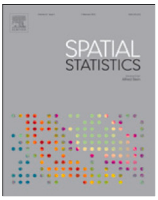
Spatial Statistics 50 (2022) 100612

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Spatial autocorrelation informed approaches to solving location–allocation problems

Daniel A. Griffith [a,*], Yongwan Chun [a], Hyun Kim [b]

[a] *School of Economic, Political, and Policy Sciences, University of Texas at Dallas, Richardson, TX, United States of America*

[b] *Department of Geography, University of Tennessee, Knoxville, TN, United States of America*

ARTICLE INFO

Article history:
Received 21 September 2021
Received in revised form 17 January 2022
Accepted 18 January 2022
Available online 31 January 2022

Keywords:
Location–allocation
p-median
Spatial autocorrelation
Spatial median
Spatial optimization

ABSTRACT

Surveying programs of study at institutions of higher learning throughout the world reveals that one natural disciplinary coupling is statistics and operations research, although these two specific disciplines currently lack an active synergistic research interface. Similarly, the development of spatial statistics and spatial optimization has occurred in parallel and nearly in isolation. This paper seeks to alter this situation by initiating transformative work at the interface of these two subdisciplines, encouraging considerably more future interaction between them. It outlines three ways spatial statistics can contribute to spatial optimization by exploiting spatial autocorrelation in georeferenced data: missing attribute value imputation (analogous to kriging); identifying colocations of local spatial autocorrelation hot spots and spatial medians; and, geographic tessellation stratified random sampling inputs to spatial optimization heuristics that successfully guide them to optimal location solutions. One contention emphasized throughout this paper is that this spatial statistics/optimization interface furnishes another vehicle for delivering spatial statistical benefits to society, which, in turn, benefits spatial statistics by providing better integration of it into novel interdisciplinary contexts.

[*] Corresponding author.
E-mail address: dagriffith@utdallas.edu (D.A. Griffith).

https://doi.org/10.1016/j.spasta.2022.100612
2211-6753/© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Spatial statistics offers benefits to science and society by furnishing methods for analysis and inference in the presence of correlated georeferenced data (Griffith, 2020), as, for example, sound statistical imputation via kriging and decision support via spatial autoregressive and Moran eigenvector spatial filtering (MESF; Griffith, 2003a,b) model specifications. Conversely, science and society increasingly recognize the existence and importance of spatial statistics, as demonstrated by a special issue of the US Centers for Disease Control's (2019) journal *Preventing Chronic Disease*, and by the emergence of such endeavors as the spatially integrated social sciences.[1] Its interface with spatial operations research, which at present is virtually nonexistent, is another emerging interdisciplinary endeavor at the frontiers of spatial statistics that should prove beneficial to both science and society. The overarching objective of this paper is to address this particular theme and crystalizing interface.

An important practical problem in spatial optimization is the location–allocation (L–A) problem, whose statistical context is nearly identical to that of geographically constrained cluster analysis. A brief overview of this L–A problem appears in Section 3.1 because it is the important exemplar used throughout this paper for illustrative purposes. Three features of this and many other spatial optimization problems potentially can benefit from spatial statistics. First is the problem of missing values. An advantageous strategy analysts confronted with incomplete data can pursue is to impute substitute values in order to decrease the uncertainty of a L–A problem's calculated optimal solution deviating from its complete data optimum. Georeferenced data imputation is a problem solved by spatial statistics. The other two spatial statistics issues discussed in this paper benefit searching for an optimal solution. The local indicators of spatial association (LISA) analysis, which appears to flag the region of a geographic landscape in which an optimal solution resides, and spatial sampling designs with their prevailing spatial autocorrelation (SA) acknowledgments, which increase optimal solution search efficiency. Educating practicing spatial statistician about these interface topics offers gains to these scholars in their repertoire of applications for delivering spatial statistical benefits to society as well as interdisciplinary scholarship.

2. Spatial optimization: a normative basis for decision making

Part of spatial statistics concerns inferential spatial decision support, often within the context of geographic map patterns characterizable by SA (e.g., Csillag and Boots, 2005). The results of such problems tend to be more descriptive/inferential than normative in nature. In contrast, spatial optimization may be defined as spatial decision support involving mathematically/computationally maximizing/minimizing a formulated objective function related to a geographic problem. It tends to be more normative than descriptive or inferential in nature. This differentiation between the two disciplinary missions, expressed in the context of map pattern, means that the primary use of SA in spatial statistics is improved descriptive (re statistical inference transfers a sample statistics description to its parent population parameters), whereas its primary potential use in spatial optimization is improved optimal solution search success (i.e., normative).

Spatial optimization embraces a wide variety of themes, from network route selection/shortest paths, through designing efficient/effective spatial sampling networks, land-use resources allocation, and regionalization like political districting, to a web of multiple demand points p-median L–A assignments (e.g., Delmelle, 2010; Tong and Murray, 2012; Ligmann-Zielinska, 2017). It shares topics with spatial statistics in each of these broad thematic areas. Common network subjects include minimum spanning trees and Delaunay triangulations and their dual Gabriel graphs vis-à-vis geographic neighbors identification (e.g., Dray et al., 2021, p. 21). Common sampling design subjects include the number and spacing across two-dimensions of monitoring locations (e.g., Müller, 2007; Mateu and Müller, 2012), such as that underlying tessellation stratified random sampling (e.g., Overton and Stehman, 1993). Common land-use resource allocation subjects include remotely sensed image classification in the presence of SA (e.g., Zhang et al., 2021). Common regionalization subjects

[1] https://escholarship.org/uc/spatial_ucsb_csiss.

include SA informed cluster analysis (e.g., Ballari et al., 2018). Finally, and the illustrative spatial optimization focus of this paper, SA latent in the geographic distribution of demand determining facility (e.g., retail outlet) locations via the L–A problem (e.g., Griffith, 2021).

The L–A problem (i.e., computing one or more supply locations in a way that most efficiently satisfies a geographically distributed set of demand points) has may variants (Church and Murray, 2018). Some versions consider facility capacities, whereas others do not. Some conceive a time sequencing for the incremental provision of multiple facilities, whereas others do not. Some minimize weighted distance to facilities – which relates to the bivariate median of spatial statistics (Small, 1990; Vardi and Zhang, 2000) – whereas others minimize distance to the furthest demand point, and yet others determine the smallest number of facilities together with their locations so that each demand point is serviced by at least one facility (i.e., the p-center problem). In other words, formulation of a spatial optimization objective function of this type may be dependent upon the goal of the optimization modeling in terms of: minimizing impedance or number of facilities, or maximizing (capacity) coverage or market share, among other criteria. The ensuing discussion focuses on this first objective for a single facility, a classical spatial optimization problem known as the p-median or L–A problem for locating p facilities.

Meanwhile, the relevant spatial optimization challenge to which spatial statistics contributes pertains to calculating an unknown optimal solution that often is very difficult (if not impossible) to determine for highly combinatorially complex location problems. The optimal solution often requires mathematical programming algorithms such as branch-and-bound (Morrison et al., 2016). This particular algorithm consists of a combinatorial tree enumeration of all possible solutions, with upper and lower optimal solution bounds applied to the branches of this tree in order to eliminate without evaluation large numbers of non-optimal solutions; if no bounds are available, it reduces to an exhaustive search. With regard to this example, this paper aspires to recommend ways to utilize local SA that effectively would improve these upper and lower estimated bounds. As such, this paper more generally aims to ascertain whether or not exploiting SA can facilitate improved spatial optimization solution efficiency by ultimately furnishing tools that increase solution speed and/or shrink the extent of a solution space, making previously intractable spatial optimization problems tractable.

3. Spatial statistics and spatial optimization: a potential for synergies

One objective for establishing an informative interface between spatial statistics and spatial optimization, especially with regard to the p-median L–A problem, is to improve the solution quality and speed in solving a class of problems sometimes encompassing a grid of facility locations—such as locating new fire stations to reduce fire loss in residentially-expanding communities, or closing existing schools in population-declining communities. Spatial scientists label these challenges L–A problems; they are very difficult to solve optimally (i.e., the selection of best locations), but still commonly formulated and somehow solved in many cases requiring planning networks of facility locations (e.g., parks, playgrounds). Church and Murray (2018) published a timely overview of this academic field. This paper examines the very common tendency for (dis)similar attribute values (i.e., weights; e.g., residential housing concentrations) related to facility service usage to cluster geographically (i.e., SA), an indispensable concept of spatial statistics. This paper demonstrates how using this available but overlooked SA information can assist in more efficiently and effectively solving these L–A problems, a topic nearly absent from the spatial sciences literature. This new tactic permits better and faster computational solutions to large L–A problems, for the location of either public or private facilities (a contribution to society).

