ELSA: a new local indicator for spatial association

Hamm NAS[*1], Naimi B[†1] and Groen TA[‡3]

[1] School of Geographical Sciences, University of Nottingham, Ningbo, China
[2]Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland
[3]Faculty of Geo-Information Science and Earth Observation, University of Twente, the Netherlands

30 March 2021

Summary

There are several local indicators of spatial association (LISA) that allow exploration of local patterns in spatial data. Despite numerous situations where categorical variables are encountered, few attempts have been devoted to the development of methods to explore the local spatial pattern in categorical data. To our knowledge, there is no indicator of local spatial association that can be used for both continuous and categorical data. We introduce ELSA, which can be used for exploring and testing local spatial association for continuous and categorical variables. We provide the R-package elsa for making these computations.

KEYWORDS: LISA, categorical data, hierarchical classification, continuous data

1 Introduction

There are several local indicators of spatial association (LISA) that allow exploration of local patterns in spatial data. Despite numerous situations where categorical variables are encountered, few attempts have been devoted to the development of methods to explore the local spatial pattern in categorical data. In this paper we introduce the entropy-based local indicators of spatial association (ELSA). ELSA can be used for exploring and evaluating local spatial association for categorical variables, including categorical variables with different levels of similarity. We also show how ELSA can be applied to continuous data. We have written an R-package elsa for making these computations, which we have made publicly available.

ELSA is elaborated in full in Naimi et al. (2019), which also gives examples based on raster data. In this paper we summarize the key aspects of ELSA and present some new developments based on point data and ordinal data.

2 Methods

2.1 The ELSA statistics

ELSA (E) is defined as:

$$E_i(h) = E_{ai}(h) \times E_{ci}(h) \tag{1}$$

and is calculated within a local neighbourhood centred on location, i. E_{ai} summarizes the dissimilarity between x_i, the attribute at location i, and its neighbours, each denoted as x_j. Hence,

$$E_{ai}(h) = \frac{\sum_j w_{ij} d_{ij}}{\max\{d\} \sum_j w_{ij}}, j \neq i \tag{2}$$

[*] nick@hamm.org
[†] naimi.b@gmail.com
[‡] groen@itc.nl

where w_{ij} is a binary weights matrix, which describes whether j is within a specific distance, h, of i and d_{ij} is the dissimilarity between the pair of observations, x_i and x_j (see Section 2.2). E_{ai} takes values between 0 and 1 inclusive, where low values indicate high similarity between x_i and its neighbours and high values indicate a low similarity.

E_{ci} is the Shannon entropy at site i, normalized by $\log_2 m_i$:

$$E_{ci}(h) = -\frac{\sum_{k=1}^{m_w} p_k \log_2(p_k)}{\log_2 m_i}, j \neq i \tag{3}$$

$$m_i = \begin{cases} m \text{ if } \sum_j w_{ij} > m \\ \sum_j w_{ij}, \text{otherwise} \end{cases} \tag{4}$$

Where m is the total number of categories in the dataset and $p_k$is the probability of obtaining category k. This term quantifies the diversity of categories within the local neighbourhood. A high value indicates high diversity. A low value of E indicates a high level of spatial association.

2.2 Dissimilarity

Consider two nominal categorical variables, x_i, and x_j. If the two attributes are the same then $d_{ij} = 0$. If the two attributes are different then set $d_{ij} = 1$. This is the most simple case.

Often categories are organized hierarchically. For example we may have two categories and several sub-categories, as illustrated in Table 1. In this case we set $d_{a1,a2} = 1$ but $d_{a1,b2} = 2$. We consider two sub-categories in the same super category to be more similar than those from different super categories. This example is simplified from CORINE 2006 land cover map and can be extended to more than two levels.

Table 1 Example of hierarchical categories.

Code	Category	Sub-category
a1	Forest	Broad-leaved forest
a2		Coniferous forest
b1	Scrub	Natural grasslands
b2		Transitional woodland-scrub

We might also consider ordered categories (ordinal scale of measurement), such as household income or air quality. For example, air quality might be categorized as very poor (rank 4), poor (3), moderate (2) or good (1). In this example the maximum difference, $d_{ij} = |c_i - c_j| = 4 - 1 = 3$.

