Knowledge and Information Systems (2022) 64:919–941
https://doi.org/10.1007/s10115-021-01640-x

REGULAR PAPER

Fast spatial autocorrelation

Anar Amgalan[1] · LR Mujica-Parodi[2] · Steven S. Skiena[3]

Received: 28 January 2021 / Revised: 3 December 2021 / Accepted: 3 December 2021 /
Published online: 20 March 2022

Abstract

Physical or geographic location proves to be an important feature in many data science models, because many diverse natural and social phenomenon have a spatial component. *Spatial autocorrelation* measures the extent to which locally adjacent observations of the same phenomenon are correlated. Although statistics like Moran's I and Geary's C are widely used to measure spatial autocorrelation, they are slow: All popular methods run in $\Omega(n^2)$ time, rendering them unusable for large datasets, or long time-courses with moderate numbers of points. We propose a new S_A statistic based on the notion that the variance observed when merging pairs of nearby clusters should increase slowly for spatially autocorrelated variables. We give a linear-time algorithm to calculate S_A for a variable with an input agglomeration order (available at https://github.com/aamgalan/spatial_autocorrelation). For a typical dataset of $n \approx 63,000$ points, our S_A autocorrelation measure can be computed in 1 second, versus 2 hours or more for Moran's I and Geary's C. Through simulation studies, we demonstrate that S_A identifies spatial correlations in variables generated with spatially-dependent model half an order of magnitude earlier than either Moran's I or Geary's C. Finally, we prove several theoretical properties of S_A: namely that it behaves as a true correlation statistic and is invariant under addition or multiplication by a constant.

Keywords Algorithm design and analysis · Computational efficiency · Autocorrelation · Biomedical informatics · Magnetic resonance · Clustering algorithms

✉ Anar Amgalan
anar.amgalan@stonybrook.edu

LR Mujica-Parodi
lilianne.strey@stonybrook.edu

Steven S. Skiena
skiena@cs.stonybrook.edu

[1] Department of Physics and Astronomy, Stony Brook University, New York, USA

[2] Department of Biomedical Engineering, Stony Brook University, New York, USA

[3] Department of Computer Science, Stony Brook University, New York, USA

1 Introduction

Physical or geographic location proves to be an important feature in many data science models, because many diverse natural and social phenomenon have a spatial component. Geographic features such as longitude/latitude, zip codes, and area codes are often used in predictive models to capture spatial associations underlying properties of interest. Some of this is for physical reasons: The current temperature at location p_1 is likely to be similar to that at p_2 if p_1 is near p_2, and the amount of synchrony between two regions in the brain is a function of the network of physical connections between them. But social and economic preferences in what people like, buy, and do also have a strong spatial component, due to cultural self-organization (homophily) as well as differential access to opportunities and resources.

Correlation measures (including the Pearson and Spearman correlation coefficients) are widely used to measure the degree of association between pairs of variables X and Y. By convention, $corr(X, Y) = 0$ signifies that X and Y are independent of each other , values $0 < corr(X, Y) \leq 1$ denote positive dependence on each other and $-1 \leq corr(X, Y) < 0$ signify inverse dependencies. The strength of dependency, and our ability to predict X given Y, increases with $|corr(X, Y)|$. *Autocorrelation* of time series or sequential data measures the degree of association of z_i and sequence elements with a lag-l, i.e., z_{i+l}. *Spatial autocorrelation* measures the extent to which locally adjacent observations of the same phenomenon are correlated.

Spatial autocorrelation proves more complex to measure than sequence autocorrelation, because the association is multi-dimensional and bidirectional. Social scientists and geoscience researchers have developed a rich array of statistics, which endeavor to measure the spatial correlation of a variable Z, including Moran's I [1], Geary's C [2], and the Matheron variogram [3]. For example, political preferences are generally spatially autocorrelated, as reflected by the notion of "Red" states and "Blue" states in the USA. There is a general sense that political preferences are increasingly spatially concentrated. Spatial autocorrelation statistics provide the right tool to measure the degree to which this and related phenomena may be happening.

These statistics are widely used, particularly Moran's I and Geary‘s C, yet our experience with them has proven disappointing. First, they are slow: All popular methods run in $\Omega(n^2)$ time, rendering them unusable for large datasets, or long time-courses with moderate numbers of points. Second, although they are effective at distinguishing spatial correlated variables from uncorrelated variables from relatively few samples, they appear less satisfying in comparing the degree of spatial association among sets of variables. Other inroads to efficient spatial data analysis primarily concern with detection of outliers and anomalies [4,5]. In this paper, continuing the naming tradition of Moran's I and Geary's C, we humbly propose a new spatial autocorrelation statistic: Skiena's A or S_A. We will primarily consider a dataset of 47 demographic and geospatial variables, measured over roughly 3,000 counties in the USA [6–10], with results reported in Table 1. The dataset was previously used in identification of socio-demographic variables determining county-level substance abuse statistics in the USA [11]. With our preferred statistic, the median-clustered S_A, the six geophysical variables measuring sunlight, temperature, precipitation, and elevation all scored as spatially autocorrelated above 0.928, whereas the strongest demographic correlation (other language) came in at 0.777, reflecting the concentration of Hispanic-Americans in the Southwestern USA.

Table 1 Spatial autocorrelation for 47 geophysical and demographic variables on US counties, sorted by their median-clustered S_A value. We note that the median-linkage agglomeration order produced the most satisfying ranking of variables by spatial autocorrelation compared to classical statistics and the weaker single-linkage aggregation order. Median-clustered S_A ranks all geophysical variables as more spatially autocorrelated than any demographic variable, and exhibits a stronger correlation with Geary's C (-0.943) and comparable with Moran's I (0.879) than they do with themselves (-0.922). For both S_A metrics, the agglomeration order was computed only once and reused for all variables

		Geary	Moran	S_A	
Variable	N	C	I	Single	Median
Maxtemp	3106	0.678	0.272	0.540	0.966
Sunlight	3106	0.684	0.258	0.519	0.965
Mintemp	3106	0.674	0.273	0.555	0.962
Precipitation	3106	0.651	0.273	0.722	0.942
Max heat index	3106	0.688	0.268	0.550	0.930
Elev	3142	0.662	0.250	0.802	0.928
Other language	3142	0.778	0.130	0.598	0.777
Med house val	3141	0.752	0.161	0.434	0.772
Log med house val	3141	0.782	0.170	0.378	0.749
Log pop density	3141	0.758	0.190	0.438	0.688
Main protestant	3113	0.820	0.162	0.395	0.675
Percent black	3142	0.835	0.194	0.049	0.672
Rep sen 2010	2115	0.804	0.161	0.406	0.668
Foreign born	3142	0.811	0.102	0.401	0.654
Percent white	3142	0.853	0.131	0.155	0.635
Evan protestant	3122	0.838	0.169	0.231	0.625
Percent physically inactive chr	3137	0.826	0.152	0.325	0.616
Rep pre 2012	3128	0.843	0.107	0.332	0.608
Catholic	2958	0.861	0.111	0.330	0.605
Total pop	3142	0.846	0.032	0.552	0.602
Percent obese chr	3137	0.837	0.131	0.350	0.588
High school	3142	0.851	0.143	0.213	0.582
Rep pre 2008	3112	0.864	0.114	0.281	0.566
Year potential life lost rate chr	2861	0.856	0.131	0.222	0.558
Percent excessive drinking chr	2591	0.860	0.145	0.252	0.539
Log med house income	3141	0.895	0.104	0.089	0.531
Percent fair or poor chr	2738	0.874	0.150	0.126	0.525
Med house income	3141	0.888	0.104	0.073	0.515
Rep hou 2010	3091	0.914	0.075	0.212	0.512
Separated	3142	0.847	0.154	0.130	0.500
Motorvehicle mortality rate chr	2828	0.888	0.086	0.281	0.493
Below poverty	3141	0.904	0.099	0.117	0.464
Percent smokers chr	2502	0.889	0.103	0.219	0.414

Table 1 continued

		Geary	Moran	S_A	
Variable	N	C	I	Single	Median
Divorced	3142	0.907	0.071	0.237	0.391
Physically unhealthy days chr	2954	0.917	0.116	0.044	0.388
Med age	3142	0.930	0.031	0.287	0.375
Bac	3142	0.918	0.075	0.074	0.357
Mentally unhealthy days chr	2953	0.924	0.096	0.100	0.321
Grad	3142	0.910	0.078	0.043	0.319
Married	3142	0.908	0.086	0.118	0.307
Agasltrate	2056	0.967	0.058	−0.021	0.179
ls 10 avg	2004	0.945	0.033	0.069	0.165
Percent male	3142	0.937	0.030	0.134	0.164
Same sex	3142	0.958	0.026	0.087	0.120
Pop density	3141	0.906	0.073	−0.598	0.116
Robberyrate	2056	0.971	0.052	−0.098	0.085
Murderrate	2056	1.029	0.019	−0.172	-0.092