Few published papers articulate relationships between SA latent in the geographic distribution of demand and corresponding L–A problem solutions. In other words, the literature describing relationships between SA and locational solutions to L–A problems is scant, with Griffith (2021) one of the very few exceptions. This paper also helps fill this gap, going well beyond the almost exclusively one-dimensional cases treated by Griffith (2021). For example, this paper documents a previously speculated relationship between (local) SA in service demand variables and solutions to L–A problems. It further documents the existence of this relationship. This investigation of the novel interface between spatial statistics (via SA) and spatial optimization (e.g., L–A) provides new knowledge and convincing evidence that should spur further spatial statistical oriented scientific research in related and neighboring disciplines, including not only operations research, but also spatial econometrics, regional science, and geography.

3.1. The L-A problem: a brief overview

The principal goal of p-median L–A problems is to find the locations of $p > 0$ central facilities in geographic space serving $n > 1$ demand points such that the cost of flows/travel between each demand point and its closest central facility is a minimum (Cooper, 1963; Hakimi, 1964); this setting is identical to determining p regional spatial medians. Publications by Scott (1970), Farahani and Hekmatfar (2009), and Eiselt and Marianov (2011) furnish a timely review of this L–A problem. As such, L–A problems involve determining locations in space, either a single location in its simplest form, or multiple locations in its more complicated form. The spatial statistical conceptualization counterparts are the bivariate median for a 1-median L–A problem (Small, 1990; Vardi and Zhang, 2000), and p regional medians for coterminous cluster analysis (Johnson and Wichern, 2008, §12.3–§12.5) determined subsets of points for a p-median L–A problem. In one of its simplest forms, a L–A problem aims to find a set of locations that minimizes the sum of distance-related weighted costs, which is its objective function. The p-median problem, one of the standard L–A problems, is widely applied in theory and in practice, and relates directly to the bivariate geometric median studied in spatial statistics (Eftelioglu, 2017). Solving the L–A problem is classified as NP-hard (Kariv and Hakimi, 1979), as well as solving it in continuous space is very numerically intensive and difficult even for small $p > 1$ (e.g., ReVelle and Eiselt, 2005); computationally calculating an optimal solution is arduous, even when feasible solutions exist. This situation parallels that for multivariate agglomerative cluster analysis techniques, many of which render a locally rather than a globally optimal solution (e.g., randomizing the order of input often produces different groupings). The k-means, k-medians, and k-medoids procedures are the most similar to a L–A problem (e.g., Koehn et al., 2010). Addressing this challenge continues to be a priority in the cluster analysis literature (e.g., Pacifico and Ludermir, 2021), with one of its difficulties being a simultaneous determination of the number of clusters (Ezugwu et al., 2021). This also continues to be a priority challenge in the L–A literature, building upon various heuristic algorithms[2] to seek global or near-global optima, with one of its difficulties being clusters comprised of coterminous points (Assunç et al., 2006; Guo, 2008). These algorithms are executed to quickly find optimal or near-optimal solutions (Mladenović et al., 2007). Meta-heuristics strategically provide a general framework to design heuristics to achieve an improved (i.e., much closer to a global optimum) solution and computational efficiency.

Inevitably, L–A problems are defined in space; for example, distances can be used as cost, and weights can be constructed as a surface (e.g., a map). Recent research discusses a relationship between spatial configurations in L–A problems and SA. Kim et al. (2019) argue that clusters of similar weight values can be utilized in solving a p-median problem. They find that demands closely located in space with similar weights tend to be assigned to the same facility, whereas facilities with very similar locations are not to be chosen together to serve an individual demand point. Furthermore, Griffith and Chun (2015) explore a relationship between local SA indices of weights and locations of p-median solutions, and show that the locations of p-median solutions move close to spatial clusters of high weights. In other words, the map pattern (i.e., SA) of weights can be informative when determining spatially optimal solutions. As such, spatial statistics offers a contribution to handling the globally optimal solution challenge of many p-median L–A problems.

[2] The application of an orderly sequence of simple, intuitive, quick, and efficacious computer calculation rules/operations based upon selected principles that allow each problem-solving set to quickly find a good/feasible but not necessarily globally optimal, solution.

3.2. The role of SA in spatial optimization

One issue is that the map pattern of weights quantifying discrete demand points distributed across a geographic landscape is not random; rather, it almost certainly exhibits positive SA, a fundamental theme in spatial statistics. When solving spatial optimization problems, SA can come into play in three distinct ways: (1) by helping to reduce potential solution spaces, and, hence, making the solving of large-size L–A problems, for example, tractable and more efficient to inventory the full range of feasible and optimal solutions; (2) by helping delineate an initial solution space that can lead to an optimal or a near optimal solution so that L–A models, for example, can efficiently render these exact or near optimal solutions based upon heuristics; and, (3) by enhancing the quality of solutions through the use of missing data values (e.g., weights in L–A problems) imputations to bolster optimization modeling results. Griffith and Paelinck (2018) illustrate these notions for p-median problems utilizing simulation experiments (essentially a proof of concept demonstration). They test the utility of SA using data for Poland where the geographic resolution is a powiat (formerly NUTS-4 areal units equivalent to US counties) and the number of units ($n = 380$) is much greater than any standard dataset pedagogically used to benchmark the performance of models or heuristics for p-median problems (e.g., the Beasley OR library[3] dataset has $n = 40$). The literature stresses that the quality of generated solutions using heuristics such as Tietz-and-Bart (TB; 1968), well-known algorithmic approaches due to their wide use and extendibility to other L–A problems, deteriorates with increasing p (Rosing et al., 1979; Brimberg and Hodgson, 2011; Daskin, 2013), and finding an exact solution also is limited to rather small n and/or p because of inherent computational complexities (Scott, 1970; Jamshidi, 2009; Daskin and Maass, 2015).

The hypothesis of this paper posits that exploiting SA can dramatically improve this computational situation. The following three avenues advance this possibility: imputation, local SA hot spot and spatial outlier analysis, and SA-informed spatial sampling. The first and third of these possibilities conceptualize demand weights as realizations of an auto-random variable field having a spatial dependence that can be modeled as SA. This particular correlation represents redundant information exploitable for spatial forecasting (e.g., kriging) and replicate sampling (e.g., tessellation stratified random sampling) purposes. Meanwhile, the second possibility conceptualizes demand as constituting a heterogeneous population in which the individual calculation terms in a SA statistic (i.e., a local SA measure) pertain to spatial outliers and/or hot/cold spots: unlike a stationary spatial random variable, some of the largest weights are nearby small weights, or the largest/smallest local SA statistics are too extreme. In the tradition of permutation-based non-parametric statistics (popularized for SA by Cliff and Ord, 1973), the L–A spatial optimization goal pertains to the specific map realization at hand, rather than inference about the existence of geographic clusters in a population, and hence local SA hot spot and spatial outlier analyses become instances of relative extreme clustering within a given geographic landscape. This context provides an opportunity for theory building using statistical outliers (e.g., Gibbert et al., 2020), not data description model identification and estimation.

A brief mathematically detailed L–A overview needs to preface any addressing of these three subjects. As a version of it, the standard p-median problem may be formulated as follows, using integer-linear programming for discrete space with n points, for which the solution is the set $\{(U_j, V_j), j = 1, 2, \ldots, p\}$:

$$\text{Min} : \sum_{i=1}^{n} \sum_{j=1}^{p} \lambda_{ij} w_i \sqrt{(u_i - U_j)^2 + (v_i - V_j)^2} \tag{1}$$

$$\text{s.t.} : \sum_{j=1}^{p} Y_j = p \tag{2}$$

$$\sum_{j=1}^{p} \lambda_{ij} = 1 \; \forall i \tag{3}$$

[3] http://people.brunel.ac.uk/~mastjjb/jeb/info.html.

$$\lambda_{ij} - Y_j \leq 0 \; \forall i, j \tag{4}$$

$$Y_j \in \{0, 1\} \tag{5}$$

$$\lambda_{ij} \in \{0, 1\} \tag{6}$$

where (u_i, v_i) is the Cartesian coordinates of demand point i, w_i is a weight (quantifying the in situ magnitude of demand) at demand point i, $d_{ij} = \sqrt{(u_i - U_j)^2 + (v_i - V_j)^2}$ is the Euclidean distance separating demand point i and facility j, λ_{ij} is 1 if demand point i is assigned to facility j, and 0 otherwise, and Y_j is 1 if facility j is selected, and 0 otherwise. Once p is given, this specification ensures that each demand point is allocated to one and only one of the p median facilities, and that at least one demand point is allocated to each facility. If $p = 1$, then the solution (U_1, V_1) is the global spatial median for the geographic distribution of weights. If $p > 1$, then the solution (U_j, V_j) is the jth of p regional medians (as noted previously, this outcome relates to the conventional multivariate statistics technique of cluster analysis).