We extend the notion of ordered categories to handle continuous data on the interval or ratio scale. ELSA works with categories so we need to bin the continuous data into ordered categories. Clearly this will lead to a loss of information. We handle this by progressively dividing the data into a larger number of bins. At each step we determine the Spearman's rank correlation between the continuous and categorized data. We continue until a threshold is reached. An example of continuous data is air quality. For example, we could measure the concentration of PM2.5 in µg m^{-3} (particulate matter less than 2.5 µm in diameter) or aggregate over different pollutants to obtain an air quality index (AQI).

3 Demonstrations

We used air quality data for central and western mainland Europe. These were obtained from Airbase (*Air* quality data*base* for the European Economic Area, and are described in detail by Hamm et al. (2015). The data for 4 April 2009 (mean: 41.9, median: 37.3, minimum: 3.0, maximum: 120.7, standard deviation: 24.2, units: µg m^{-3}) are shown in Figure 1. This day is characterized by a high pollution event over north-east France, Belgium, the Netherlands and northern Germany. The rest of Europe has comparatively lower PM10 concentrations.

This example was chosen because it supports the evaluation of ELSA for both ratio scale continuous data and for ordered categories.

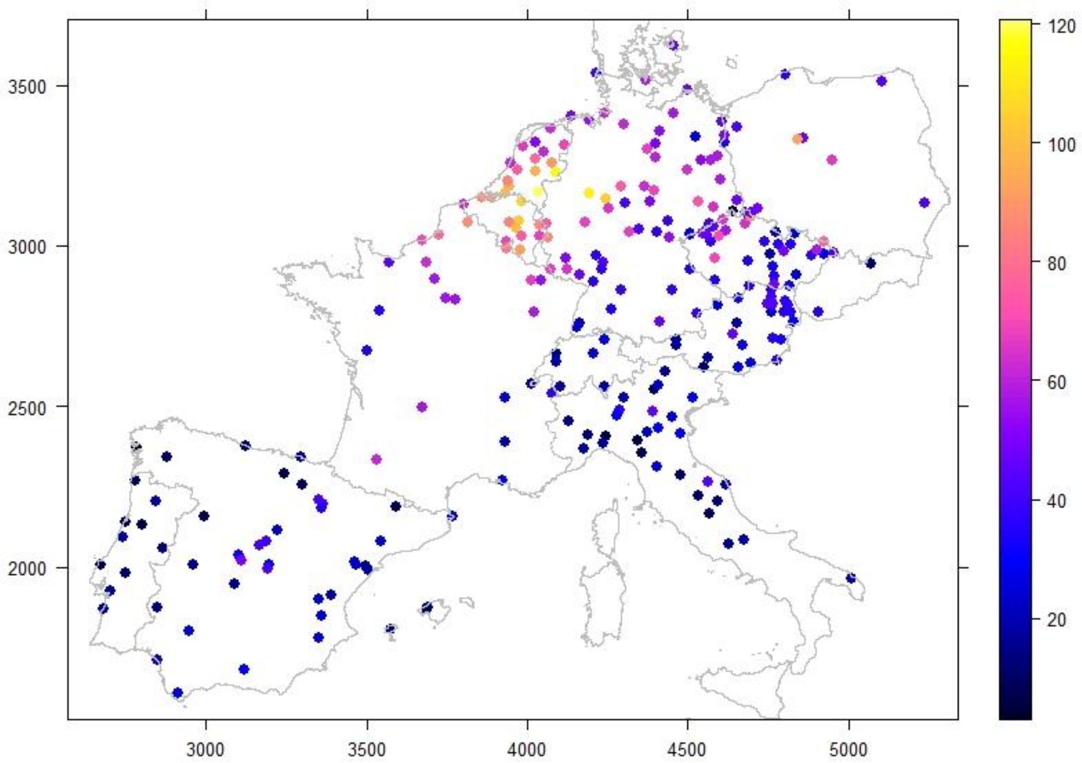

Figure 1 PM10 data from Airbase for 2009-04-04. Units are µg m^{-3}. The map projection is the ETRS89 Lambert Azimuthal Equal-Area (LAEA) projection (EPSG: 3035), with unit km.

We categorized the PM10 measurements into three levels – low, moderate and high – based on WHO (2005) and European Union guidelines. According to these guidelines PM10 should not exceed 20 µg m^{-3} on average over the year and should not exceed 50 µg m^{-3} on any given day. The upper threshold should not be exceeded more than 18 times in a year. This is illustrated in Figure 2.

