high suitability and the similarity of an unknown location to this sample in geographic configuration is high, then the habitat suitability at the unknown location would also be high. If the known location is of low habitat suitability and the similarity in geographic configuration is high between the two locations, then the unknown location would have low habitat suitability. However, when the known location is of high suitability, but the similarity in geographic configuration between the two locations is low, we cannot say that the unknown location is of low or high suitability due to the fact that these two points are not similar in geographic configuration relevant to monkey habitat, which means that we do not have sufficient information to say much about the suitability at the unknown location because the information we have is not relevant to that unknown location.

Two observations can be made about using the Third Principle. The first observation is that the analysis is based on the representation of an individual sample and only applied to areas (or locations) that are similar
to this sample. The second observation is that the level of similarity can be used to measure the level of confidence (opposite of uncertainty) about our assessment at the unknown location. The more similar the unknown location is to the known location, the less uncertain we are about our assessment at the unknown location. The lower the similarity, the more uncertain we are about our assessment at the unknown location. Thus, we can provide an uncertainty value to every location in the study area based on its similarity to the known location. If we have a set of known points (samples) in the study area, we can compute for every location the overall uncertainty given the set of known samples. The collection of overall uncertainty at each location across the study area forms a map showing how well each location is represented by the set of samples we currently have. These two observations lead to an important distinction between the Third Principle and the first two principles that spatial analysis based on the Third Principle of Geography may not have specific requirements on the number and the distribution of samples and may not demand the stationarity assumption to hold because methods based on the Third Principle of Geography do not take the ‘best fit approach’ or ‘averaging’. The removal of these requirements is important for geographic analysis.

4. Implications for geography broadly

By discussing the nexus of these three principles, we hope to open up analytical/conceptual space to think more broadly about geographical phenomena and our understanding of them in quantitative and qualitative ways. First, due to the vastness (in geographic space) and the complexity (in feature space) of geographic phenomena, samples are one important source of information geographers depend on for studying geographic phenomena and their processes. With existing geographic analytical frameworks, the samples used need to meet two important requirements: they are sufficient in number and adequately distributed over space so that the samples are representative over the study area and the resulting findings are meaningful (representative and generalizable). However, such requirements are rarely met in many geographic analyses due to the financial and human resources required to collect such kind of sample sets. Thus, findings from insufficient and biased sample sets run the risks of drawing incomplete or biased conclusions. In addition, these requirements post a major challenge for geographers to make good use of the new form of geospatial data provided through spatial big data (Graham and Shelton 2013) or data rich-environments (Goodchild 2007; Miller and Goodchild 2015) because these geospatial data (particularly samples) are often not representative over the study area (Goodchild and Li 2012; Zhu et al. 2015b). The comparative nature as stipulated in the Third Principle of Geography, which does not require samples to be of specific distribution and in prescribed sufficient number, might provide an effective means to handle this type of data. Examples in habitat suitability assessment, landslide susceptibility mapping and soil mapping have shown that the Third Principle of Geography is effective in improving the accuracy of mapping using spatially biased data (Zhang and Zhu 2020; Zhu et al. 2019, 2015a,b).

The second implication is that the complexity of geographic phenomena and processes not only exhibit themselves over the spatial domain in a discernable pattern but also in the interaction of driving factors. Deriving simple functional relationships to characterize such complexity is not only difficult but might also be insufficient. This is particularly the case for many non-material variables of interest to human geographers – variables that do not diffuse, show discernable spatial patterns and that span but do not cover space – very much shape the spatial patterns of geographical phenomena (e.g. Leitner et al., 2002). The reluctance in exploring geography as holding more generalizable patterns and processes may reflect the early emphasis on simply the space dimension as a unifying frame as envisioned through the highly reduced dyad of spatial autocorrelation and spatial heterogeneity (even prior to the formal invocation of the Second Principle). Reducing geography to simply the study of the spatial patternning of geographical features in relation to spatial parameters is, for most geographers, uninteresting and divorced from what they seek to understand.

The similarity-based comparative nature, as expressed in the Third Principle of Geography, retains the individuality of geographic phenomena while still providing a structure to assess the interplay or covariance of a broader set of factors. This would provide an angle that is similar to how many geographers think of the spatial framing of the processes they study. Whether we speak of site, ecological community, or place, physical and human geographers alike point to these locations as conjunctures in time and space where various biophysical and social processes or networks interact with key properties emerging from these interactions (e.g. Marston et al., 2005; Whittaker 1975). Put more simply, historical and geographical contexts attached to places matter in human and nonhuman experience. At the same time, there is a view that all places or sites are not unique and that spatial structure exists, reflecting the fact that variables constituting geographical
configurations in Third Principle parlance show different degrees of spatial autocorrelation at different scales. This lies at the heart of ongoing debates concerning hierarchical treatments of scale in ecology, physical geography and human geography (Allen and Hoekstra 1992; Manson 2008; Birkenholtz 2011; Brown and Purcell 2005; Swyngedouw 2004; Neumann 2009).