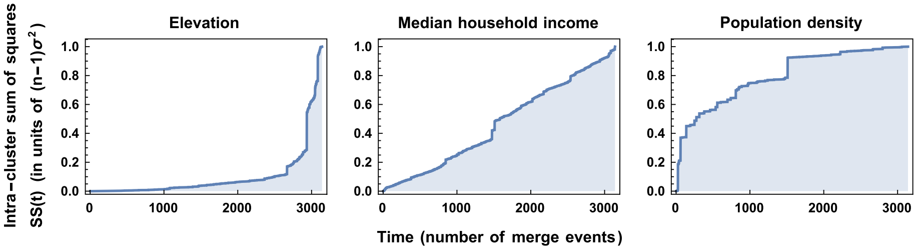


Fig. 1 Representative traces of within-cluster sum of squared deviations $SS(t)$ as a function of the number of merging events, for selected US county variables: elevation, household income and population density. Intuitively, in left panel, the initial slower increase indicates that counties being clustered together early-on have similar elevation values. In right panel, the variability in population densities of counties are shown to quickly take off as nearby counties are clustered

Our statistic is based on the notion that spatially autocorrelated variables should exhibit low variance within natural clusters of points. In particular, we expect the variance observed when merging pairs of nearby clusters to increase less the more spatially autocorrelated the variable is. The within-cluster sum of squares of single points is zero, while the sum of squares of a single cluster after complete agglomerative clustering is $(n-1)\sigma^2$. The shape of this trajectory from 0 to $(n-1)\sigma^2$ during the $n-1$ merging operations defines the degree of spatial autocorrelation, as shown in Fig. 1. The linearly-transformed (to enforce a range on the value of the final statistic) version of the trajectory shown in Fig. 2 then becomes the basis for the S_A statistic.

Our major contributions in this paper include:

- *Linear-time spatial correlation*—The complexity to calculate S_A for a variable defined by n points and an input agglomeration order is $O(n)$, where traditional measures such as Moran's I and Geary's C require quadratic time. This matters: for a typical dataset of

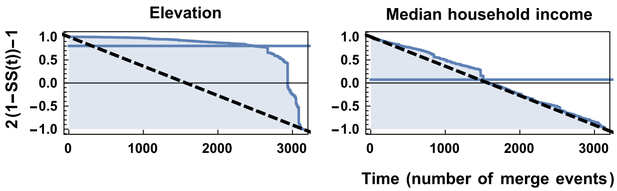


Fig. 2 Representative traces of the single-linkage S_A statistic (sum of squared deviations $SS(t)$ scaled with $L(x) = 2(1 - x) - 1$ to be in range $[-1, 1]$) as a function of the number of merging events, for selected US county variables. The area under the curve shows *Elevation* as strongly spatially correlated (S_A = **0.802**), *Median Income* as uncorrelated (S_A = **0.073**), and *Population Density* as spatially anti-correlated (S_A = **−0.598**)

Statistic	Number of data points				
	100	1000	10000	39810	63095
Moran I	≤ 1	≤ 1	60	1036	6784
Geary C	≤ 1	2	169	3112	11901
S_A single	≤ 1	≤ 1	≤ 1	≤ 1	≤ 1
S_A median	≤ 1	≤ 1	≤ 1	≤ 1	≤ 1

$n \approx 63,000$ points, our S_A autocorrelation measure can be computed in 1 second, vs. 2 hours for Moran's I and Geary's C. Times shown are in seconds.

For points in two dimensions, the single-linkage agglomeration order can be computed in $O(n \log n)$. Constructing more robust agglomeration orders like median-linkage may take quadratic time; however, this computation needs to be performed only once when performing spatial analysis over m distinct variables or time points.

We demonstrate the practical advantages of this win in an application on a brain fMRI time series data—analyzing the results of a dataset roughly 36,000 times faster than possible with either Moran's I or Geary's C, had they not run out of memory in the process.

- *Greater sensitivity than previous methods*—We assert that the median-clustered S_A captures spatial correlations at least as accurately as previous statistics. Through simulation studies, we demonstrate that it identifies spatial correlations in variables generated with spatially dependent model half an order of magnitude earlier than either Moran's I or Geary's C (Fig. 11). On the US county data, we show that median-clustered S_A correlates more strongly with Geary's C (−0.943) and comparably with Moran's I (0.879) than they do with themselves (−0.922).
- *Theoretical analysis of statistical properties*—We demonstrate a variety of theoretical properties concerning S_A. We prove that it behaves as a true correlation statistic, ranging from $[-1, 1)$ with an expected value of 0 for any *i.i.d.* random variable generated independently of coordinates. We show that $S_A(X) = S_A(a + X) = S_A(a \cdot X)$, meaning it is invariant under addition or multiplication by a constant applied to the variable z. Further, we show that S_A measures increased spatial correlation as the sampling density increases, as should be the case for samples drawn from smooth functions—but is not true for either Moran's I or Geary's C.

The implementation of our statistic is available at https://github.com/aamgalan/spatial_autocorrelation. This paper is organized as follows. Section 2 introduces previous work on spatial autocorrelation statistics, with descriptions of six such statistics including the popular

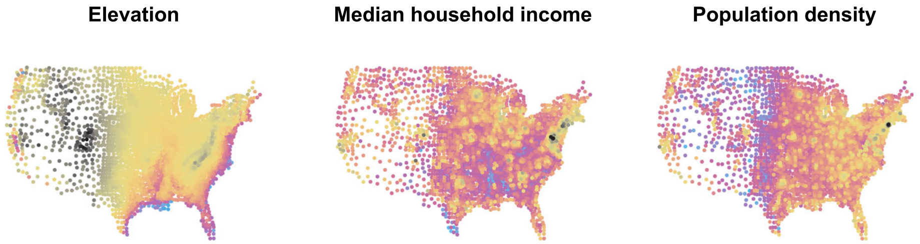


Fig. 3 The spatial distributions of the demographic variables (US counties dataset)

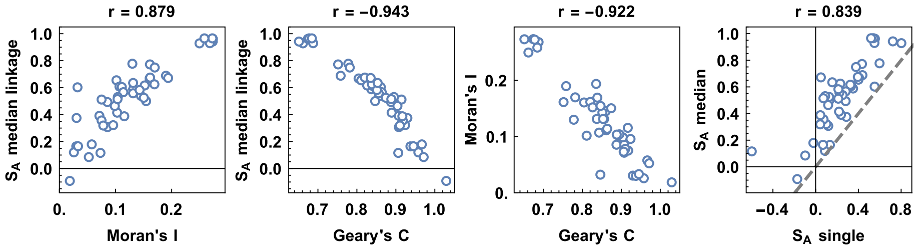


Fig. 4 Scatter plot of variables Moran's I, Geary's C, and S_A. US counties dataset

Moran's I and Geary's C. Our new S_A agglomerative clustering statistic, with a fast algorithm to compute it, is presented in Sect. 3. Theoretical and experimental results are presented in Sects. 4 and 5, respectively.

2 Previous work

2.1 Moran's *I*

The most well known of spatial autocorrelation metrics, Moran's I [1], has been around for more than 50 years. Originally proposed as a way of capturing the degree of spatial correlation between neighboring elements on a two-dimensional grid data from agricultural research, it calculates the following in its current form:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(z_i - \overline{z})(z_j - \overline{z})}{\sum_i (z_i - \overline{z})^2}$$

where z_i is the value of random variable z at each of the N spatial locations, w_{ij} is the weight between spatial locations i and j, with $W = \sum_{i,j} w_{ij}$ and $\overline{z} = \sum_i z_i / N$. Moran's I provides a global measure of whether the signed fluctuations away from the mean of quantity of interest z at a pair of spatial locations correlates with the weight (frequently the inverse distance is used) between the locations. The metric found extensive use in fields that concern mapped data: econometrics [12], ecology [13], health sciences [14], geology, and geography [15]. Statistical distributions and their moments for Moran's I under various conditions have been derived [16–18].

2.2 Geary's *C*

Another early contender in the field is the Geary's C, originally named the *contiguity ratio* [2]. First demonstrated as a viable metric of spatial correlation on the example of demographic and agricultural data from counties of Ireland, it is defined as

$$C = \frac{N-1}{2W} \frac{\sum_i \sum_j w_{ij}(z_i - z_j)^2}{\sum_i (z_i - \overline{z})^2}$$

Moran's I and Geary's C have several features in common: Both take the form of an outer product weighted by the spatial weights between the locations and both are normalized by the observed variance of z and the sum of all spatial weights. The distinction between them is the exact outer product operations carried out: Moran's I multiplies the signed fluctuations away from the mean of z: $(z_i - \overline{z})(z_j - \overline{z})$, whereas Geary's C takes the square of difference between values of z at spatial locations i and j: $(z_i - z_j)^2$. As such, Geary's C takes on a large value for a variable that displays large variation among closely neighboring (large weight w_{ij}) spatial locations, whereas Moran's I is large when the neighboring values fluctuate from the mean in the same direction.