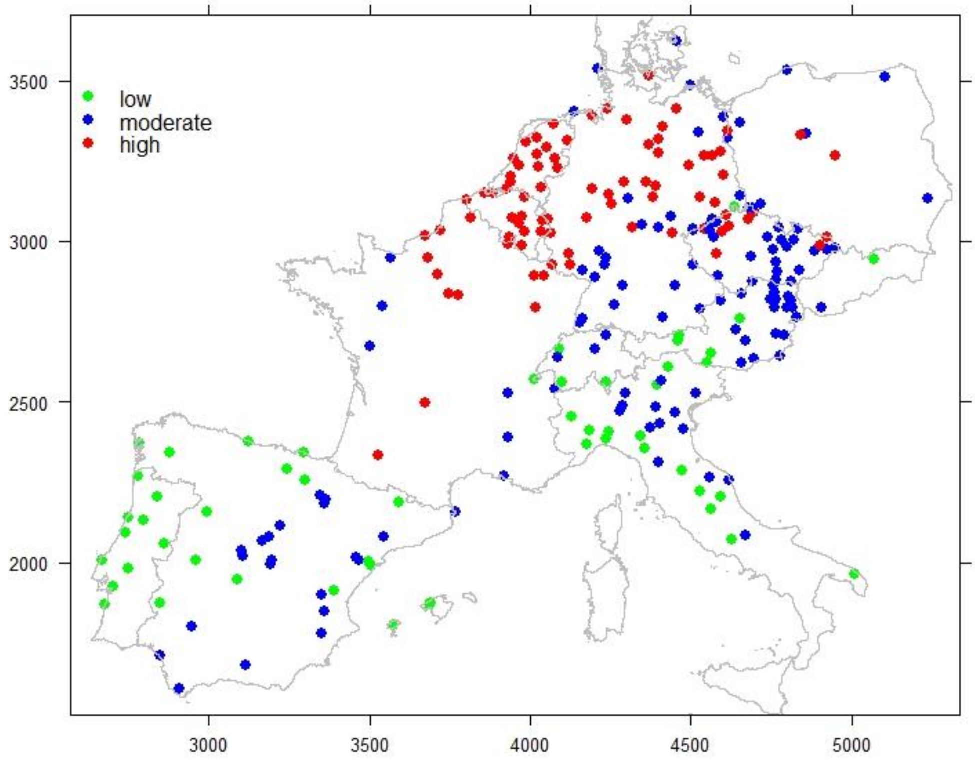

Figure 2 PM10 by ordered category. Low (< 20 µg m^{-3}), moderate (between 20 and 50 µg m^{-3}) and high (>50 µg m^{-3}) PM10. Projection same as Figure 1.

We first considered the calculation of ELSA within a local neighbourhood of $h = 150$ km. This is illustrated first for the continuous data in Figure 3. E_a summarizes the dissimilarity between an observation and its neighbours. This was lowest in Spain and Portugal and largest in central Germany and the Netherlands. E_c summarizes the composition or diversity of values within the neighbourhood. This showed a larger range of values. The largest values of E_c were found in northern Germany, the Netherlands and Belgium where there was a large range of high PM10 values. The lowest E_c values were found in Spain and Portugal, except for central Spain where there were some high E_c values. Finally ELSA (E) showed the lowest level of local spatial association in northern France and Germany, Belgium and the Netherlands and the highest spatial association was found in Portugal.

Next we repeated this exercise for the categorized data. This revealed much clearer patterns (Figure 4). The lowest values of E_a were found in Portugal, Belgium and the Netherlands reflecting a cluster of low and high values respectively. Indeed there were 70 locations where $E_a = 0$, which means that c_i had the same value as its neighbours. E_c showed the full range of values. The larger values tended to occurs at the borders between different air pollution classes. Finally ELSA showed that the highest degree of spatial association was found in western Spain and Portugal (low PM10 values), eastern Austria and the Czech Republic (moderate PM10 values) and Belgium, the Netherlands and north-west Germany (high PM10 values).

