Spatial Autocorrelation Analysis with Graph Convolutional Neural Network

Pengyuan Liu*[1] and Stefano De Sabbata†[1]

[1]School of Geography, Geology and Environment, University of Leicester

March 31, 2021

Summary

Spatial autocorrelation statistics have a long-standing history being used by geographers to determine whether identifiable spatial patterns exist in data. However, existing research has identified that solely relying on p-values can be problematic when working with large datasets. This paper introduces a generalised model that can capture geographical data's spatial patterns using a graph convolutional network (GCN). The preliminary analysis demonstrates that GCN can capture the localities among areas in local-scale datasets by processing the data features and the spatial information separately into the graph network.

KEYWORDS: spatial autocorrelation, spatial statistics, graph convolutional network, deep learning, generalised model.

1 Introduction

Determining whether or not identifiable spatial patterns exist is a crucial step in spatial data analysis. According to Haining (2001, p. 14763), "spatial autocorrelation refers to the presence of systematic spatial variation in a mapped variable". Spatial autocorrelation is defined as positive if adjacent observations have similar values and negative if adjacent observations have contrasting values. The concept of spatial autocorrelation plays an important role in defining the discipline of spatial analysis. Since the mid-1990s, the idea of spatial autocorrelation was extended to local variation, which led to the development and use of local statistics (Getis, 2008). The latter can be defined as descriptive statistics whose value is calculated for each entity in a spatial dataset and focus on the relationship between each entity and its neighbours. The concept behind the statistical measurements for calculating spatial autocorrelation has a close connection to statistical theory, and a local statistic can be derived from "almost any standard statistic" (O'Sullivan and Unwin, 2014, p. 222). Some of the more commonly utilized algorithms in the study of spatial autocorrelation include: local Moran's I (Anselin, 1995), local Gi and Gi* statistic (Getis and Ord, 1992; Ord and

*pl164@leicester.ac.uk

†s.desabbata@leicester.ac.uk

Getis, 1995) and LISA (Anselin, 1995). Similar to many standard statistics, interpreting the results of local statistics involves the expected values of the statistical model and its statistical inference, which often uses p-value to quantify the idea of statistical significance as evidence that whether spatial autocorrelation exists in the data.

However, Lin et al. (2013) raised concerns about the use of p-value associated with large datasets, as p-values quickly tend to zero. Thus, solely relying on p-values and commonly used significance thresholds can lead to no sufficient grounds to support the results of statistical models. In geographic information analysis, those same concerns limit the scalability of local statistical models.

This paper aims to investigate the potential of semi-supervised learning approaches, and in particular deep learning neural networks for modelling spatial patterns (i.e., local spatial autocorrelation) in a large geographic area. To overcome the issues related to common approaches to assess the significance of statistical methods when working with large databases, we propose a deep learning model named graph convolution network (GCN) to investigate local object's features of input data. The source code used for this paper is available on GitHub[1].

2 Method: Graph Convolutional Neural Network

The overall workflow of the proposed method is shown in Figure 1. We frame the problem of classifying local statistics in such spatial matrix as a graph-based semi-supervised learning task (see next section for more detailed explanations) and graph convolution network (GCN) (Kipf and Welling, 2016) is suitable for modelling complex spatial patterns in geographic data. Spatial weights are a key component in the measurement of spatial autocorrelation statistics (O'Sullivan and Unwin, 2014). The weights encode the neighbouring relationship between the observations as a $n \times n$ matrix $\mathbf{A}$ where the elements A_{ij} of the matrix are the spatial weights.

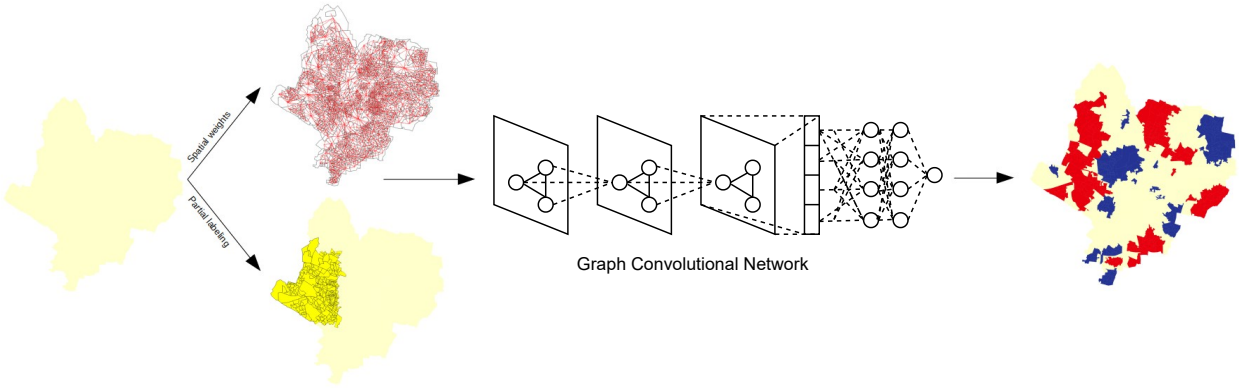


Figure 1: The workflow of the proposed method. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0. Contains OS data © Crown copyright and database right 2021.