4. Spatial statistics, data imputation, and spatial optimization

A practical problem frequently encountered by many empirical researchers is the presence of missing values in their datasets; for geospatial datasets, thematic maps contain gaps or holes. Effectively handling this data analysis complication in general has a long history (e.g., Schafer, 2000; Little and Rubin, 2002), with one popular procedure being to replace missing values by substituting comparable known/calculated values in their dataset places. Methodologically speaking, the expectation–maximization (E–M) algorithm supplies a sound statistical foundation for tackling this problem (e.g., McLachlan and Krishnan, 2008). One spatial statistical counterpart to this routine is geostatistical kriging; another more akin to standard E–M algorithm implementations relies on spatial autoregression and MESF (Griffith and Liau, 2021).

Pursuing spatial auto-normal model specifications, suppose $\mathbf{Y}_o$ denotes the n_o-by-1 ($n_o = n - n_m$) vector of observed response attribute values, and $\mathbf{Y}_m$ denotes the n_m-by-1 vector of missing response values. Let $\mathbf{X}_o$ denote the vector of predictor values for the set of observed response values, and $\mathbf{X}_m$ denote the vector of predictor values for the set of missing response values. Let $\mathbf{0}_o$ denote an n_o-by-n_m matrix of zeros, and $\mathbf{0}_m$ denote an n_m-by-1 vector of zeros. Partition vector $\mathbf{1}$ into $\mathbf{1}_o$, denoting the vector of ones for the set of observed response values, and $\mathbf{1}_m$, denoting the vector of ones for the set of missing response values. Let $\mathbf{I}_m$ denote an n_m-by-n_m identity matrix. Finally, let $\mathbf{W}$ denote the row-standardized spatial weights matrix, with $\mathbf{W}_{oo}$ containing the spatial weights for the pairs of known attribute value locations, $\mathbf{W}_{om}$ and $\mathbf{W}_{mo}$ containing the spatial weights for the pairs of known with unknown value locations, and $\mathbf{W}_{mm}$ containing the spatial weights for the pairs of unknown value locations. The spatial statistical autoregressive response (AR; i.e., a spatial econometrics spatial lag[4]) model specification may be written, using partitioned matrix notation, for imputation purposes as

$$\begin{aligned} \begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} &= \beta_0 \begin{pmatrix} \mathbf{1}_o \\ \mathbf{1}_m \end{pmatrix} + \begin{pmatrix} \mathbf{X}_o \\ \mathbf{X}_m \end{pmatrix} \boldsymbol{\beta} \\ &\quad + \rho \begin{pmatrix} \mathbf{W}_{oo} & \mathbf{W}_{om} \\ \mathbf{W}_{mo} & \mathbf{W}_{mm} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_o \\ \mathbf{Y}_m \end{pmatrix} + \begin{pmatrix} \mathbf{0}_o \\ -\mathbf{I}_m \end{pmatrix} (\mathbf{Y}_m) + \boldsymbol{\varepsilon}, \end{aligned} \tag{7}$$

where β_0 is the intercept term, $\boldsymbol{\beta}$ is the p-by-1 vector of regression coefficients for p covariates, ρ is the SA parameter, and $\boldsymbol{\varepsilon}$ is an p-by-1 vector of independent and identically (i.e., iid) normal random errors. In this specification, $\mathbf{Y}_m$ is treated as a vector of parameters during estimation (after all, it is

[4] Estimating Eq. (7) involves a Jacobian term that is more complex than the standard spatial statistical Jacobian term. It is given by $-\frac{2}{n-n_m}\left[\sum_{i=1}^{n} LN\,(1-\rho\lambda_i) - \sum_{k=1}^{n_m} LN\,(1-\rho\omega_k)\right]$, where LN denotes the natural logarithm, the first set of n eigenvalues (λ_i) is from matrix $\mathbf{W}$, and the second set of m eigenvalues (ω_k) is from matrix $\mathbf{W}_{mm}$.

a vector of conditional expectations), which requires nonlinear regression techniques for estimation purposes. As such, $\boldsymbol{\varepsilon}_m \equiv 0_m$. These imputations are those discussed by Griffith et al. (1989), and are equivalent to those obtained with kriging (Griffith and Layne, 1997).

Pursuing MESF (Griffith, 2003a) specifications circumvents the preceding complications, which are more severe for non-normal random variables (e.g., Poisson, negative binomial, Bernoulli, and binomial). It replaces the spatial weights matrix components in Eq. (7) with an eigenvector spatial filter (ESF). This ESF – which is a linear combination of selected spatial weights matrix eigenvectors – specification permits the calculation of standard imputations for a linear model using the following estimation equation:

$$\begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} = \beta_0 \begin{pmatrix} \mathbf{1}_o \\ \mathbf{1}_m \end{pmatrix} + \begin{pmatrix} \mathbf{X}_o \\ \mathbf{X}_m \end{pmatrix} \boldsymbol{\beta}_X + \begin{pmatrix} \mathbf{0}_o \\ -\mathbf{I}_m \end{pmatrix} (\mathbf{Y}_m) + \sum_{k=1}^{K} \begin{pmatrix} \mathbf{E}_{o,k} \\ \mathbf{E}_{m,k} \end{pmatrix} \beta_{E_k} + \boldsymbol{\varepsilon}, \qquad (8)$$

where $\boldsymbol{\beta}_X$ denotes the p-by-1 vector of regression coefficients for p covariates, K is the number of eigenvectors selected from the candidate vector set for constructing an ESF, $\mathbf{E}_{o,k}$ and $\mathbf{E}_{m,k}$ respectively denote the parts of eigenvector k associated with the observed and missing values, and $\boldsymbol{\beta}_{E_k}$ denotes the regression coefficient for eigenvector $\mathbf{E}_k$ used to construct the ESF in question. Imputations here are based upon the generalized linear model (GLM) counterpart to Eq. (8), outlined in Griffith and Haining (2010), for binomial random variables, and in Griffith (2013), for Poisson random variables.

Although few published papers articulate relationships between SA latent in the geographic distribution of demand and corresponding L–A problem solutions, Griffith (1997, 2003b) constitutes one of the few exceptions. He exploits SA in the geographic distribution of demand having missing values to compute imputations for them for $p = 1$ and 2 spatial median continuous space exact solution problems. Without these approximate values, demand becomes zero, causing its point location to disappear from the L–A problem. Although these two papers do not focus on the relationship between the latent level of SA and L–A spatial optimization, they show that a map pattern affects optimal L–A solutions. With regard to societal concerns, this type of solution is important when a L–A problem formulation is in terms of attributes governed by confidentiality/privacy restrictions. Government agencies rarely release income figures, disease counts, genetic marker recordings, and the such, for sufficiently small geographic areas when these quantities become sensitive data (e.g., too few households/firms inhabit an areal unit polygon to ensure anonymity through aggregation).

For illustrative purposes, Fig. 1 portrays the geographic distribution of industrial employment location quotients (LQs) as weights, w_i ($i = 1, 2, \ldots, 380$), by powiaty (roughly equivalent to US counties) across Poland. SA is visibly observable in the two maps. Furthermore, Fig. 2 presents and pinpoints some of the $p = 1, 2, \ldots, 10$ median solutions for this Poland dataset; their calculations were the L–A solutions obtained with IBM CPLEX Optimizer 12.8.[5] Among them, solutions for $p = 1$, 2, 3, and 4 appear on the Fig. 2a map. Because both n and p are relatively small, all instances were solved to optimality within several minutes.