2.3 Matheron's Variogram and γ

Another metric is the variogram method of Matheron [3] intended to quantify the typical variation of spatial data points as a function of the distance separating them. Empirical variogram is often utilized in practice and is defined as follows:

$$\hat{\gamma}(h \pm \delta) = \frac{1}{|N(h \pm \delta)|} \sum_{(i,j) \in N(h \pm \delta)} |z_i - z_j|^2$$

where h is the distance between spatial locations with allowed tolerance δ, $N(h \pm \delta)$ is the set of all pairs of points (i, j) such that distance between them lies in range $h \pm \delta$, and z_i and z_j are the values of the variable of interest at locations indexed i and j, respectively. Variogram analysis results in intuitive quantities: *sill* and *range* extracted from the curve of $\hat{\gamma}(h \pm \delta)$, where *sill* indicates the eventual level of variability reached at asymptotic length scales, and *range* denotes the length scale required to reach variability indistinguishable from the eventual *sill*. Variogram is extensively used in geology as part of *kriging* in mineral surveillance process [19].

2.4 Γ index and local Γ index

The global Γ index, proposed in 1967, as a generalized method for identifying time-space clustering of cancer cases and other geographically labeled incidence data, tries to capture not only spatial, but also temporal information, albeit on binary variables indicating whether an incidence occurred or not [10]. Γ index considers two matrices: a_{ij} and b_{ij}, one containing the measure of spatial similarity between incidences and the other containing the temporal similarity information. The statistic then is

$$\Gamma = \sum_{i<j} a_{ij} b_{ij}.$$

The local version of Γ statistics avoids summation over index i, making it a quantity specific to the observation: $\Gamma_i = \sum_j a_{ij} b_{ij}$, and when summed equal to the global Γ, a property discussed in 2.6.

2.5 Getis-Ord G_i^*

The 1990s saw a rapid development in the field and what some termed the global-to-local transition [20]. Known as the G and G_i statistics, a class of metrics first formalized by Getis and Ord [21] appeared as circumventing the shortcomings of the Moran's I statistics. The local version of the statistic, G_i, is defined as follows:

$$G_i(d) = \frac{\sum_{j \neq i} w_{ij}(d) z_j}{\sum_{j \neq i} z_j}$$

where d is the length scale of concentration of variable z being tested, and w_{ij} is a binary weight indicating whether spatial locations i and j are within distance d of each other. A minute modification of including the ith element in the summation over index j turns the metric into $G_i^*(d)$. The global version of Getis-Ord statistics, measuring the overall level of concentration of variable z in the neighborhood of linear scale d, is defined as follows:

$$G(d) = \frac{\sum_i \sum_j w_{ij}(d) z_i z_j}{\sum_i \sum_j z_i z_j}$$

The global $G(d)$ differs from Moran's and Geary's measures by taking the cross-product by multiplying the variables at locations i and j together: $z_i z_j$, instead of $(z_i - \overline{z})(z_j - \overline{z})$ in the case of Moran's and $(z_i - z_j)^2$ in the case of Geary's. For the G_i^* statistic, a negative number reveals proximity of low values of the variable, and a positive number—proximity of high values. The first empirical use case of G statistics was to rule out significant spatial correlation on the county-level data of sudden infant death syndrome for US state of North Carolina and to reveal concentration of low home values in San Diego County beyond what Moran's I would have indicated [21].

2.6 Anselin's LISA and local Moran and Geary

Anselin proposed a generalized procedure for localizing the contribution of individual measurements on the global measure of spatial autocorrelation termed local indicators of spatial association (LISA). The method also serves to identify hot spots or pockets of local variation in the mapped variable. LISA, broadly defined using two requirements: (i) The statistic for a specific measurement should report whether similar values are clustered around it and (ii) the sum over all measurements should be proportional to a global statistic of spatial autocorrelation, generalizes the localized Moran's I_i and Geary's c_i statistics, also defined by Anselin [22]:

$$I_i = (z_i - \overline{z}) \sum_j w_{ij}(z_j - \overline{z}) \quad \text{and} \quad c_i = \sum_j w_{ij}(z_i - z_j)^2$$

Both local statistics are, in fact, proportional to their global counterparts with straightforward proportionality constants, when summed up over all spatial locations. LISA's (specifically local Moran's I_i) first demonstrated usage was on dataset of international conflict among African nations, quantitatively identifying the hotbed of instability in Northeastern Africa.

In the same category of techniques is the Moran's scatter plot, also outlined by Anselin [23], which disassociates low spatial autocorrelation into *quadrants* of low values surrounded by high values and high values surrounded by low values, as well as high value of spatial autocorrelation into *quadrants* of low values among other low values and high values among other high values. See Getis [20] for a thorough history of spatial autocorrelation analysis.

3 The S_A Algorithm and statistic

Our proposed method, which we term S_A, produces a measure of spatial autocorrelation given a particular agglomeration order of n locations $\{\hat{x}_i\}$ embedded in Euclidean space and values of random variable $\{z_i\}$ (with variance σ^2) paired with them. S_A is agnostic to the exact clustering used, provided it is agglomerative and two clusters of spatial locations are merged at each step.

S_A exploits the fact that the total sum of squared deviations ($SS(t)$) from the cluster mean of the variable z_i increases monotonically as clusters are joined (proof in Sect. 4.1). This quantity is traced at a cost of constant time per merge event, starting when the first pair of observations are joined into a cluster and reaching $(n-1)\sigma^2$ when all observations are in a single cluster. We are interested in how quickly during the agglomeration process this trace of sum of within-cluster squares takes off and reaches its eventual value of $(n-1)\sigma^2$.

Formally, computation of S_A starts with all coordinates as their own singleton clusters and keeps track of the geographic centroids of clusters ($\bar{\hat{x}}_{C_1}$ and $\bar{\hat{x}}_{C_2}$), their sizes ($|C_1|$ and $|C_2|$), means ($\bar{z}_{C_1}$ and $\bar{z}_{C_2}$), and the total sum of squares over all clusters: $SS(t) = \sum_{C_k \in C(t)} \sum_{i \in C_k} \left(z_i - \bar{z}_{C_k}\right)^2$ where $C(t)$ denotes the set of all clusters at time t of the agglomeration order. During a merge event, clusters C_1 and C_2 are joined into a new cluster C_{12} ($C_{12} \leftarrow C_1 \cup C_2$), with size $|C_{12}| \leftarrow |C_1| + |C_2|$, coordinate centroid

$$\bar{\hat{x}}_{C_{12}} \leftarrow (|C_1|\bar{\hat{x}}_{C_1} + |C_2|\bar{\hat{x}}_{C_2})/|C_{12}|$$

and mean $\bar{z}_{C_{12}} \leftarrow (|C_1|\bar{z}_{C_1} + |C_2|\bar{z}_{C_2})/|C_{12}|$. The trace of sum of squares is updated as

$$SS(t) \leftarrow SS(t-1) + |C_1|(\bar{z}_{C_{12}} - \bar{z}_{C_1})^2 + |C_2|(\bar{z}_{C_{12}} - \bar{z}_{C_2})^2$$

Sample traces are shown in Fig 1. It is then normalized by its final value, averaged over all agglomeration steps, and linearly transformed with $L(x) = 2(1-x) - 1$ to give the S_A value:

$$S_A = 2\left(1 - (\sum_{t \le n-1} SS(t))/((n-1) \cdot SS(n-1))\right) - 1$$

with $n-1$ indicating the total number of merge events. Sample traces of the linearly transformed sum of squares are shown in Fig. 2, with their corresponding geographic distributions for the variables shown in Fig. 3.

Just like conventional correlation coefficients, S_A can range in the interval from -1 to 1. It will take 0 value when there is no spatial structure, larger value when similar values of z_i are spatially nearby and negative values if neighboring values are anti-correlated. Intuitively, both nearby locations with very different values of feature z_i and distant locations with similar values will decrease S_A, while nearby locations with similar values and distant locations with differing values will contribute to the increase in S_A. We note here that each update in the total sum of within-cluster squares due to a joining event is done in constant time, making calculation of S_A for variable z_i and any particular pre-specified agglomeration order an $O(n)$ algorithm. The required pre-computation of an agglomeration order can be performed in $O(n \log n)$ time, using single-linkage clustering in the plane.