[1] https://github.com/PengyuanLiu1993/GISRUK2021_GCN_spatial_autocorrelation

In its simplest form, the A_{ij} will be 1 if two locations are adjacent and 0 if they are not, where the whole area is represented as nodes and their connections. Therefore, such adjacency can be seen as a graph-based generalization of a spatial matrix.

3 Case Study

3.1 Regular Grid: Hexagons

The first step to validate the proposed approach is to test it on data of known properties. Therefore, our first experiment aimed to test whether our approach can correctly identify values that have been artificially generated to be spatially autocorrelated. We tested our model on a dataset that consists of 2700 hexagons. The use of hexagons has a long-standing history in GIScience for investigating spatial patterns at various geographical scales. Compared with irregular geometry shapes (e.g., output areas, postal areas), hexagons are the polygons closest to a circular shaped polygon that can be tessellated as an evenly-spaced grid. Furthermore, their shape reduces the sampling bias that can emerge from edge effects. Thus, hexagons can be a useful tessellation for the preliminary testings for our proposed model.

Using the hexagonal grid as a base, we followed the method introduced by Goodchild (1986) to generate values (one per hexagon) displaying local spatial autocorrelation, as shown in Figure 2(a). We used standardized local G value (z-score) and p significance value from the standard Getis-Ord Gi* statistic (Gi*) (Getis and Ord, 1992) for each output area to identify and label hot spot areas ($z>0$, $p<0.01$), cold spot areas ($z<0$, $p<0.01$) and non-significant areas ($p>=0.01$).

Having a set of spatially autocorrelated data and a set of labels defining whether each hexagon is part of a hot or cold spot (or neither), this first experiment will assess whether a GCN trained on a subset of those data can learn the relationship between the values of each hexagon and its neighbours, and the labels assigned through the Gi* analysis.

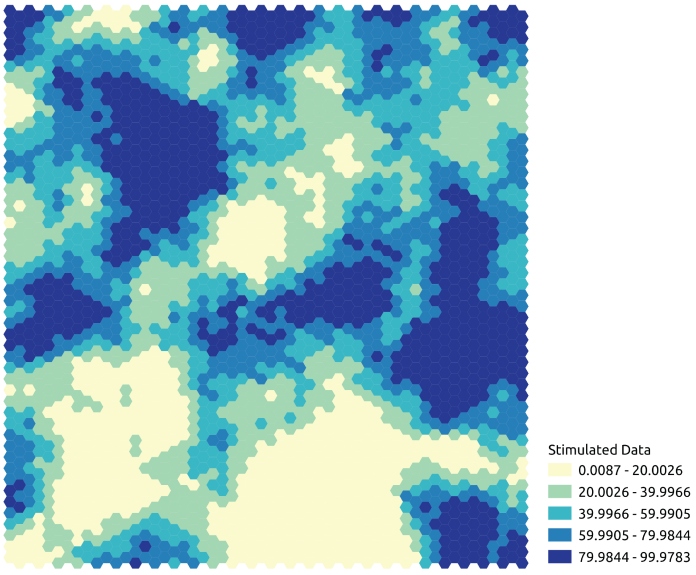
The code is developed in Keras[2] with Tensorflow[3] as the backend. We randomly selected 500 hexagons as training samples for the two-layer GCN's hyperparameter optimisation. We choose a dropout rate of 0.2 for all layers, $L2$ regularisation factor for the first GCN layer and 16 as the number of hidden units. We train the GCN model for a maximum of 3000 epochs (training iterations) using Adam (Kingma and Ba, 2014) with a learning rate of 0.01 on a cross-entropy loss function, and early stopping with a window size of 300, that is the model stop training if the validation loss does not decrease for 300 consecutive epochs. Trainable weights initialisation and feature vectors normalisation remain the same as in Kipf and Welling (2016). As shown in Table 1, GCN achieves a high accuracy (92.77%) in node classification task using partially labelled data. Figure 2(b) demonstrates that the labels generated by the GCN model mostly match the labels assigned during the Gi* analysis. The result shows a significant association existed between the GCN outputs and labels assigned based on Gi* statistics ($X^2(4)=3252.2$, $p<0.01$). This experiment indicates our proposed model has the potential to be implemented on large size datasets for the spatial autocorrelation analysis.