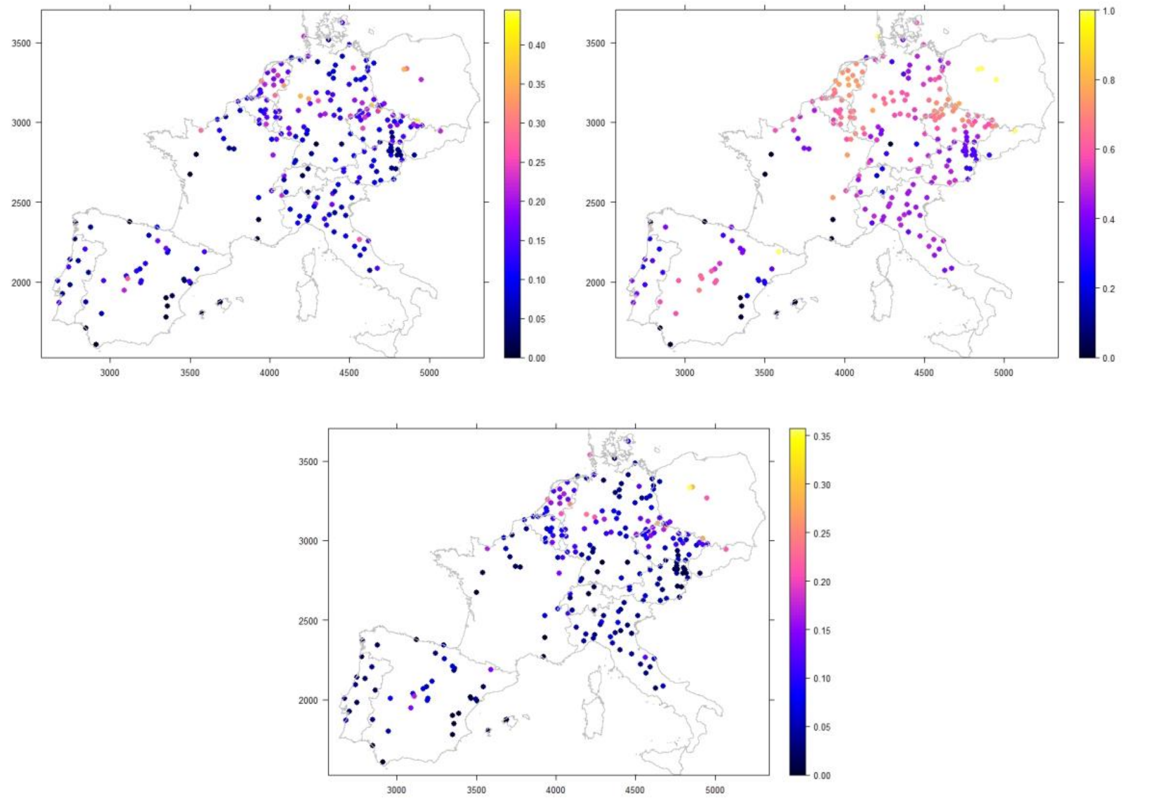

Figure 3 ELSA statistics computed for the continuous PM10 data (Figure 1) and $h = 150$ km: E_a (top left), E_c (top right), E (ELSA, bottom). Projection same as Figure 1.