3.1 Dependence on agglomeration order

Multiple agglomerative clustering criteria are in common use, reflecting a trade-off between computational cost and robustness. In this paper, we investigate four distinct criteria and their impact on observed spatial autocorrelations:

- *Single linkage*—Here, the distance between clusters C_1 and C_2 is defined by the closest pair of points spanning them:

  $$d(C_1, C_2) = \min_{z_1 \in C_1, z_2 \in C_2} ||z_1 - z_2||$$

  This is akin to the criteria of Kruskal's algorithm for finding minimum spanning trees, and runs in $O(n \log n)$ time for the primary use case of points in the plane. The $O(n \log n)$ time is due to the disjoint set data structure with complexity bound of $O(\alpha(n))$ on merge/search operations. α is an extremely slowly increasing inverse Ackermann function and is a small constant for all practical purposes.
- *Average linkage*—Here, we compute distance between all pairs of cluster-spanning points and average them for a more robust merging criteria than single link:

  $$d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{z_1 \in C_1} \sum_{z_2 \in C_2} ||z_1 - z_2||$$

  This will tend to avoid the skinny clusters of single link, but at a greater computational cost. The straightforward implementation of average link clustering is $O(n^3)$, because each of the n merges will potentially require touching $O(n^2)$ edges to recompute the nearest remaining cluster.
- *Median linkage*—Here, we maintain the centroid of each cluster and merge the cluster pair with the closest centroids. The new merged cluster's centroid is given by the average of the centroids of the clusters being merged. This has two main advantages. First, it tends to produce clusters similar to average link, because outlier points in a cluster get overwhelmed as the cluster size (number of points) increases. Second, it is much faster to compare the centroids of the two clusters than test all $|C_1||C_2|$ point pairs in the simplest implementation.
- *Furthest linkage*—Here, the cost of merging two clusters is the farthest pair of points between them:

  $$d(C_1, C_2) = \max_{z_1 \in C_1, z_2 \in C_2} ||z_1 - z_2||$$

  This criterion works hardest to keep clusters round, by penalizing mergers with distant outlier elements. Efficient implementations of furthest linkage clustering are known to run in $O(n^2)$ time.

All linkage methods except for single linkage, produce similar results, while single linkage produces a slightly lower S_A autocorrelation. This is natural as single linkage method merges only locally and suffers from what is known as the *chaining phenomenon*. The larger linear dimensions of the single linkage clusters reach the variability of the variable z_i earlier driving the sum of squares up and the S_A down (Fig. 5).

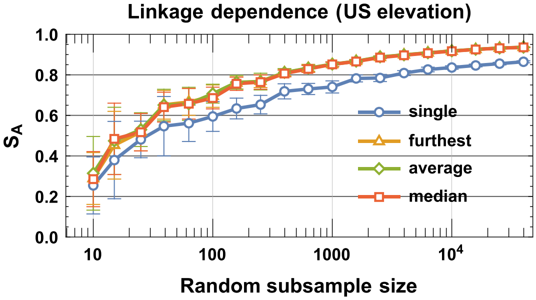


Fig. 5 S_A calculated on subsample of the elevation data with different agglomeration methods. All methods considered produce similar values of S_A, except for single linkage

3.2 Comparison with Moran's *I* and Geary's *C*

The comparison of median clustered S_A with Geary's and Moran's can be seen in the scatter plot of Fig. 4 with each point representing a feature in the US counties dataset. All three pairwise comparisons show large magnitude correlations $|r| > 0.8$. In the bottom panel, single and median linkage methods are compared for S_A.

4 Analysis of statistical properties

In this section, we prove three important properties of S_A, namely (1) monotonicity under merging, (2) that it is a well-defined correlation measure with zero corresponding to no spatial correlation, and (3) invariance under addition and multiplication by a constant.

4.1 Monotonicity

For demonstration of the monotonicity of the total sum of within-cluster squared deviations from the mean of variable z_i, it suffices to show that an arbitrary cluster C_1 merging with another (C_2) would have non-decreasing squared deviation from the new cluster's mean $\overline{z}_{C_{12}}$ compared to the original mean $\overline{z}_{C_1}$. Setting the mean shift equal to $\delta_z = \overline{z}_{C_{12}} - \overline{z}_{C_1}$, we compute the difference between the sum of squared deviations from mean for z_i values in cluster C_1 before and after the merge event as

$$\sum_{i \in C_1} (z_i - \overline{z}_{C_{12}})^2 - (z_i - \overline{z}_{C_1})^2 = \overline{z}^2_{C_{12}} - 2z_i\overline{z}_{C_{12}} - \overline{z}^2_{C_1} + 2z_i\overline{z}_{C_1}$$

Substituting the mean shift δ_z and simplifying, we obtain

$$\sum_{i \in C_1} 2\overline{z}_{C_1}\delta_z + \delta_z^2 - 2z_i\delta_z = \sum_{i \in C_1} \delta_z^2 = |C_1|\delta_z^2 \geq 0$$

where we have used the definition of mean to eliminate z_i and $\overline{z}_{C_1}$. The change in sum of squared deviations for the clusters C_1 and C_2 being merged is, therefore, nonnegative for all merge events, making the trace of $SS(t)$ a monotonic quantity. Its monotonicity, coupled with a suitable agglomeration order, which merges close-by coordinates earlier on, enables us to single out the area under its curve as a measure of spatial autocorrelation indicating how early/late in the agglomeration the variability increases from 0 to $(n-1)\sigma^2$.

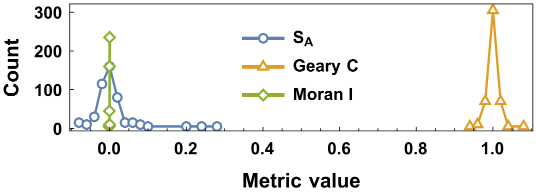


Fig. 6 Random shuffling of the county labels gives a distribution of S_A values centered on 0. Each variable from the US counties dataset was shuffled 10 times, and the statistic values were computed. The distributions for all variables were combined into a single histogram

4.2 Expected value

Intuitively, S_A is the (linearly transformed with $L(x) = 2(1-x)-1$) mean of the (monotonically increasing) sum of squared deviations of values of z_i from their cluster means while the observations are gradually merged into a single cluster made up of all coordinates $\hat{X}$. Under lack of spatial dependence, the sum of squared deviations will increase in even steps with no particular time structure and produce a mean over time equal to half its eventual value ($(n-1)\sigma^2/2$). After normalization and a linear transformation to flip the sign and adjust the range ($L(x) = 2(1-x)-1$), we will obtain 0. Empirically, we verify that the random reshuffling of the county data z_i values, while keeping the spatial locations intact, produces a narrow distribution centered at 0, similar in width to the distribution of Geary's C (see Fig. 6).

For a formal proof, let us first consider n real numbers $Z = \{z_1, \dots z_n\}$ with mean $\overline{z}$ and Euclidean coordinates $\hat{X} = \{\hat{x}_1, \dots \hat{x}_n\}$. Let $A(\hat{X}) = \{e_1, \dots e_{n-1}\}$ a merge order that determines an agglomerative clustering on the symmetric weighted graph (with no self-edges) induced by a similarity metric on coordinates $\hat{X}$. Define the stages of this agglomeration at time t as $A(\hat{X}, t) = \{e_1, \dots e_t\}$ (with a shorthand $A(t)$) such that $A(\hat{X}, n-1) = A(\hat{X})$. Let $C(t)$ denote the set of disjoint clusters present at time t of agglomeration process such that $C(0) = \{\{1\}, \{2\}, \dots \{n\}\}$ and $C(n-1) = \{\{1, 2, \dots n\}\}$.

Definition 4.1 S_A. Define the S_A statistic as

$$S_A(A(\hat{X}), Z) = 2\left(1 - \frac{\sum_{t=1}^{n-1} SS(A(t), Z)}{(n-1)\sum_{i=1}^{n}(z_i - \overline{z})^2}\right) - 1$$

where $SS(A(t), Z) = \sum_{C_k \in C(t)} \sum_{i \in C_k} (z_i - \overline{z}_{C_k})^2$ (with a shorthand notation $SS(t)$) denoting the sum of within-cluster squared deviations at time t of the agglomeration given by $A(t)$.