Table 1 summarizes output from a simulation experiment utilizing the same Polish data, but to explore impacts of randomly missing data on the $p = 1$ L–A solution. Eq. (8) furnishes the imputation instrument, and suppression was for a range of occurrences between 0% and 90%, in 10% increments. LQs are rescaled (i.e., divided by the matching coarser level reference percentage) georeferenced binomial random variables, population density commonly is an effective social science covariate, and imputation exploits SA in its role as redundant attribute information. For this specimen empirical example, log-population density (as the weighted sum of its linear and quadratic versions) accounts for roughly 14% of the geographic variation across Poland powiaty in the 2013 industrial LQs specified as a binomial regression response variable, with 29 MESF adjusted

[5] This advanced computer optimization package provides flexible, high-performance mathematical programming solving procedures for linear, mixed integer, quadratic, and quadratically constrained problems. Its efficiency rests upon limiting an optimal solution to a predefined set of potential solution points on a plane.

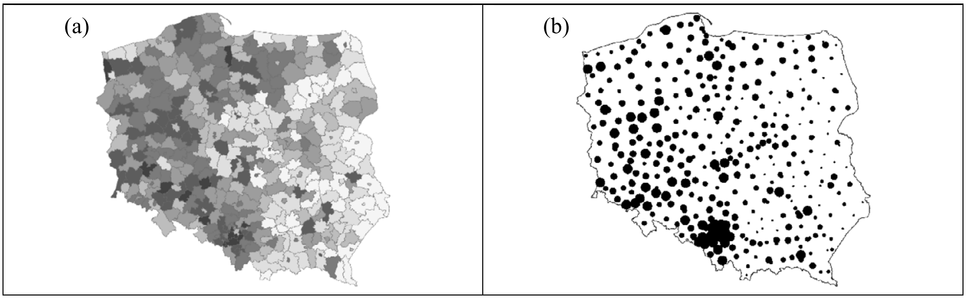
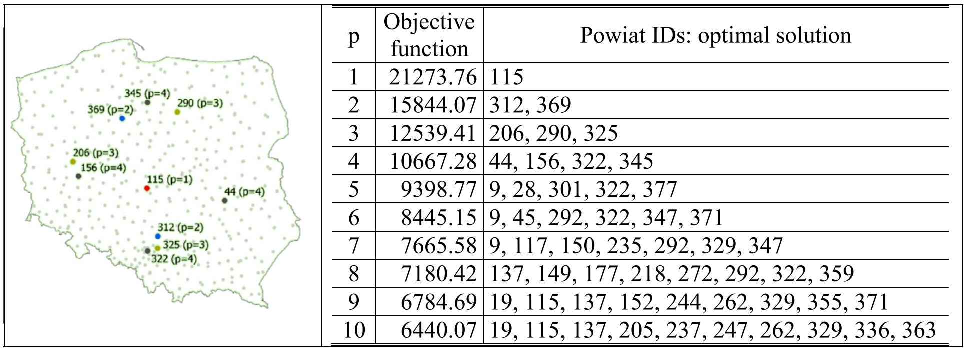


Fig. 1. The 2013 Polish industrial employment LQs geographic distribution. (a) by powiat; LQ magnitudes are directly proportional to grayscale darkness. (b) converted to weights attached to powiat centroids; LQs are directly proportional to their filled circle size.



p	Objective function	Powiat IDs: optimal solution
1	21273.76	115
2	15844.07	312, 369
3	12539.41	206, 290, 325
4	10667.28	44, 156, 322, 345
5	9398.77	9, 28, 301, 322, 377
6	8445.15	9, 45, 292, 322, 347, 371
7	7665.58	9, 117, 150, 235, 292, 329, 347
8	7180.42	137, 149, 177, 218, 272, 292, 322, 359
9	6784.69	19, 115, 137, 152, 244, 262, 329, 355, 371
10	6440.07	19, 115, 137, 205, 237, 247, 262, 329, 336, 363

Fig. 2. Selected p-median solutions determined by CPLEX; the set of potential solution points is the 380 areal unit centroids.

spatial weights matrix eigenvectors – 22 depicting positive and nine depicting negative SA, selected from respective candidate sets containing 85 and 119 vectors – accounting for another roughly 45% of this geographic variance. These LQs exhibit moderate positive SA (i.e., Moran Coefficient = 0.45, Geary Ratio = 0.57), whereas the binomial regression residuals exhibit only trace SA (with their individual index null hypothesis probabilities exceeding 0.2). The accompanying excess binomial variation decreases from 1916.7 to 1454.7 by including the population density covariates, and then to 680.6 by including the MESF eigenvectors; although this overdispersion still is excessive, its reduction corroborates the prominent role SA plays in these geospatial data.

One important reinforced finding from Table 1 is that even though the $p = 1$ solution is unbiased with or without acknowledging latent SA for random data value suppression, involving this source of duplicate information in the solving of spatial optimization problems reduces uncertainty associated with the final solution. This contribution can be a windfall when confronting societal issues.

5. Spatial statistics, hot spot analysis, and spatial optimization

The widely known L–A majority theorem supplies a critical ingredient for a convincing demonstration of an important role LISA (e.g., Anselin, 1995) statistics can play in solving L–A problems. To begin,

MAJORITY THEOREM (MT): For an n destinations $p = 1$ source location–allocation (i.e., p-median) problem in continuous space, with $n > 1$ and Euclidean distance as the metric, if a single weight $w_k > \frac{\sum_{i=1}^{n} w_i}{2}$, then the demand point (u_k, v_k) is the optimal location (i.e., spatial median) solution.

Table 1
Missing and imputed weight results for the 2013 Poland industrial LQs after converting the country's coordinates into those on a unit square; random sampling without replacement weights suppression; 1000 simulation replications.

% suppressed	Missing weights set to 0			Missing weights imputed		
	n	U_1	V_1	n	U_1	V_1
0	380	0.44548	0.44324	380	0.44548	0.44324
10	342	0.44528 (0.00554)	0.44358 (0.00684)	380	0.44620 (0.00128)	0.44332 (0.00144)
20	304	0.44564 (0.00877)	0.44326 (0.01069)	380	0.44689 (0.00168)	0.44331 (0.00190)
30	266	0.44485 (0.01093)	0.44402 (0.01359)	380	0.44739 (0.00193)	0.44354 (0.00226)
40	228	0.44553 (0.01409)	0.44416 (0.01685)	380	0.44814 (0.00212)	0.44343 (0.00237)
50	190	0.44504 (0.01677)	0.44383 (0.02060)	380	0.44875 (0.00221)	0.44364 (0.00242)
60	152	0.44509 (0.02043)	0.44494 (0.02500)	380	0.44936 (0.00207)	0.44356 (0.00239)
70	114	0.44441 (0.02477)	0.44307 (0.03222)	380	0.45006 (0.00193)	0.44348 (0.00225)
80	76	0.44389 (0.03273)	0.44487 (0.04213)	380	0.45071 (0.001172)	0.44375 (0.00203)
90	38	0.44303 (0.04940)	0.44492 (0.05965)	380	0.45133 (0.00128)	0.44374 (0.00145)

NOTE: standard errors are in parentheses.

PROOF: see Witzgall (1964).

In other words, for the global spatial median case, if the weight at any location is more than half of the total sum of all n weights (each of which must be positive), then the spatial median collocates with that demand point. In this context, weights function in a fashion similar to frequencies of observations (i.e., repeated/tied attribute values that may occur many times) in classical statistics. This result relates to the breakdown point (i.e., a measure of resistance to misbehavior of observations in a dataset) of a median, which is 50%.

The MT describes a data anomaly, one that is both a univariate and a geographic outlier. Its local SA Moran value is a function of $w_n = \sum_{i=1}^{n-1} w_i + 1$ (hence, it is a univariate outlier), whose designation as the majority weight is without loss of generality because it occupies a random location, coupled with the minimum majority situation of

$$\frac{(n-2)\,w_n+1}{n} c_{nj} \left(w_j - \frac{2w_n+1}{n} \right), \tag{9}$$

where $\frac{2w_n+1}{n}$ is the mean of the weights, c_{nj} denotes the cell (n, j) entry in the accompanying spatial weights matrix, $\mathbf{C}$, which is a substantial negative quantity (i.e., local negative SA), indicating a spatial outlier. The neighboring values for location n have local Moran values that are a function of

$$\left(w_i - \frac{2w_n+1}{n} \right) c_{ij} \left(w_j - \frac{2w_n+1}{n} \right) + \left(w_i - \frac{2w_n+1}{n} \right) c_{in} \left(\frac{(n-2)\,w_n+1}{n} \right), j \neq n. \tag{10}$$

The second term in expression (10) is negative and larger than the first term, which is positive. Therefore, if the remaining $(n-1)$ weights are identically distributed, then the only local SA cluster that would emerge by other than chance is for the location affiliated with weight w_n; both this location n and its neighbors would exhibit significant negative local SA. These results characterize all MT geographic landscapes.