[2]https://keras.io/

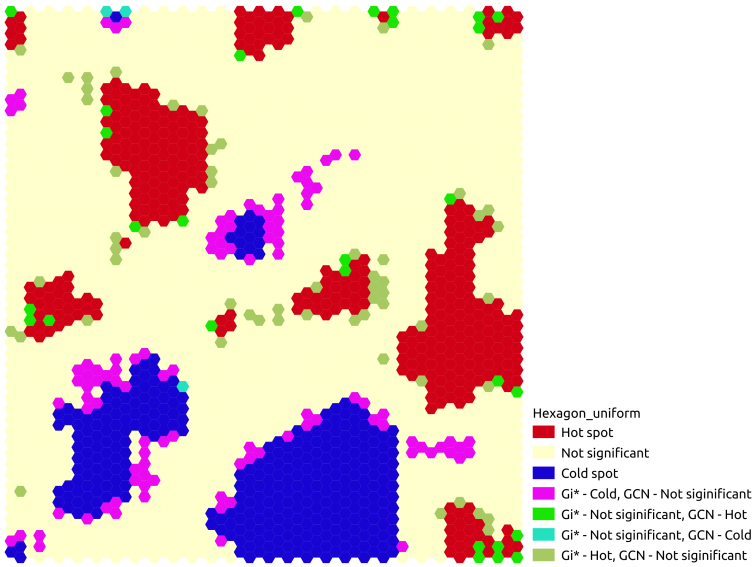
[3]https://www.tensorflow.org/

Data	Accuracy
Hexagons: Simulated Data	92.77%

Table 1: Results of the GCN accuracy.



(a) Simulated data

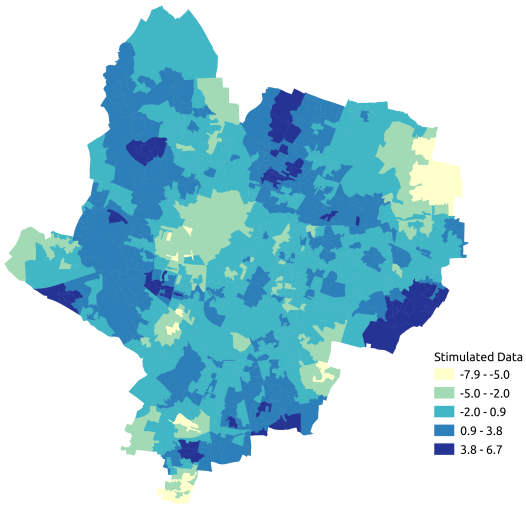


(b) Comparisons between classification results and pre-assigned labels

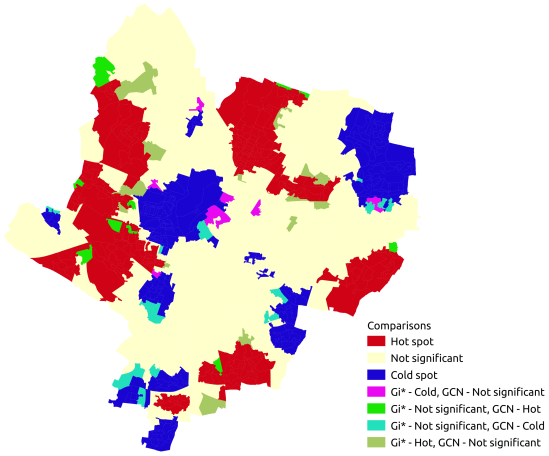
Figure 2: Simulated dataset for hexagons and the results produced by GCN.

3.2 Leicester: Artificial Dataset

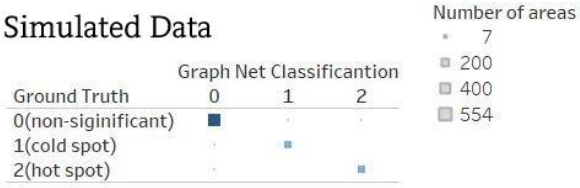
Having tested our model on regular grids, we further tested our model on an irregular geometry using output areas. Using the geometry of the city of Leicester (969 output areas) as a base, we adopted the similar approach as described in the previous section to artificially generate attributes displaying local spatial autocorrelation, as shown in Figure 3(a). We subdivided Leicester's city into six broad areas to ensure internal variation, aggregating neighbouring 2011 Census Middle-Super Output Areas (MSOAs). We generated a set of spatially autocorrelated values (one per broad area) and used that value as a starting point to generate spatially autocorrelated values for each OA within each broad area separately. We used the standard Gi* statistic to identify the hot and cold spots created by the artificial, generative process just described, defining neighbour OA based on the 12 nearest neighbours. Same to the previous experiment, we used the standardized local G value (z-score) and p significance value for each output area to identify and label hot spot areas ($z>0$, $p<0.01$), cold spot areas ($z<0$, $p<0.01$) and non-significant areas ($p>=0.01$).