Theorem 4.1 *Let $Z = \{z_1, z_2, \dots z_n\}$ be a set of normal i.i.d. random variables with mean 0 and variance σ^2 and $\hat{X} = \{\hat{x}_1, \hat{x}_2, \dots \hat{x}_n\}$ their coordinates in Euclidean space. Then, the random variable $S_A(A(\hat{X}), Z)$ converges to zero in limit of large n:*

$$\lim_{n \to \infty} \mathbb{E}[S_A(A(\hat{X}), Z)] = 0$$

Proof We proceed by considering the contribution of each cluster joining event on the eventual metric S_A. During a given merge event, clusters C_1 and C_2 with sizes n_1 and n_2 and means $\overline{z}_{C_1}$ and $\overline{z}_{C_2}$ join to make the cluster C_{12} with size $n_{12} = n_1 + n_2$ and mean $\overline{z}_{C_{12}}$. At the same time, the running sum of within-cluster squares changes as follows (see Sect. 4.1):

$$\begin{aligned}\delta_{SS}(t+1) &= SS_{C_{12}}(t+1) - (SS_{C_1}(t) + SS_{C_2}(t)) \\ &= n_1(\overline{z}_{C_{12}} - \overline{z}_{C_1})^2 + n_2(\overline{z}_{C_{12}} - \overline{z}_{C_2})^2\end{aligned}$$

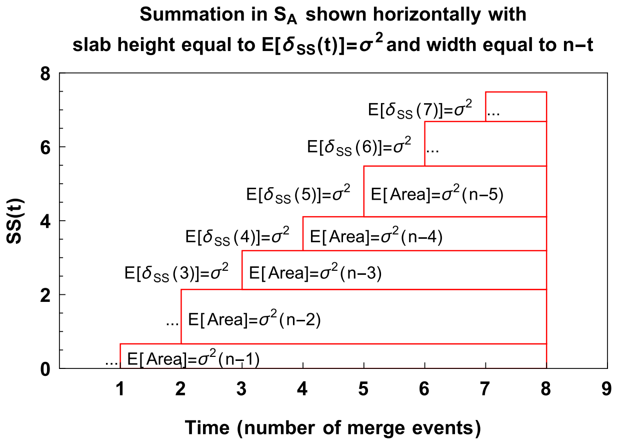


Fig. 7 Summation carried out in "horizontal slabs," each with height in expectation equal to σ^2 and deterministic width of $n - t$

The expectation of change in sum of squared deviations due to merge event $\mathbb{E}[\delta_{SS}(t+1)]$ is then given by the difference in the expectations of sum of squares before and after the merge.

$$\begin{aligned}\mathbb{E}[\delta_{SS}(t+1)] &= \mathbb{E}[SS_{C_{12}}(t+1)] - (\mathbb{E}[SS_{C_1}(t)] + \mathbb{E}[SS_{C_2}(t)])\\ &= (n_{12}-1)\sigma^2 - ((n_1-1)\sigma^2 + (n_2-1)\sigma^2)\\ &= (n_{12}-n_1-n_2+1)\sigma^2 = \sigma^2\end{aligned}$$

Here, we use the fact that for a given cluster C, SS_C—its sum of squared deviations from mean, is an estimate of the population variance biased by a factor of $n-1$. The summation in definition of S_A can then be carried out "horizontally," by considering the jump in the global sum of squares times the number of time intervals for which this jump contributes to the metric as shown in Fig. 7.

It then follows that

$$\begin{aligned}\mathbb{E}[S_A(A(\hat{X}), Z)] &= 2\left(1 - \mathbb{E}\left[\frac{\sum_{t=1}^{n-1} SS(t)}{(n-1)SS(A(\hat{X}), Z)}\right]\right) - 1\\ &= 2\left(1 - \frac{\sum_{t=1}^{n-1}\mathbb{E}\left[SS(t)\right]}{(n-1)SS(A(\hat{X}), Z)}\right) - 1\\ &= 2\left(1 - \frac{\sum_{t=1}^{n-1}(n-t)\mathbb{E}[\delta_{SS}(t)]}{(n-1)(n-1)\sigma^2}\right) - 1\\ &= 2\left(1 - \frac{((n-1)n - (n-1)n/2)\,\sigma^2)}{(n-1)(n-1)\sigma^2}\right) - 1\\ &= -\frac{1}{n-1}\end{aligned}$$

Here, we use the fact that the distribution of overall sum of squares in the denominator is related to the sampling distribution of sample variance:

$$\frac{SS(A(\hat{X}), Z)}{\sigma^2} = \frac{\sum_{i=1}^{n}(z_i - \overline{z})^2}{\sigma^2} \sim \chi^2(n-1)$$

making $SS(A(\hat{X}), Z)$ a self-averaging quantity with mean $(n-1)\sigma^2$ and variance $2(n-1)\sigma^4$, and hence vanishing relative variance in the limit of large n:

$$\lim_{n\to\infty} \frac{Var[SS(A(\hat{X}), Z)]}{\mathbb{E}[SS(A(\hat{X}), Z)]^2} = \lim_{n\to\infty} \frac{2(n-1)\sigma^4}{(n-1)^2\sigma^4} = 0$$

This lets us treat $SS(A(\hat{X}), Z)$ in denominator as a constant factor and taking the limit of large n of $\mathbb{E}[S_A(A(\hat{X}), Z)]$, we obtain

$$\lim_{n\to\infty} \mathbb{E}[S_A(A(\hat{X}), Z)] = \lim_{n\to\infty} \left(-\frac{1}{n-1}\right) = 0$$

as desired. □

4.3 Invariance

The S_A statistic has the nice property of invariance under addition and multiplication by a constant. Letting Z a spatial variable with $S_A(Z) = s$ and considering $S_A(Z+c)$ with $c \in \mathbb{R}$, we note that the sum of squared deviations is unaffected by addition of a constant, making our statistic invariant to addition of a constant c.

$$\begin{aligned} SS(T(t, \hat{X}), Z + c) &= \sum_{\substack{C_k \in C(t) \\ e_i \in C_k}} \left(z_i + c - \frac{\sum_{e_j \in C_k} z_j + c}{|C_k|} \right)^2 \\ &= \sum_{\substack{C_k \in C(t) \\ e_i \in C_k}} \left(z_i - \frac{\sum_{e_j \in C_k} z_j}{|C_k|} \right)^2 \\ &= SS(T(t, \hat{X}), Z). \end{aligned}$$

Considering multiplication of variable Z by an arbitrary constant $c \in \mathbb{R}$, we note that a factor of c^2 appears both in denominator and in numerator due to the squared deviation from the mean being considered, canceling each other and returning the same value as the original variable $S_A(c \cdot Z) = S_A(Z)$.

5 Experimental evaluation

Here, we present the results of simulations which demonstrate (1) the running time of S_A is indeed an order of magnitude faster to compute than competing statistics, (2) S_A identifies substantially weaker spatial correlations in synthetic data than Moran's and Geary's statistics, (3) S_A appears to be influenced less by non-uniform sampling than competing statistics, and finally (4) S_A appropriately reports increased autocorrelation with greater sampling density while still converging to a limit below the perfect autocorrelation of 1.

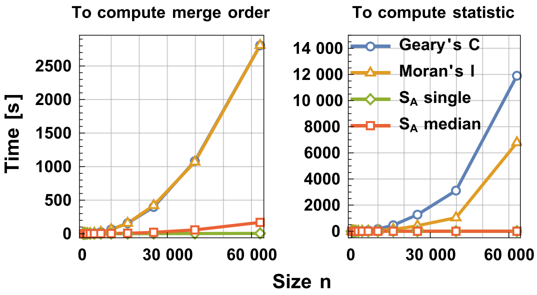


Fig. 8 Experiments concerning running time. Single-link and median-link agglomeration orders cost less to compute than single runs of Moran I and Geary's C metrics. S_A outperforms I and C drastically given the merge order on a dataset of size ≈ 63000

5.1 Running time

S_A substantially outperforms both Moran's and Geary's metrics in computation time, in establishing the agglomerative merging order both to use and to compute the statistics. In our experiments, computing a single median-linkage agglomeration order costs approximately 10% of a *single* Moran or Geary computation on the same points, as shown in Fig. 8 (left). By reusing this agglomeration order, we can save a linear factor of running time on subsequent autocorrelation analyses. Figure 8 (right) shows that for a typical dataset of $n \approx 63,000$ points, our S_A autocorrelation measure can be computed in 1 second, versus 2 hours or more for Moran's I and Geary's C.

Timing experiments were done as follows: Starting from coordinates, agglomeration order was computed using Kruskal's routine with disjoint set structure (for S_A single), scipy's linkage tool (for S_A median) and numpy's linear algebra toolbox with vectorization (for weight matrix of Moran's I and Geary's C) and metrics were computed using our streaming tool (S_A) and pysal library for Python (Moran's I and Geary's C). All tools were written in Python 3.7.