Simulation experiments explored a wide range of point distributions (i.e., uniform, normal, skewed, and sinusoidal; see Fig. 3b) for a $p = 2$ continuous space L–A solution, with a challenging second-largest weight being nearly as large as the largest weight, and in every case, the majority weight is in the optimal solution set. Fig. 3c portrays an example showing that the majority weight point is a hot spot; in that particular simulated case, the second largest weight demand point also constitutes a hot spot (formed by it and one of its very close demand points), but with lower statistical significance. Consequently, a conjecture can be made that, for $p = 2$ median continuous

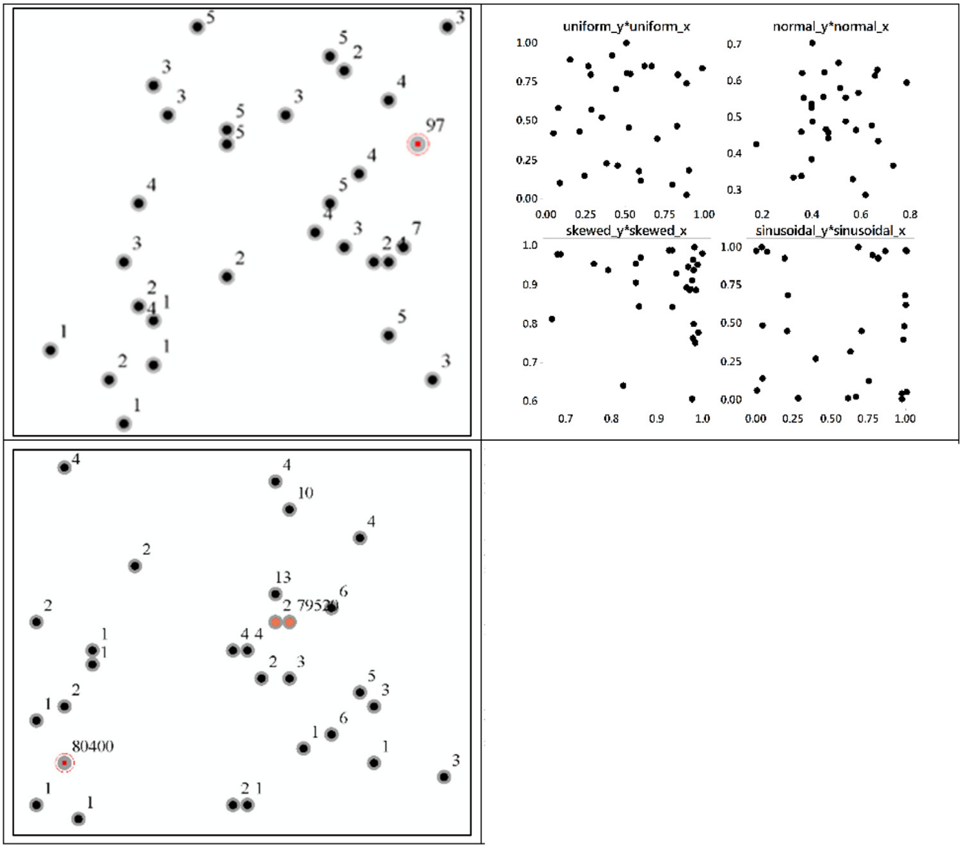


Fig. 3. MT illustrations for $p = 1$ and $p = 2$; numbers near points are their respective demand weights. Top left (a): a 30-by-30 grid, $p = 1$, $n = 30$, Moran Coefficient = 0.02, $w_{\max} = 50.34\%$ of the total weights; red denotes the optimal location and hot spot; black denotes nonsignificant local SA indices. Top right (b): specimen unit square point pattern distributions for determining $p = 2$ solutions; 1000 replications; $w_{max} = 50.25\%$ and $w_{2nd\text{-largest}} = 49.70\%$. Bottom left (c): a sample uniform distribution of points for $p = 2$; red denotes the dominant hot spot, and orange denotes the second largest hot spot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

space L–A problems with $n > p$ demand points and Euclidean distance, if a weight $w_k > \sum_{i=1}^{n} w_i/2$, then the demand point (u_k, v_k) is in the optimal solution set of locations.

For illustrative purposes, a simulation experiment evaluated the $p = 1$ case for the aforementioned various random variable types. Table 2 reports employed nearest neighbor threshold values whose judicious selection ensured that each demand point had at least one neighboring point. As Fig. 3 entries and the preceding discussion demonstrate, geographic distributions of LISA tend to constitute two groups, the pair coinciding with the MT optimal locations, and essentially all others; these were the criteria used to classify the LISA for comparison purposes (see Table 2). Not surprisingly, the LISA groups neither conform to bell-shaped curves nor display constant variance (i.e., they are non-normal random variables), compelling the application of non-parametric statistical techniques. Table 2 tabulates an analysis of variance type of output for Kruskal–Wallis treatments of these data. The computed chi-square statistics based upon 1000 observations are roughly a thousand times greater than even the usual extreme null hypothesis $\alpha = 0.01$ critical

Table 2
Kruskal–Wallis results for the difference of MT and all other LISA, random weight types, and demand point patterns; $n = 1000$ and 1000 replications.

Statistical distribution of weights	Underlying geographic distribution of points					
	Uniform			Skewed		
	Threshold distance identifying neighbors	n_i range	Chi-square (1 df)	Threshold distance identifying neighbors	n_i range	Chi-square (1 df)
Uniform	0.10	2–64	5398.9 ($p < 0.0001$)	0.03	1–64	4964.5 ($p < 0.0001$)
Normal	0.10	3–59	5190.1 ($p < 0.0001$)	0.03	1–55	4891.2 ($p < 0.0001$)
Poisson	0.10	2–63	5754.8 ($p < 0.0001$)	0.03	1–57	5428.3 ($p < 0.0001$)

NOTE: df denotes degrees of freedom; n_i denotes the number of neighboring points for demand point location i.

value of 6.635; in other words, the LISA results are tremendously inconsistent with a null hypothesis stating that they are the same for the MT optimal and the $(n - 1)$ other demand point locations, and certainly satisfy a criterion such as being at least four times greater than the designated critical value (Ryan, 2009). In practical terms, this conclusion basically implies that LISA statistics identify a negative hot spot that is (nearly) equivalent to the MT optimal solution.

An insightful finding about SA corroborated here is that the MT $p = 1$ optimal solution coincides with a dramatically statistically significant LISA outlier value; in other words, these two values collocate at an exclusive negative SA hot spot on a map. The potential academic contribution underscored here is for empirical spatial scientists to exploit LISA and Getis–Ord (e.g., Ord and Getis, 1995) G_i^* statistics (which focus on positive SA hot spots) to better determine spatial optimization solutions, such as those for L–A problems, when addressing societal issues.