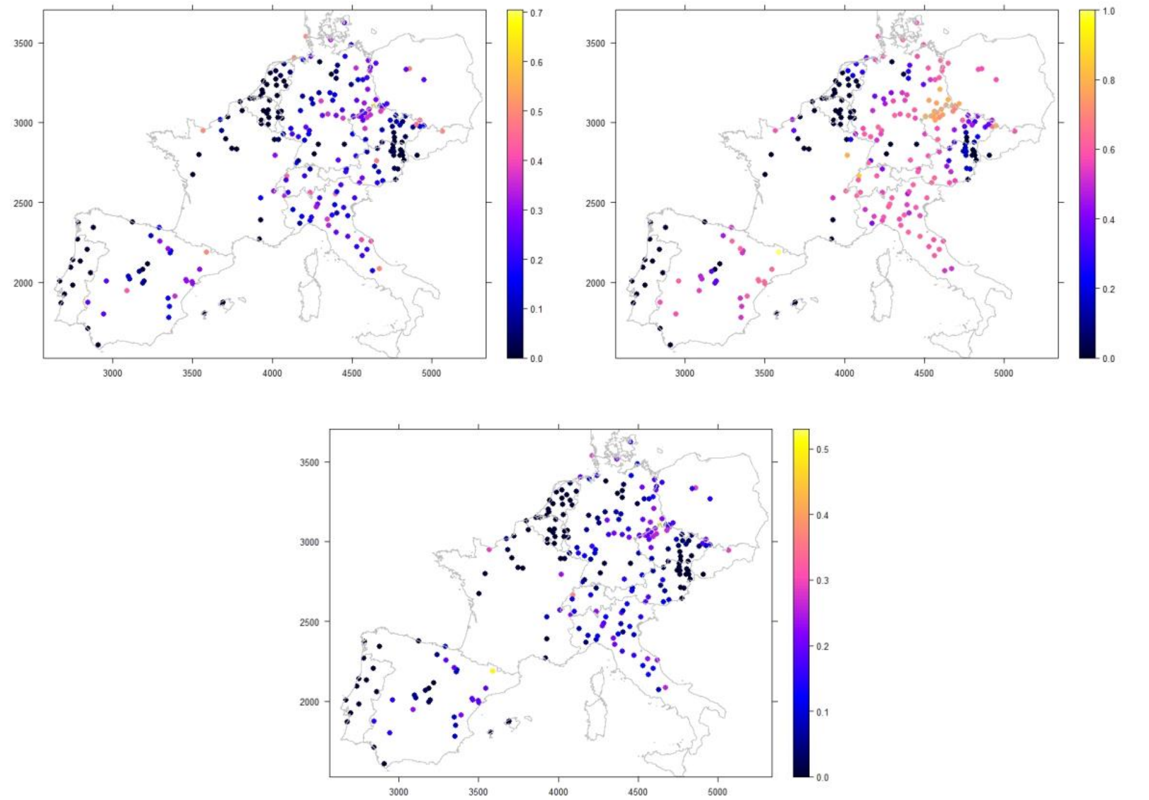

Figure 4 As for Figure 3, but computed for the categorized PM10 data (Figure 2).

Categorizing the data did lead to information loss. Notably the range of values in the high category (50 to 120 μg m^{-3}) was larger than range of values covering the low (0 to 20 μg m^{-3}) and moderate (20 to 50 μg m^{-3}) categories. However, it did allow us to better visualize the patterns of in these categories, which are important from both regulatory and health perspectives.

Finally, we explored the impact of changing, h. Increasing h increases the size of the window within which the ELSA statistics are calculated. This wass illustrated for ELSA (E) for $h = 150, 300, 450$ and 500 km (Figure 5) for the categorized data. Following Tobler's First Law of Geography, we expected that measurements that are close together in space to be more similar that distant measurements. Hence, increasing h was expected to lead to increased heterogeneity within the local window. We could then identify the scale at which similar categories tend to be clustered. As discussed above, for $h = 150$ km we identified three clusters. These could still be identified when we set $h = 300$ km. For $h = 450$ km the cluster of high values over Belgium and the Netherlands was clear, although the cluster over Portugal was less clear. For $h = 600$ km none of the clusters were clear. These changes reflected the size of the areas with low, moderate and high PM10 concentrations. This was approximately 300 km, except for the cluster over Belgium and the Netherlands. The shift of ELSA towards higher values with increasing h is illustrated in Figure 6.