5.2 Reusing agglomeration order: fMRI time series analysis

Much of the efficiency gains of S_A accrue from its ability to reuse a once-computed agglomeration order for new data points arriving from the same spatial coordinates. We demonstrate this with an application to functional neuroimaging data (fMRI), which gives a time series readout for each spatial location in the brain. In order to study the dynamics of brain networks, neuroscience is concerned with extracting summary statistics from the brain images of potentially $> 10^6$ voxels (3D pixels) at the resolution of sampling period. Statistical tools are then used in downstream prediction and classification tasks of clinical significance. In this experiment, we used a publicly available fMRI neuroimaging dataset with 36 fMRI scans (12 human subjects $\times$ 3 experimental conditions) with each scan consisting of $2 \times 2,320 = 4,640$ repeated measurements of the entire brain at $0.8s$ sampling period [24]. We focused on the grey matter data, which consist of readings from $n = 133,000 \pm 13,000$ (mean $\pm$ std) voxels at each time point. To compute S_A, we constructed a single agglomeration order for each scan, using k-d tree structure by treating the grey matter voxels of brain as points in space to be partitioned into singletons. We cycled through the three axes of brain recursively splitting each partition between its median pair of planes perpendicular to the axis until all partitions reached size of 1. The splitting events then define an agglomeration order in reverse. The time complexity of partitioning space using k-d tree structure is $O(n)$ in the case of unbalanced

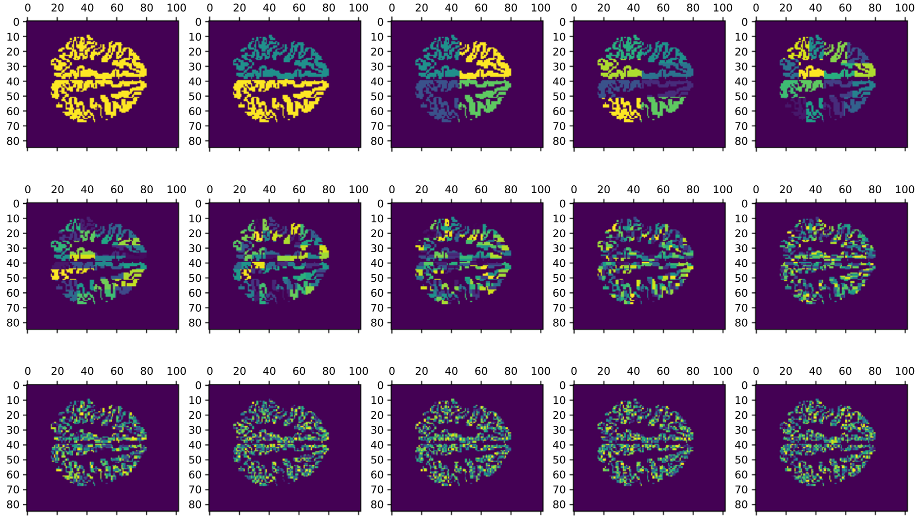


Fig. 9 The divisive clustering routine applied to MRI brain image voxels (3D pixels): We cycle through the three Cartesian axes, splitting each existing cluster along its medial wall perpendicular to the current axis. The whole brain at the top left panel eventually partitions into individual voxels. The partition steps progress left-to-right, followed by top-to-bottom

tree, and $O(n \log n)$ for a balanced tree using a median finding subroutine. See Fig. 9 for a demonstration of the individual partition steps.

Due to the highly irregular shape of the grey matter, we resorted to finding the medians for balanced partitions, with the average time to establish an agglomeration order of two minutes, but it can be reused for each of the m time points of a given scan. This reduces the run time from $O(mn^2)$ for Moran's I and Geary's C to $O(n \log n + mn)$. In our case, with $m = 4640$ time points and $n \approx 133,000$ coordinates, S_A took 3500 ± 300 seconds, or 0.75 ± 0.07 seconds per feature (time step). On the other hand, we were not able to compute Moran's I and Geary's C for $133,000$ coordinates on an average workstation hardware using the standard implementation (pysal), due to space limitations. We give a linear-time algorithm to calculate S_A for a variable with an input agglomeration order (available at https://github.com/aamgalan/spatial_autocorrelation).

Extrapolation from computations of Moran's I and Geary's C on smaller samples indicates that if memory requirements were lifted, it would take more than 7.5 and 13.5 hours, respectively, for each time step of the time series data, or roughly $36,000$ times longer than S_A. Figure 10 shows representative autocorrelation time series from brain fMRI data. This shows that S_A not only improves computation for each data feature, but also processes each additional feature in linear time by reusing the agglomeration order once it is computed. S_A's runtime for each time step is comparable to the sampling period of the fMRI data. This permits future applications in closed-loop systems that process data and provide feedback stimuli or electromagnetic stimulation to the brain in real time for improved clinical intervention.

5.3 Sensitivity to true autocorrelation: synthetic data

Ground truth on the degree of spatial autocorrelation can only be obtained from simulation results, where we explicitly generate data with specified amount of spatial autocorrelation and

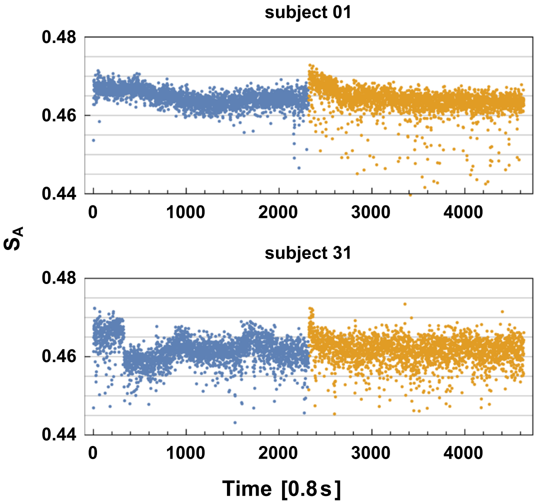


Fig. 10 Spatial autocorrelation (measured by S_A) time series for fMRI data, showing visibly different degrees of coherence on two different human subjects. We estimate that this computation would have taken roughly 36,000 times as long using either the Moran's I or Geary C statistic. The two colors indicate the two halves of the scanning session, with a short break in the middle

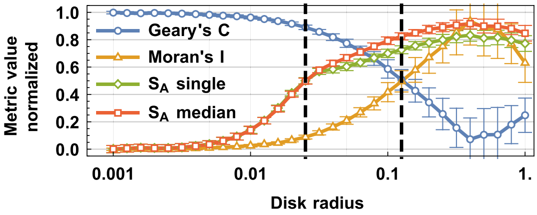


Fig. 11 S_A is more sensitive to true autocorrelation than Moran's I and Geary's C, on a "disk-averaging" generative model as a function of disk radius. Moran's I, Geary's C values are rescaled to match the range of S_A. S_A detects the autocorrelation $>$ 0.5 order of magnitude earlier. Note the entire range of [0, 0.9] is covered with S_A within 2 orders of magnitude of the disk radius. Vertical dashed lines indicate disk radii where metrics reach half of their ranges

see how much bias must be added for statistics to identify the phenomenon. For this purpose, we carry out a disk-averaging experiment, whereby a normally distributed independently sampled random variable z_i is assigned to uniformly distributed coordinates and undergoes an averaging procedure. The averaging takes all values of z_j for locations within disk of radius r around coordinate $\hat{x}_i$, and reassigns the average of the within disk values to it: $z_i \leftarrow mean(\{z_j \,|\, d(\hat{x}_i - \hat{x}_j) < r\})$. See Algorithm 1 for details of the averaging routine. The S_A statistic of the disk-averaged z_i values were computed and compared to Moran's I and Geary's C. Random sampling, disk-averaging and statistic computation were each repeated 100 times.

Figure 11 summarizes the results of these experiments for 1000 points. S_A (both single and median linkage) demonstrates far greater sensitivity, identifying significant and rapidly increasing amounts of spatial autocorrelation for disk radii half an order of magnitude smaller than that of Geary's C and Moran's I. Although both Moran and Geary statistics support problem-specific weight matrices to tune their sensitivity, the interesting autocorrelation distance scales are a priori unknown and difficult to determine, so methods without tunable parameters are preferred.

```
input  : radius of disk r
input  : number of samples m
output: Spatial data $z_r$ with spatial correlation given by radius r
for $i \leftarrow 1$ to m do
    $z[i] \leftarrow$ random sample from $\mathcal{N}(0, 1)$;
    $\hat{x}[i] \leftarrow$ random sample from Uniform([0, 1]$^2$);
end
for $i \leftarrow 1$ to m do
    $z_r[i] \leftarrow$ mean({$z[j]$ | d($\hat{x}[i]$ - $\hat{x}[j]$) < r});
end
```

Algorithm 1: Disk-averaging algorithm for inducing spatial autocorrelation on synthetic data

5.4 Sensitivity to sample size and coordinate subsampling: US elevation data

Spatial autocorrelation depends on the exact sampling of the coordinates as well as the spatial distance/weight matrix. We note that for historical and demographic reasons, US counties are not of equal size and shape, but generally smaller and more irregular in the east rather than the west. A spatial autocorrelation statistic should ideally report similar values on the same underlying geographic variable regardless of the details of the sampling method.