6. Spatial statistics, SA-informed sampling, and spatial optimization

The combinatorial nature of L–A problems, like that for most spatial optimization research questions needing answered, results in a massive number of possible feasible solutions, often too many to evaluate in a sensible amount of time, let alone almost instantly, for reasonably modest values of n and p. Discrete L–A problems can circumvent this complication to a large degree by limiting the infinite number of solution points on a plane to a relatively small finite number of designated points. A mixed integer optimizer such as CPLEX can solve such discrete L–A problems. The count of different ways to allocate n demand points to p central facilities such that each facility serves at least one demand point is given by $[{}_nC_p] \times [p \cdot (n - p)]$—location level × allocation level. Theoretically, an exact solution may not be obtainable in a reasonable timeframe because this problem is from a NP-hard class. Heuristic algorithms arose in response to this timing challenge. In the past, the TB heuristic formulated by Teitz and Bart (1968) has been a popular L–A problem solver, in part because it was the first such heuristic. It is a greedy algorithm (i.e., it makes a locally optimal decision at each iteration) with substitution. It begins with a random set of facility locations selected from the n demand points—its solution space is a set of discrete points. Next, it allocates these n demand points to these selected facilities using the shortest distance between each pair (i.e., the Euclidean metric), and then computes the objective function value. Finally, for each point A in a current solution, and for each point B not in a current solution, the objective function is recalculated after swapping B and A (i.e., substitution). If the new objective function value improves (i.e., less than the incumbent value, further decreasing toward a lower minimum), then B and A replace each other in the solution; otherwise, the solution retains these points' previous allocations. This heuristic found the optimal solution for every 49-demand-points p-median problem that Rosing et al. (1979) submitted to it, regardless of the inputted starting solution. Furthermore, Fotheringham et al. (1995) restarted TB only once for a set of 120 different L–A problems they tested. Meanwhile, ALTERN (also ALT and ALA), formulated by Cooper (1964), solves continuous versions of the L–A problem. It embraces the following two principles: (1) if an allocation of n demand points to p spatial medians is known, then the $p = 1$ problem can be solved for each grouping of demand points (i.e., regional medians); and, (2) if the locations of p spatial medians are known, then n demand points can be allocated to their respective closest ones. This heuristic alternates between these two principles until the computed objective function value of weighted distance no longer decreases. ALTERN optimally solved (except for trivial rounding error) an $n = 500$ and $p = 2$ problem, using random initial demand point allocations, 485 out of 1000 times; it also found nine different local optima, a particular one, whose objective function is 0.1% greater than the optimal solution objective function, 329 times—this pair of locations perceptively deviates from the optimal pair, but with positions very close to them. Because both heuristics are sensitive to their initial solutions, which often are selected at random, TB and ALTERN are not guaranteed to produce a global optimum; each easily can be trapped at a local optimum location. Consequently, the practice of executing multiple initiations with random starting solutions is widespread, with the best output selected as the final solution. This same approach characterizes attempts to secure globally optimal conventional multivariate cluster analysis results, too (van der Kloot et al., 2005).

This local–global optima dilemma motivated strategy developments to resolve it. Meta-heuristics in terms of simulated annealing (SiAn) also have been applied to solve L–A problems—for the most part, discrete space ones. Murray and Church (1996a) argue that SiAn is a competitive stratagem for solving the p-median problem. Their computational results for 40 heuristic solutions employing selected OR-Library datasets from Beasley show that their approach to the p-median problem successfully finds (near) optimal solutions for a set of test datasets, with the gaps between their solutions and the optimal solutions being very small. Some other popular meta-heuristics include genetic algorithms (Hosage and Goodchild, 1986; Chaudhry et al., 2003; Salhi and Gamal, 2003), Tabu search (Crainic et al., 1993; Rolland et al., 1996; Salhi, 2002), heuristic concentration (Rosing and ReVelle, 1997; Rosing and Hodgson, 2002), and ant colony (Levanova and Loresh, 2004; Arnaout, 2013). Mladenović et al. (2007) present a comprehensive review of meta-heuristic schemes for solving L–A problems. This section adds a treatment of SA-informed sampling to this perspective by presenting a summary of an examination of a $p = 2$ case for illustrative purposes.

Fig. 4 portrays results for a skewed distribution of points, a point pattern concentrating in the southeastern part of the geographic landscape. The non-uniqueness quality of $p = 2$ optimal solutions generates, in this setting, north–south coupled with east–west, or northeast-southwest coupled with northwest-southeast, types of solution pairings; various of the portrayed complete data regional spatial median scatterplots align with each of these groupings, with their orthogonal groups removed for visual clarity (yielding variable replication numbers). SA, exploited by a geographic tessellation stratified random sampling design (Overton and Stehman, 1993) based upon a 10-by-10 grid, dampens this non-uniqueness trait, although all Fig. 4 portrayals still exhibit some degree of circular formation for their collective regional spatial medians. This non-uniqueness spawned regional spatial median circular point pattern persists, in part, because of the intermingling of SA and skewed point distribution effects. These findings supplement as well as corroborate those appearing in Griffith (2021).

One enlightening finding about SA uncovered here is that it can help simplify the $p = 2$ L–A spatial optimization problem. Positive SA characterizes much, if not most, georeferenced data, and renders collections of attribute values cohabitating a tessellation stratification areal unit more similar, on average, than attribute values between these areal units. This relative homogeneity of weight values allows repeated samples from a point pattern population to deliver similar regional spatial median locations. This methodology enables solutions for excessively large $p = 2$ problems, say $n = 100,000$, which cannot be solved exactly in a practical amount of real time; even CPLEX would find such a problem not only intractable, but also nearly unsolvable. One way to strengthen such a sampling result is to input it as an initial solution for a heuristic algorithm, such as ALTERN (see Tables 3 and 4). This plan rivals the current routine of randomly initiating a sizeable number of heuristic executions, and then selecting the solution with the smallest objective function value; still, this exercise is not immune to suffering from the $p = 2$ non-uniqueness complication.

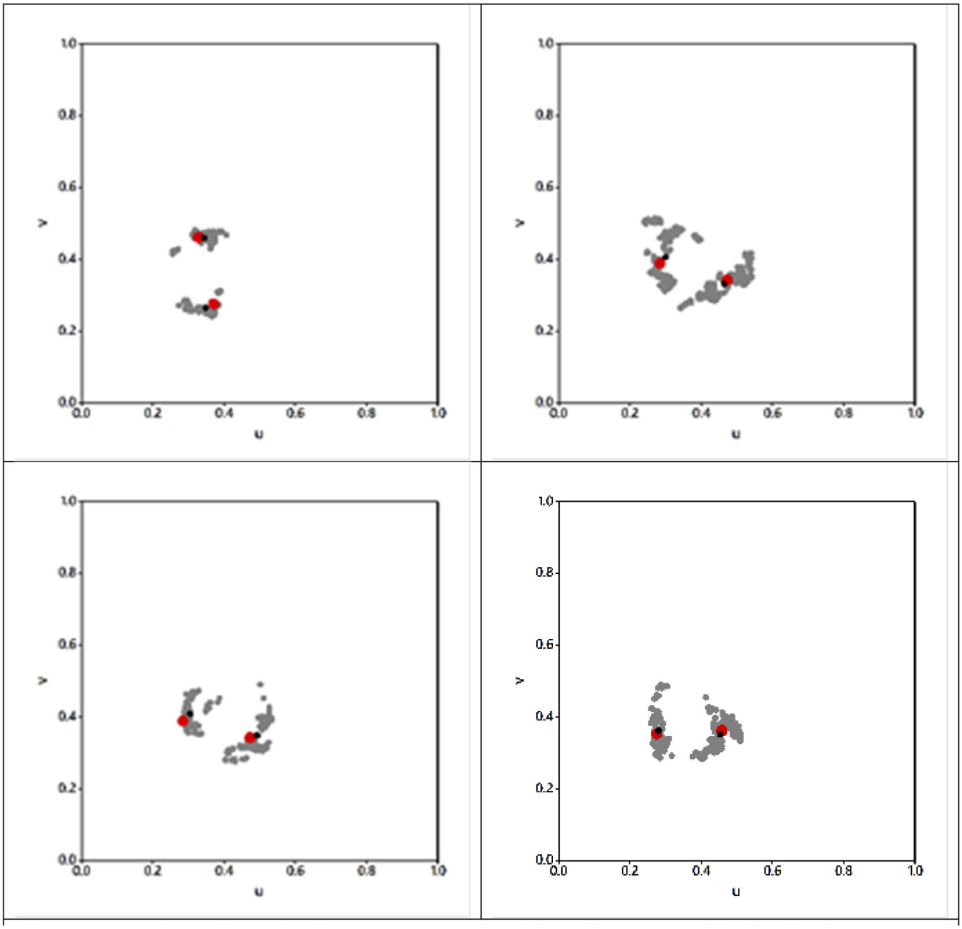


Fig. 4. Specimen skewed point pattern $p = 2$ spatial median geographic distributions [gray, red, and black filled circles respectively denote sample ($n \leq$ 1500), complete data, and averaged sample spatial medians); these latter two essentially collocate. Top left (a): random independent weights. Top right (b): linear weights gradient. Bottom left (c): quadratic weights gradient. Bottom right (d): periodic weights geographic distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7. Conclusions and implications

This paper both presents some new, and reviews and extends some existing, contributions of spatial statistics to the spatial optimization arena, with special reference to L–A problems, advocating the need for a research active interface between these two subfields. Synergies emerging from such an interface offer valuable contributions to society in the form of advancing the frontiers of feasible spatial optimization solution spaces, a particularly important improvement for the realm of L–A solutions to public facility provisions. The three specific interfaces surveyed here, all of which exploit SA latent in geospatial data, are spatial statistical imputation assistance, understanding LISA and Getis–Ord G_i^* statistics within the pre-analysis context of potentially isolating optimal solution locations, and adapting spatial resampling to better initiate and guide heuristic L–A algorithms. Future research should expand investigations of these three themes well beyond their rudimentary

Table 3
Skewed distribution of demand points $p = 2$ simulation experiment output summaries; a single parent sample of $n = 500$ for each situation, and 1500 resamples (without replacement) from it of size $n = 100$.