In sum, the differences of the Third Principle from the first two open up new analytical space more likely to engage with broader understandings of place-based variation of geographic phenomena held in physical and human geography. To realize this potential, several issues need to be addressed or explored. First, the Third Principle-inspired analyses use the similarity of geographical configuration to estimate the potential for certain geographical outcomes to occur across geographical space through examining the similarity implicitly in the processes leading to these geographical outcomes between a known location and an unknown location. To make this principle feasible for explicitly seeking a greater understanding of the processes and causal pathways that lead to these geographical outcomes: an understanding that motivates much physical and human geography work, similarity measures that can explicitly measure geographic processes or pathways needs to be developed.

Second, the current application of the Third Principle is based on the similarity of variables that can be measured easily using existing quantitative methods. Application of the Third Principle in human geography would call for similarity being measured on variables that are on ordinal or cardinal scales and ‘variables’ that are less reducible to parameters due to their relational nature (e.g. power, vulnerability, and identity), and contingency. With similarity under the Third Principle measured in a comparative nature, rather than in the rigid quantitative relationships, it is possible to measure similarity on variables that expressed as qualitative terms in natural language (Liu and Zhu 2009; Gao et al. 2017). The development of these types of measures would vastly improve the application of the Third Principle to realist epistemologies where key processes, relations or networks that are not easily quantifiable using the methods from the quantitative revolution of the 1960s in geography.

Third, as outlined in this paper, the employment of similarity in geographic configuration reflects a rejection of mechanistic regression approaches in spatial analysis that have real problems of causal inference. This stance resonates with concerns expressed by geographers about causal inference by realist accounts of geographical phenomena as insightfully explored by Andrew Sayer (1984). There is a major dilemma facing applications of the Third Principle and geospatial analysis more broadly: how to use its quantitative descriptions of similarity to contribute to the building of greater understandings of the processes that produce these patterns without falling into the causal inference trap. This is an important question to be considered, given that history has shown that even if the Third Principle formally refers to covariation of target and geographical configuration variables, it is likely to be used to make causal statements about the world. Within a realist framing, quantitative spatial analysis has significant limitations of causal inference unless coupled with substantive work on the nature of relationships in place. One could imagine that the Third Principle-inspired spatial analysis could be combined with such substantive work in a myriad of ways such as using similarity analyses to uncover patterns useful to identify new lines of substantive inquiry or to bring various types of place-based knowledge of relations (expert or lay) into similarity analysis. From a human geographical perspective, such work would have the potential to add to our understandings. One possible way to address this issue is to thoroughly examine the processes and pathways from the knowns (the examples) that the Third Principle relies on and then seek the dis-similarity with the unknown locations for clues to processes and pathways that are different from these knowns. This type of effort would mitigate the concerns that many geographers have with big data analysis that seeks to predict patterns of human or physical geographical outcomes from data sets without attention to the underlying processes that produce these outcomes.

5. Summary

The contrast between the first two principles of Geography and the Third Principle of Geography is strong, with substantial differences. Spatial dimension (variation) is an important property of geographic phenomena and is one of the foundational elements of geography as a field. It is natural and important to highlight the general principles geographic phenomena exhibit along this dimension, just as the First and Second Principles of Geography stipulated. However, geographic phenomena/processes are impacted by more factors in a more interactive fashion than by a single spatial dimension. The Third Principle of Geography provides a way (in the form of geographic context and similarity to the knowns) to consider this
interactive and integrative nature and thus, together with the first two principles forms a more complete theoretical framework of geography.

The Third Principle of Geography also explores a new direction of spatial analysis by emphasizing similarity in a comparative approach other than the traditional relationship approach. The similarity in a comparative approach is not dependent on a specific relational function to characterize the expressed principle when the principle is applied in spatial analysis. In addition, the comparative approach is instance-oriented and does not seek for functional relationships to be representative of the population over the study area. Thus, the similarity-based comparative approach as stipulated in the Third Principle is more flexible when applied because it does not require samples to be over a certain number and with particular spatial distributions. The similarity measure also provides uncertainty information about the assessment at a location and thus would naturally provide explicit information on how well the currently available information (samples) cover the geographic phenomena over the area of interest. These flexibilities provide a potential for the application of the Third Principle of Geography to realist epistemologies where key processes, relations, or networks that underlie target outcomes of interest are expressed with variables that cannot be measured using conventional means.