(a) Simulated data in the city of Leicester



(b) Comparisons between classification results and pre-assigned labels



(c) Visual Chi-Square statistical tests

Figure 3: Simulated dataset for Leicester and the results produced by GCN. Map boundaries source: Office for National Statistics licensed under the Open Government Licence v.3.0. Contains OS data © Crown copyright and database right 2021.

Data	Accuracy
Leicester: Simulated Data	96.25%

Table 2: Results of the GCN accuracy.

We selected one of the six broad areas containing 204 OAs (roughly, the west part of Leicester) as training samples for the two-layer GCN's hyperparameter optimisation. The rest of 795 OAs are used as the validation set and not be used during the training. The classifications' accuracy is summarised in Table 2, where GCN still achieves good performance (96.25% accuracy) in node classification task using partially labelled data on the irregular geometry.

This experiment's outcome is illustrated in Figure 3(b), which shows how the labels generated through the GCN model for the 795 OAs in the validation set largely match the labels assigned during the Gi* analysis. Figure 3(c) can be interpreted as visual representations of the Chi-Square statistical tests, showing the correspondence between the classification produced by the GCN and during the Gi* analysis. The test clearly shows that there is a significant association between the GCN outputs and labels assigned based on Gi* statistics ($X^2(4)$=1647.2, p<0.01). This experiment's results provide a first clear indication that GCN is capable of learning to identify local patterns of spatial autocorrelation through message exchanging between nodes in the geographic neighbours' network.

4 Discussions

We adopt a semi-supervised GCN to model the spatial statistics for areas where its nature limits traditional Gi* statistics. The p-value of significance in traditional spatial statistics is simulated through multiple comparisons, and it can lead to uncertainties for large datasets. Therefore, we introduce a generalised model that can capture geographical data's spatial patterns using graph convolutional networks. Our preliminary analysis demonstrates that a GCN can capture the correspondence between the labels assigned through a classic Gi* analysis and the spatial autocorrelation of values by processing the data features and the spatial information separately through the graph network. In our future research, we will test the framework on larger (e.g., national) scales to test the scalability and robustness of the GCN-based method.

5 Biography

Pengyuan Liu. PhD in the School of Geography, Geology and Environment, University of Leicester. His research interests broadly cover urban analytic, digital geographies, GIScience, spatio-temporal modelling and GeoAI.

Stefano De Sabbata. Lecturer in Quantitative Geography, University of Leicester. His research focuses on geographic information science, critical GIS, and quantitative human geography.

References

Anselin, Luc (1995). "Local indicators of spatial association—LISA". In: *Geographical analysis* 27.2, pp. 93–115.

Getis, Arthur (2008). "A history of the concept of spatial autocorrelation: a geographer's perspective". In: *Geographical Analysis* 40.3, pp. 297–309.

Getis, Arthur and J Keith Ord (1992). "The analysis of spatial association by use of distance statistics". In: *Geographical analysis* 24.3, pp. 189–206.

Goodchild, Michael F (1986). *Spatial autocorrelation*. Vol. 47. Geo Books.

Haining, R.P. (2001). "Spatial Autocorrelation". In: *International Encyclopedia of the Social Behavioral Sciences*. Ed. by Neil J. Smelser and Paul B. Baltes. Oxford: Pergamon, pp. 14763–14768. ISBN: 978-0-08-043076-8. DOI: https://doi.org/10.1016/B0-08-043076-7/02511-0. URL: https://www.sciencedirect.com/science/article/pii/B0080430767025110.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kipf, Thomas N and Max Welling (2016). "Semi-supervised classification with graph convolutional networks". In: *arXiv preprint arXiv:1609.02907*.

Lin, Mingfeng, Henry C Lucas Jr, and Galit Shmueli (2013). "Research commentary—too big to fail: large samples and the p-value problem". In: *Information Systems Research* 24.4, pp. 906–917.

Ord, J Keith and Arthur Getis (1995). "Local spatial autocorrelation statistics: distributional issues and an application". In: *Geographical analysis* 27.4, pp. 286–306.

O'Sullivan, David and David Unwin (2014). *Geographic information analysis*. John Wiley & Sons.