Demand points distribution	Spatial median coordinate	Map pattern of weights			
		Random ($n = 635$)	Linear gradient ($n = 1500$)	Quadratic gradient ($n = 749$)	Periodic (i.e., SINE function; $n = 1341$)
Region #1	Complete data U	0.32793	0.47314	0.47314	0.45584
	Complete data V	0.45947	0.34043	0.34043	0.36090
	Sampled data U	0.34342 (0.02984)	0.46323 (0.05422)	0.49156 (0.03073)	0.45324 (0.02980)
	Sampled data V	0.45773 (0.01411)	0.33083 (0.03364)	0.34734 (0.04257)	0.34928 (0.03265)
	ALTERN U	0.33879	0.47309	0.48286	0.45255
	ALTERN V	0.45970	0.33929	0.33414	0.35550
Region #2	Complete data U	0.37115	0.28405	0.28405	0.27519
	Complete data V	0.27462	0.38864	0.38864	0.35211
	Sampled data U	0.34797 (0.03035)	0.29990 (0.02951)	0.30443 (0.01425)	0.27951 (0.01330)
	Sampled data V	0.26370 (0.01562)	0.40449 (0.06778)	0.40760 (0.03035)	0.36105 (0.04159)
	ALTERN U	0.36234	0.28454	0.29634	0.27222
	ALTERN V	0.26560	0.38971	0.40565	0.35736

NOTE: standard errors (the input for standard distances) are in parentheses.
Map pattern generators (all demand weight specifications include adding 1 to eliminate the prospect of $w_i = 0$; all map-wide averages are roughly five): (1) random—Poisson ($\mu = 4$) + 1; linear gradient—$9(u + v) + 1 + 0.01\times$ Normal(0, 1); quadratic gradient—$5(u + v)^2 + 1 + 0.01\times$ Normal(0, 1); and, periodic—5[SIN($u{\cdot}\pi$)+ 2SIN($v{\cdot}\pi$)] +1 + 0.01× Normal(0, 1).
Resampling procedure: initially, a 10-by-10 square grid tessellation superimposed on a unit square created geographic strata, and a sample of size five was drawn at random from each stratum; each replication randomly selected one sample from each stratum.

Table 4
Objective function values for Table 3 specimen analysis.

Spatial median	Random	Linear gradient	Quadratic gradient	Periodic (i.e., SINE function)
Exact	297.32380	474.27980	223.22160	808.62540
Heuristic from sample	297.39392	474.28241	223.22569	808.74801
Sample average	297.88550	476.81080	224.44630	809.48510

examinations described here. This interface also should expand to include other undertakings, such as the translation of latent SA into strong valid inequality conditions and variable reduction, reducing the number of constraints spatial optimization problems require.

With regard to this newly mentioned interface activity, Kim et al. (2019) already outline the p-median problem with SA treatments, which focuses on ensuring optimality via a mixed integer programming formulation, while reducing the computational complexity of the standard p-median problem. Their novel formulation consists of the following two components that seek to enhance p-median problem solving capabilities: (1) construction of strong and effective valid inequality conditions; and, (2) exploration and identification of the best spatial extent for a given problem in order to reduce the number of assignment variables between its n demand points and p candidate spatial medians.

In conclusion, spatial statistics offers benefits to science and society by furnishing sound statistical imputation and decision support in the presence of SA, which should only encourage science and society to continue to, as well as increasingly, recognize the existence and importance of spatial statistics and its applications.

Funding acknowledgment

This research was supported by the National Science Foundation, United States, grant BCS-1951344. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

References

Anselin, L., 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27, 93–115.

Arnaout, J.P., 2013. Ant colony optimization algorithm for the euclidean location–allocation problem with unknown number of facilities. J. Intell. Manuf. 24 (1), 45–54.

Assunç, R., Neves, M., Câ, G., Freitas, C., 2006. Efficient regionalisation techniques for socio-economic geographical units using minimum spanning trees. Int. J. Geogr. Inf. Sci. 20 (7), 797–811. http://dx.doi.org/10.1080/13658810600665111.

Ballari, D., Giraldo, R., Campozano, L., Samaniego, E., 2018. Spatial functional data analysis for regionalizing precipitation seasonality and intensity in a sparsely monitored region: unveiling the spatio-temporal dependencies of precipitation in Ecuador. Int. J. Climatol. 38, 3337–3354. http://dx.doi.org/10.1002/joc.5504.

Brimberg, J., Hodgson, J., 2011. Chapter 15: Heuristics for location models. In: Eiselt, H., Marianov, V. (Eds.), Foundations of Location Analysis. Springer., Berlin, pp. 335–355.

Chaudhry, S., He, S., Chaudhry, P., 2003. Solving a class of facility location problems using genetic algorithm. Expert Syst. 20, 86–91.

Church, R., Murray, A., 2018. Location Covering Models: History, Applications and Advancements. Springer, Cham, Switzerland.

Cliff, A., Ord, J., 1973. Spatial Autocorrelation. Pion, London.

Cooper, L., 1963. Location-allocation problems. Oper. Res. 11 (3), 331–343.

Cooper, L., 1964. Heuristic methods for location–allocation problems. SIAM Rev. 6, 37–53.

Crainic, T., Gendreau, M., Soriano, P., Toulouse, M., 1993. A tabu search procedure for multicommodity location/allocation with balancing requirements. Ann. Oper. Res. 41 (4), 359–383.

Csillag, F., Boots, B., 2005. A framework for statistical inferential decisions in spatial pattern analysis. Can. Geogr. 49, 172–179.

Daskin, M., 2013. Network and Discrete Location: Models, Algorithms and Applications, second ed. Wiley, NY.

Daskin, M., Maass, K., 2015. The p-median problem. In: Laporte, G., Nickel, S., da Gama, F. Saldanha (Eds.), Location Science. Springe, Berlin, pp. 21–45.

Delmelle, E., 2010. Spatial optimization methods. In: Wharf, B. (Ed.), Encyclopedia Of Human Geography. Sage, Thousand Oaks, CA, pp. 2657–2659.

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., Wagner, H., 2021. Package 'adespatial', CRAN. https://cran.r-project.org/web/packages/adespatial/adespatial.pdf.

Eftelioglu, E., 2017. Geometric median. In: Shekhar, S., Xiong, H., Zhou, X. (Eds.), Encyclopedia of GIS, second ed. Springer., Cham, Switzerland, pp. 701–704.

Eiselt, H., Marianov, V., 2011. Foundations of Location Analysis. Springer, NY.

Ezugwu, A., Shukla, A., Agbaje, M., Oyelade, O., José-García, A., Agushaka, J., 2021. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature. Neural Comput. Appl. 33, 6247–6306. http://dx.doi.org/10.1007/s00521-020-05395-4.

Farahani, R., Hekmatfar, M., 2009. Facility Location: Concepts, Models, Algorithms and Case Studies. Physica-Verlag, Berlin.

Fotheringham, A., Densham, P., Curtis, A., 1995. The zone definition problem in location-allocation modeling. Geogr. Anal. 27, 60–77.

Gibbert, M., Nair, L., Weiss, M., 2020. Using outliers for theory building. Organ. Res. Methods 24 (12), 172–181. http://dx.doi.org/10.1177/1094428119898877.

Griffith, D., 1997. Using estimated missing spatial data in obtaining single facility location–allocation solutions. L'Espace GÉ 26, 173–182.