Acknowledgements

Supports through the Vilas Associate Award, the Hammel Faculty Fellow Award, the Manasse Chair Professorship, the Distinguished Professorship from the University of Wisconsin-Madison are greatly appreciated.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work reported here was supported by grants from the National Natural Science Foundation of China (Project No.: 41871300) and from the 111 Program of China (Approved Number: D19002).

References

Allen, T. F. H., and T. W. Hoekstra. 1992. Toward a Unified Ecology. New York: Columbia University Press.

Anselin, L. 1988. Spatial Econometrics: Methods and Models, 119. New York: Springer.

Baddeley, A., and E. Rubak. 2016. Spatial Point Patterns: Methodology and Applications with R. Boca Raton; London; New York: CRC Press: Taylor & Francis Group.

Barnes, T. J. 2004. “A Paper Related to Everything but More Related to Local Things.” [J] Annals of the Association of American Geographers 94 (2): 278–283. doi:10.1111/j.1467-8306.2004.09402004.x.

Birkenholtz, T. 2011. “Network Political Ecology: Method and Theory in Climate Change Vulnerability and Adaptation Research.” Progress in Human Geography 36 (3): 295–315. doi:10.1177/0309132511421532.

Brown, J. C., and M. Purcell. 2005. “There’s Nothing Inherent about Scale: Political Ecology, the Local Trap, and the Politics of Development in the Brazilian Amazon.” Geoforum 36: 607–624.

Cresswell, T. 2013. Geographic Thought: A Critical Introduction. New York: Wiley-Blackwell.

Dokuchaev, V.V., 1883/1948/1967. Russian Chernozem, in, Selected works of V. V. Dokuchaev, and Moscow, 1883/1948/1967, 1, 14–419. Jerusalem: Israel Program for Scientific Translations Ltd. (for USDA-NSF), Publ. by S. Monson, 1967. (Transl. into English by N. Kaner).

Fotheringham, S., C. Brunsdon, and M. Charlton. 2000. Quantitative Geography: Perspectives on Spatial Data Analysis, 270. Los Angeles, CA 91320 United States: Sage Publications.

Fotheringham, S., and M. E. O’Kelly. 1989. Spatial Interaction Models: Formulations and Applications. Boston: Kluwer Academic Publishers.

Gao, S., K. Janowicz, D. R. Montello, Y. Hu, J.-A. Yang, G. McKenzie, Y. Ju, L. Gong, B. Adams, and B. Yan. 2017. “A Data-synthesis-driven Method for Detecting and Extracting Vague Cognitive Regions.” International Journal of Geographical Information Science 31 (6): 1245–1271. doi:10.1080/13658816.2016.1273357.

Gatreil, A. C., T. C. Bailey, P. J. Diggle, and B. S. Rowlingson. 1996. “Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology.” Transactions of the Institute of British Geographers 21 (1): 256–274. doi:10.2307/622936.

Getis, A. 2010. “Spatial Autocorrelation.” In Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications, edited by M. M. Fischer, and A. Getis. London, UK: Springer-Verlag Berlin Heidelberg 255–278. doi:10.1007/978-3-642-03647-7_14.

Goodchild, M. F., and L. Li. 2012. “Assuring the Quality of Volunteered Geographic Information.” Spatial Statistics 1: 110–120. doi:10.1016/j.spasta.2012.03.002.

Goodchild, M. F. 2004. “The Validity and Usefulness of Laws in Geographic Information Science and Geography.” Annals of the Association of American Geographers 94 (2): 300–303. doi:10.1111/j.1467-8306.2004.09402008.x.

Goodchild, M. 2007. “Citizens as Sensors: The World of Volunteered Geography.” GeoJournal 69 (4): 211–221. doi:10.1007/s10708-007-9111-y.

Graham, M., and T. Shelton. 2013. “Geography and the Future of Big Data, Big Data and the Future of Geography.” Dialogues in Human Geography 3 (3): 255–261. doi:10.1177/2043820613513121.

Griffith, D. A. 2019. “Negative Spatial Autocorrelation: One of the Most Neglected Concepts in Spatial Statistics.” Stats 2 (3): 388–415. doi:10.3390/stats2030027.