To interrogate whether S_A computed on subsamples of real data differs from Moran and Geary's statistics in its dependence on the exact subsample of coordinates, we use the following procedure. n random data points are drawn from the US elevation data (itself sampled at $1km^2$) [25], and S_A, Moran's I and Geary's C are computed from their coordinates $\hat{x}_i$ and elevation values z_i. Performing the experiment at sample sizes up to 40,000 points (limited by the $O(n^2)$ running time of Moran and Geary's), we compute autocorrelation metrics and compare them with the values obtained from the elevation column of the US counties dataset

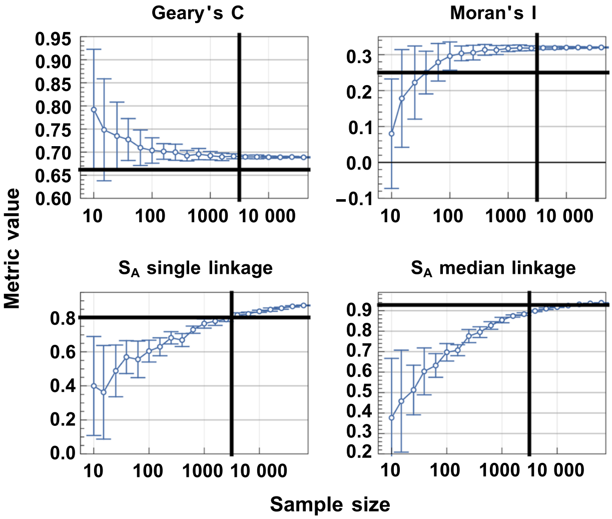


Fig. 12 S_A reveals autocorrelation independent of the exact coordinates sampled. The random subsampling experiment on $1km^2$ scale US elevation data carried out up to subsample size 40000. Vertical and horizontal lines indicate the number of counties in the US counties dataset and the value of metric computed from them, respectively

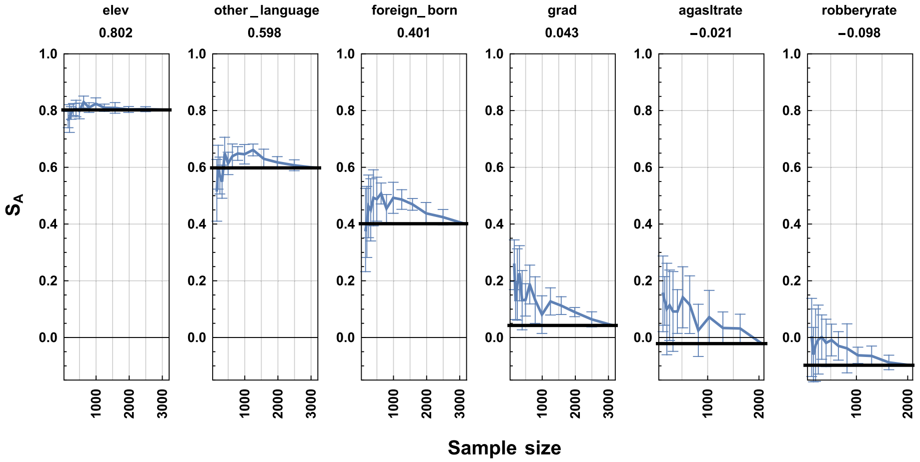


Fig. 13 The S_A estimate quickly converges to its full sample value as a function of the subsample size. The mean and standard deviation (error bar) of the S_A at the subsample size from 10 repeats are shown. S_A reveals autocorrelation independent of the exact coordinates sampled. The variables from the US counties data shown are in a descending order in S_A, but with all estimates at small subsample size starting out close to their eventual full sample values

at a sample size of 3142. The details of the routine are described in Algorithm 2, and the results are shown in Fig. 12.

```
input  : $z_{elev}$ (US elevation data at $1km^2$ resolution)
input  : m (number of random samples to take from US elevation data)
output: $S_A$ ($S_A$ value from US elevation data $z_{elev}$ at m samples)
for $i \leftarrow 1$ to m do
    choice $\leftarrow$ RandChoice($\{1, \ldots |\hat{x}_{elev}|\}$);
    $\hat{x}[i] \leftarrow \hat{x}_{elev}$[choice];
    $z[i] \leftarrow z_{elev}$[choice]
end
$S_A \leftarrow S_A(\hat{x}, z)$
```

Algorithm 2: Routine for subsampling elevation on the continental US data.

Both Moran's I and Geary's C report different values when the coordinates are sampled uniformly, compared to the irregular sample of coordinates given by US counties' locations. On the other hand, both single- and median-linkage S_A report similar values with equal number of uniformly sampled coordinates as it did with coordinates of US counties, showing robustness to changes in the exact subsampling of coordinates.

5.5 Sensitivity to sample size: US Counties Demographic data

In order to further validate the subsample stability of S_A estimates on variables other than the US elevation, we repeated the procedure on the rest of the variables from the US counties dataset. We randomly subsampled the counties and computed the S_A statistic for an indication of whether the initial small subsample estimates were close to the eventual full sample value of S_A. Our observation indicates that for a range of S_A autocorrelation, the estimates can quickly converge to its eventual value. See Fig. 13 for details.

5.6 Convergence evaluation and analytical fit

To test convergence of S_A, Moran's, and Geary's metrics, we perform the following sampling procedure on grids of random values of varying sizes. For a rectangular grid of finite size, e.g., k-by-k, we assign a uniformly random z_{ij} value to each of the k^2 grid cells, then randomly sample n real-valued coordinates from the support given by $[0, k]^2$, and take their corresponding cell's z_{ij} values to compute S_A. This procedure locks a particular correlation length into the data by choosing the number of grid cells and forces the metrics to capture it as number of sample coordinates increases. See Algorithm 3 for full description of the grid sampling routine. We expect $1/k^2$th of all samples to fall in each grid cell, thus taking on the same z value, and raising the autocorrelation as the number of samples increases to a natural limit, because there will also be nearby pairs of points that sit across a grid boundary and take different z values. Thus, a meaningful metric should converge to a large value (but less than the maximum possible 1) that decreases for shorter autocorrelation lengths induced by larger number of grid cells.

Figure 14 (left) reports that Moran's I converges to values increasingly closer to 0 as the grid size increases, indicating that it captures the de-correlated structure of large number of random grid cell entries z_{ij}. Geary's C does similarly, reporting values increasingly closer to 1. But S_A *sees* the coarser, more correlated structure of smaller grids using fewer samples, reporting *earlier* increase for 10-by-10 grid than for 100-by-100 (Fig. 14, right panel).

```
input : n (number of grid cell along each axis)
input : m (number of samples)
output: $S_A$ ($S_A$ value for grid size $n^2$ at m samples)
for $i \leftarrow 1$ to n do
    for $j \leftarrow 1$ to n do
        $z_{grid}$ $[i, j] \leftarrow$ sample from Uniform($[0, 1]^2$);
    end
end
for $i \leftarrow 1$ to m do
    $\hat{x}[i] \leftarrow$ sample from Uniform($(0, n]^2$);
    $z[i] \leftarrow z_{grid}$ [Ceiling($\hat{x}[i, 1]$), Ceiling($\hat{x}[i, 2]$)];
end
$S_A \leftarrow S_A(\hat{x}, z)$
```

Algorithm 3: Routine for randomly sampling from a grid of values.

In order to estimate the asymptotic value of the S_A metric, we fit the following log-sigmoidal functional form to the observed values of S_A as a function of samples taken: $S_A(n) = S_{max}/(1 + e^{-a(\log n - b)})$. The parameter S_{max} has a natural interpretation of the asymptotic value of S_A at unlimited number of samples turning the task of finding the asymptote into a parameter estimation for S_{max}. In Fig. 15, we report that with sample size $> 10^5$, the confidence interval for estimated S_{max} includes the eventually best estimate (dark horizontal thick line) computed using 10^7 samples. None of the estimates of S_{max} in Figs. 14 and 15 includes the value of 1.