Griffith, D., 2003a. Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization. Springer-Verlag, Berlin.

Griffith, D., 2003b. Using estimated missing spatial data with the 2-median model. Ann. Oper. Res. 122, 233–247.

Griffith, D., 2013. Estimating missing data values for georeferenced Poisson counts. Geogr. Anal. 45, 259–284.

Griffith, D., 2020. A family of correlated observations: from independent to strongly interrelated ones. Stats 3, 166–184.

Griffith, D., 2021. Articulating spatial statistics and spatial optimization relationships: Expanding the relevance of statistics. Stats 4 (4), 850–867.

Griffith, D., Bennett, R., Haining, R., 1989. Statistical analysis of spatial data in the presence of missing observations: a methodological guide and an application to urban census data. Environ. Plan. A 18, 1511–1523.

Griffith, D., Chun, Y., 2015. Spatial autocorrelation in spatial interactions models: geographic scale and resolution implications for network resilience and vulnerability. Netw. Spat. Econ. 15, 337–365.

Griffith, D., Haining, R., 2010. Analyzing small geographic area datasets containing values having high levels of uncertainty. In: Tate, N., Fisher, P. (Eds.), Accuracy 2010, Proceedings of the 9th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences (University of Leicester, 20-23 July), Leicester, UK. MPG Books Group, pp. 289–292.

Griffith, D., Layne, L., 1997. Uncovering relationships between geo-statistical and spatial autoregressive models. In: The 1996 Proceedings on the Section on Statistics and the Environment. American Statistical Association, pp. 91–96.

Griffith, D., Liau, Y.-T., 2021. Imputed spatial data: cautions arising from response and covariate imputation measurement error. Spatial Stat. 42, 100419, 1-12.

Griffith, D., Paelinck, J., 2018. Chapter 2.6: Relationships between spatial autocorrelation and solutions to location–allocation problems. In: Morphisms for Quantitative Spatial Analysis. Berlin, In: Advanced Studies in Theoretical and Applied Econometrics Series, Springer-Verlag, pp. 18–22.

Guo, D., 2008. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). Int. J. Geogr. Inf. Sci. 22 (7), 801–823. http://dx.doi.org/10.1080/13658810701674970.

Hakimi, S.L., 1964. Optimum location of switching centers and the absolute centers and medians of a graph. Oper. Res. 12, 450–459.

Hosage, C., Goodchild, V., 1986. Discrete space location–allocation solutions from genetic algorithms. Ann. Oper. Res. 6 (2), 35–46.

Jamshidi, M., 2009. Median location problem. In: Farahani, R., Hekmatfar, M. (Eds.), Facility Location: Concepts, Models, Algorithms and Case Studies. Physica-Verlag, Heidelberg, pp. 177–191.

Johnson, R., Wichern, D., 2008. Applied Multivariate Statistical Analysis, sixth ed. Prentice-Hall, Upper Saddle River, NJ.

Kariv, O., Hakimi, S., 1979. An algorithmic approach to network location problems I: the p-centers. SIAM J. Appl. Math. 37, 513–538.

Kim, H., Chun, Y., Griffith, D., 2019. Spatial autocorrelation for solving p-median problem. In: Paper Presented At the Annual Meeting of American Association of Geographers. AAG, April 5, 2019, Washington D.C.

Koehn, H., Steinley, D., Brusco, M., 2010. The p-median model as a tool for clustering psychological data. Psychol. Methods 15 (1), 87–95. http://dx.doi.org/10.1037/a0018535.

Levanova, T., Loresh, M., 2004. Algorithms of ant system and simulated annealing for the p-median problem. Autom. Remote Control 65, 431–438.

Ligmann-Zielinska, A., 2017. Spatial optimization. In: Richardson, D., Castree, N., Goodchild, M., Kobayashi, A., Liu, W., Marston, R. (Eds.), The International Encyclopedia of Geography. Wiley, NY, pp. 1–6. http://dx.doi.org/10.1002/9781118786352.wbieg0156.

Little, R., Rubin, D., 2002. Statistical Analysis with Missing Data, second ed. Wiley, NY.

Mateu, J., Müller, W. (Eds.), 2012. Spatio-Temporal Design: Advances in Efficient Data Acquisition. Wiley, Chicester, UK.

McLachlan, G., Krishnan, T., 2008. The EM Algorithm and Extensions, second ed. Wiley, Hoboken.

Mladenović, N., Brimberg, P., Moreno-Pérez, J.A., 2007. The p-median problem: A survey of metaheuristic approaches. European J. Oper. Res. 179 (3), 927–939.

Morrison, D., Jacobson, J., Sewell, E., 2016. Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning. Discrete Optim. 19, 79–102. http://dx.doi.org/10.1016/j.disopt.2016.01.005.

Müller, W., 2007. Collecting Spatial Data, third ed. Springer, Berlin.

Murray, A., Church, R., 1996a. Applying simulated annealing to location-planning models. J. Heuristics 2 (1), 31–53.

Ord, J., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and an application. Geogr. Anal. 27, 286–306.

Overton, S., Stehman, S., 1993. Properties of designs for sampling continuous spatial resources from a triangular grid. Commun. Stat. 22, 251–264.

Pacifico, L., Ludermir, T., 2021. An evaluation of k-means as a local search operator in hybrid memetic group search optimization for data clustering. Nat. Comput. 20, 611–636. http://dx.doi.org/10.1007/s11047-020-09809-z.

ReVelle, C., Eiselt, H., 2005. Location analysis: a synthesis and survey. European J. Oper. Res. 165, 1–19.

Rolland, E., Schilling, D., Current, J., 1996. An efficient tabu search procedure for the p-median problem. European J. Oper. Res. 96, 329–342.

Rosing, K.E., Hillsman, E., Rosing-Vogelaar, H., 1979. A note comparing optimal and heuristic solutions to the p-median problem. Geogr. Anal. 11 (1), 86–89.

Rosing, K., Hodgson, M., 2002. Heuristic concentration for the p-median: an example demonstrating how and why it works. Comput. Oper. Res. 29, 1317–1330.

Rosing, K., ReVelle, C., 1997. Heuristic concentration: two stage solution construction. European J. Oper. Res. 97, 75–86.

Ryan, T., 2009. Modern Regression Analysis. Wiley, New York.

Salhi, S., 2002. Defining tabu list size and aspiration criterion within tabu search methods. Comput. Oper. Res. 29, 67–86.

Salhi, S., Gamal, M., 2003. A genetic algorithm based approach for the uncapacitated continuous location–allocation problem. Ann. Oper. Res. 123 (1–4), 203–222.

Schafer, J., 2000. Analysis of Incomplete Multivariate Data. Chapman and Hall/CRC, Boca Raton, FL.

Scott, A., 1970. Location–allocation systems: a review. Geogr. Anal. 2, 95–119.

Small, C., 1990. A survey of multidimensional medians. Internat. Statist. Rev. 58 (3), 263–277.

Teitz, M., Bart, P., 1968. Heuristic methods for estimating the generalized vertex median of a weighted graph. Oper. Res. 16 (5), 955–961.

Tong, D., Murray, A., 2012. Spatial optimization in geography. Ann. Am. Assoc. Geogr. 102 (6), 1290–1309.

US Centers for Disease Control, 2019. Population health, place, and space, special issue of preventing chronic disease, 16. https://www.cdc.gov/pcd/issues/2019/19_0237.htm.

van der Kloot, W., Spaans, A., Heiser, W., 2005. Instability of hierarchical cluster analysis due to input order of the data: The permuclUSter solution. Psychol. Methods 10 (4), 468–476. http://dx.doi.org/10.1037/1082-989X.10.4.468.

Vardi, Y., Zhang, C.-H., 2000. The multivariate L_1-median and associated data depth. Proc. Natl. Acad. Sci. USA 97, 1423–1426.

Witzgall, C., 1964. Optimal Location of a Central Facility: Mathematical Models and Concepts. NBS Report 8388, U. S. National Bureau of Standards, Washington DC.

Zhang, H., Wang, Z., Yang, B., Chai, J., Wei, C., 2021. Spatial–temporal characteristics of illegal land use and its driving factors in China from 2004 to 2017. Int. J. Environ. Resour. Public Health 18 (1336), http://dx.doi.org/10.3390/ijerph18031336.