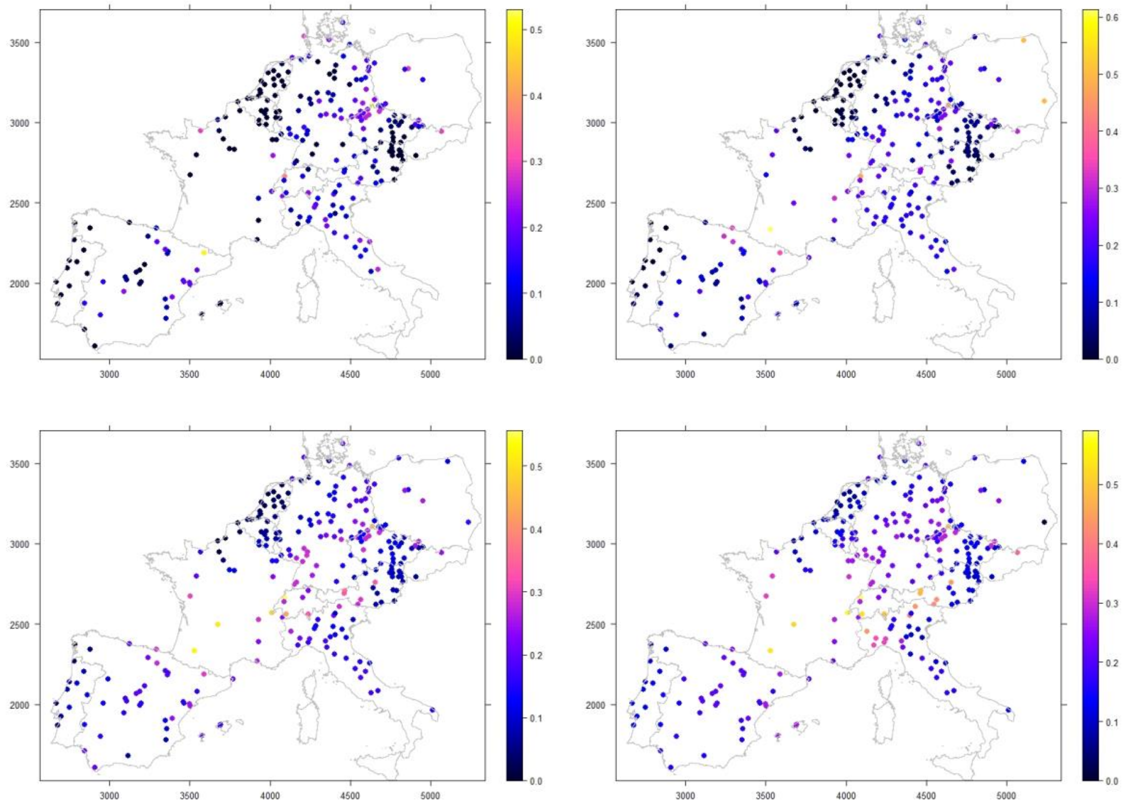

Figure 5 The ELSA (E) statistic for the categorized data for h = 150 km (top left), 300 km (top right), 450 km (bottom left) and 600 km (bottom right).

4 Conclusions

In this paper we introduced the ELSA statistics. We illustrated how these can be used to explore patterns in point observations of air pollution represented as ratio and ordinal data. This adds to our previous work (Naimi et al., 2019) which considered only raster data and did not consider ordinal data.

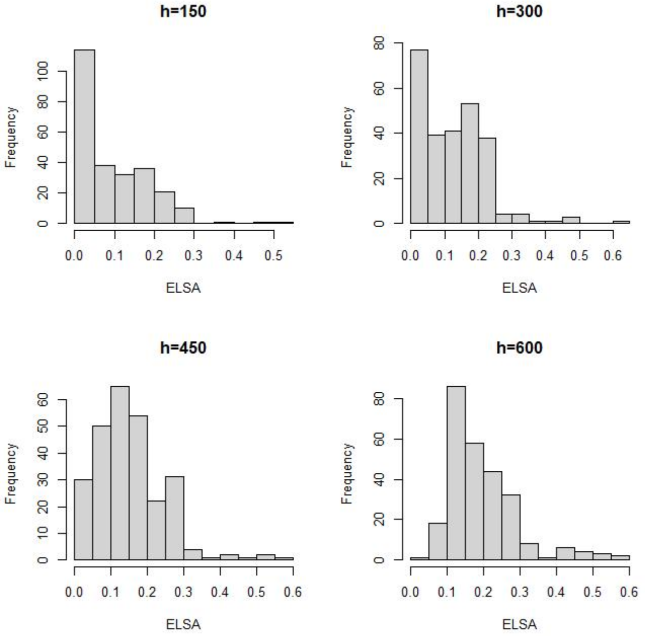

Figure 6 Histograms for the data shown in Figure 5.

5 Acknowledgements

We acknowledge Arjo Segers (TNO Netherlands) for his support with data acquisition. We are grateful to the partial support of the European Union, Erasmus Mundus programme External Cooperation Windows (2007/1139/001-001 MUN ECW).

References

Hamm NAS, Finley AO, Schaap M and Stein A (2015). A spatially varying coefficient model for mapping PM10 air quality at the European scale, Atmospheric Environment, 102, 393-405.

Naimi B, Hamm NAS, Groen TA, Skidmore AK, Toxopeus AG and Alibakhshi S (2019). ELSA: Entropy-based local indicator of spatial association. *Spatial Statistics*, 29, 66-88.

WHO (2005) WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Global update 2005: summary of risk assessment, World Health Organization.

Biographies

Dr Nicholas Hamm is an associate professor in the School of Geographical Sciences at the University of Nottingham, Ningbo, China. His research interests are in geospatial data science – in particular geostatistics, geospatial uncertainty and spatial data quality.

Dr Babak Naimi is a researcher in the Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland. His interests are in species distribution modelling, spatial data science, geoinformatics and spatial temporal analysis.

Dr Thomas Groen is an associate professor in the Department of Natural Resource Sciences, Faculty of Geo-Information Science (ITC), University of Twente, the Netherlands. His research interests are in species distribution modelling, remote sensing and ecology.