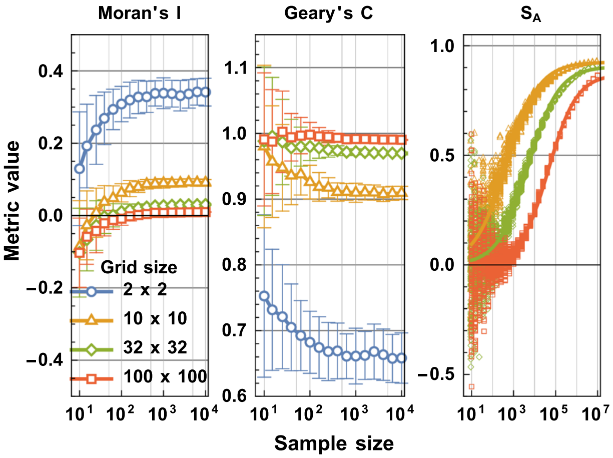
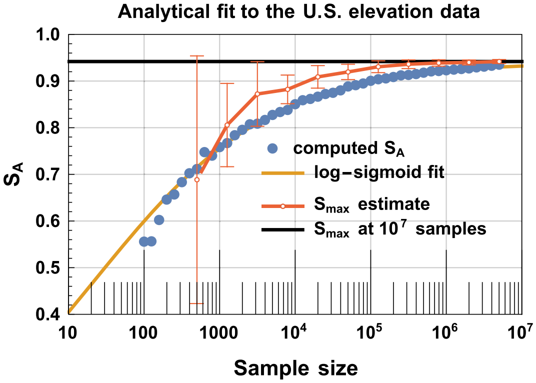


Fig. 14 Asymptotic behavior of spatial autocorrelation metrics. Random coordinates are sampled at increasing sample size from square grid of independent random values from [0, 1] interval. *Left*: Moran's I, *center*: Geary's C, and *right*: S_A, solid lines represent best fit of log-transformed sigmoid curves for data drawn from grids of size 10 x 10. Note the asymptote of single-linkage S_A converging to values < 1: $S^{10}_{max} = 0.925$, $S^{32}_{max} = 0.905$, and $S^{100}_{max} = 0.87$

Fig. 15 Estimation of S_{max}. Yellow curve: best fit to S_A as a function of sample size $S_A(N) = \frac{0.942}{1+\exp(-0.361(\log N - 3.035))}$ using data points in blue. Red curve: the confidence interval of parameter estimation for the asymptotic value S_{max} using the S_A computed only up to the sample size on the x-axis. US elevation data



6 Conclusion

The Skiena's A (S_A) algorithm and statistic we propose provide an efficient, improved sensitivity procedure for computing the spatial autocorrelation, running in linear time after computing the agglomeration order (implementation available at https://github.com/aamgalan/spatial_autocorrelation). Separating the computation into two steps: i) obtaining the agglomeration order and ii) computing of the statistic, provides additional improvements by reusing the agglomeration order for new data that arrive from the same coordinates. S_A achieves run time of $O(n \log n + mn)$ for m separate features, improving upon the standard $O(mn^2)$. As demonstrated in the fMRI example, it can be thousands of times faster in natural time series applications of spatial autocorrelation than previous methods. Even for single-shot applications in the plane where we can compute single-linkage agglomeration in $O(n \log n)$ run time, we beat previous $O(n^2)$ algorithms. We have also shown that S_A has the convenience of converging to 0 for random data, invariance under linear transforms uniformly applied to data, making it an attractive addition to standard toolbox for analysis of spatial data irrespective of the domain.

Acknowledgements The research described in this paper was partially funded by the NSF (IIS-1926751, IIS-1927227, and IIS-1546113 to S.S.S.), the W. M. Keck Foundation (L.R.M.-P.), and the White House Brain Research Through Advancing Innovative Technologies (BRAIN) Initiative (NSFNCS-FR 1926781 to L.R.M.-P.).

References

1. Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika 37(1–2):17–23. https://doi.org/10.1093/biomet/37.1-2.17
2. Geary RC (1954) The contiguity ratio and statistical mapping. The Inc Stat 5(3):115–146 (**11-01**)
3. Matheron G (1963) Principles of geostatistics. Econ Geol 58(8):1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246 (**[Online]. Available:**)
4. Lu C, Chen D, Kou Y (2003) "Algorithms for spatial outlier detection," in *Third IEEE International Conference on Data Mining*, pp. 597–600
5. Pei Sun Chawla S (2004) "On local spatial outliers," in *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pp. 209–216
6. for Disease Control C (2017) Prevention *et al.*, "Brfss prevalence & trends data," *Internet site:* http://www.cdc.gov/brfss/brfssprevalence/(Accessed July 22, 2015),
7. Bureau U (2010) "Profile of general population and housing characteristics: 2010," *Bureau USC*,
8. for Disease Control C, Prevention NC (2018)for Health Statistics *et al.*, "Multiple cause of death 1999–2016 on cdc wonder online database, released december, 2017."
9. Grieco EM, Acosta Y, De La Cruz GP (2012) *The foreign-born population in the United States: 2010*.US Department of Commerce, Economics and Statistics Administration, US
10. Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27(2):209–220 (**Part 1**)
11. Curtis B, Giorgi B, Buffone AE, Ungar LH, Ashford RD, Hemmons J, Summers D, Hamilton C, Schwartz HA (2018) Can twitter be used to predict county excessive alcohol consumption rates? PloS one 13(4):e0194290
12. Anselin L (2001) "Spatial econometrics," *A companion to theoretical econometrics*, 310330,
13. Legendre P, Fortin MJ (1989) Spatial pattern and ecological analysis. Vegetatio 80(2):107–138
14. Waller LA, Gotway CA (2004) Applied spatial statistics for public health data. John Wiley, Berlin
15. Burrough PA, McDonnell R, McDonnell RA, Lloyd CD (2015) Principles of geographical information systems. Oxford University Press, UK
16. Sen A (1976) Large sample-size distribution of statistics used in testing for spatial correlation. Geogr Anal 8(2):175–184. https://doi.org/10.1111/j.1538-4632.1976.tb01066.x (**[Online]. Available:**)
17. Kelejian HH, Prucha IR (2001) On the asymptotic distribution of the moran i test statistic with applications. J Econ 104(2):219–257
18. Getis A (1995) Cliff a d and ord j k 1973 Spatial autocorrelation london. Pion Prog Human Geogr 19(2):245–249. https://doi.org/10.1177/030913259501900205
19. Davis J (1986) Statistics and data analysis in geology. Wiley, Hoboken
20. Getis A (2008) A history of the concept of spatial autocorrelation: a geographer's perspective. Geogr Anal 40(3):297–309. https://doi.org/10.1111/j.1538-4632.2008.00727.x (**[Online]. Available:**)
21. Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. Geogr Anal 24(3):189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x (**[Online]. Available:**)
22. Anselin L (1995) Local indicators of spatial association-lisa. Geogr Anal 27(2):93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x (**[Online]. Available:**)
23. Anselin L (1996) "The moran scatterplot as an esda tool to assess lo-cal instability in spatial association," *Spatial Analytical Perspectives on GIS in Enviromental and Socio-Economic Sciences*,
24. Mujica-Parodi LR, Amgalan A, Sultan SF, Antal B, Sun X, Skiena S, Lithen A, Adra N, Ratai E-M, Weistuch C et al (2020) Diet modulates brain network stability, a biomarker for brain aging, in young adults. Proc Natl Acad Sci 117(11):6170–6177
25. Survey UG, Centre for Topographic Information (Sherbrooke) NRC (2007), "North america elevation 1-kilometer resolution grid," *Internet site:* https://www.sciencebase.gov/catalog/item/4fb5495ee4b04cb937751d6d

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Anar Amgalan is a postdoctoral researcher at the Leonard Davis School of Gerontology at the University of Southern California. He received his Ph.D. from the Physics and Astronomy Department at Stony Brook University. His current research interests include quantification of brain aging process in health and diseases, particularly in Alzheimer's and traumatic brain injury.

LR Mujica-Parodi is Director of the Laboratory for Computational Neurodiagnostics, and Professor in Stony Brook University's Department of Biomedical Engineering. In addition, she is Research Staff Scientist and Lecturer in the Department of Radiology at Massachusetts General Hospital and Harvard Medical School (Martinos Center for Biomedical Imaging). She received her Ph.D. (Whiting Fellow) at Columbia University, studying mathematical logic and foundations of physics. She consequently trained in Clinical Neuroscience at Columbia University's College of Physicians and Surgeons. She was subsequently promoted to Assistant Professor there, where she performed research until being recruited by Stony Brook University. She is the recipient of the National Science Foundation's Career Award, the White House's Presidential Early Career Award in Science and Engineering, and the Fulbright Distinguished Scholar Award. Her research interests focus on the extension of control systems engineering and dynamical systems to human neuroimaging, with neurodiagnostic applications to neurological and psychiatric disorders.

Steven S. Skiena is Distinguished Teaching Professor of Computer Science and Director of the Institute for AI-Driven Discovery and Innovation at Stony Brook University. His research interests include data science, bioinformatics, and algorithms. He is the author of six books, including "The Algorithm Design Manual", "The Data Science Design Manual", and "Who's Bigger: Where Historical Figures Really Rank", and over 150 technical papers. Skiena received his B.S. in Computer Science from the University of Virginia and his Ph.D. in Computer Science from the University of Illinois in 1988. He is a Fellow of the American Association for the Advancement of Science (AAAS), a former Fulbright scholar, and recipient of the University of Virginia Engineering Distinguished Alumni Award (WahooWa!), the ONR Young Investigator Award and the IEEE Computer Science and Engineer Teaching Award. More info is available at http://www.cs.stonybrook.edu/skiena/.