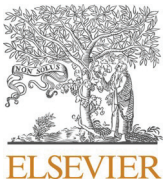
Spatial Statistics 59 (2024) 100810



Contents lists available at ScienceDirect

Spatial Statistics

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

Impacts of spatial imputation on location-allocation problem solutions

Dongeun Kim, Yongwan Chun, Daniel A. Griffith *

School of Economic, Political and Policy Sciences, The University of Texas at Dallas, United States

ARTICLE INFO

Keywords:
Spatial imputation
Spatial autocorrelation
Moran eigenvector spatial filtering
P-median
Location-allocation

ABSTRACT

Georeferenced data often contain missing values, and such missing values can considerably affect spatial modeling. A spatial location model can also suffer from this issue when there are missing values in its geographic distribution of weights. Although general imputation approaches have been developed, one distinguishing fact here is that spatial imputation generally performs better for georeferenced data because it can reflect a fundamental property of those data, that is, spatial autocorrelation or spatial dependency. This paper explores how spatial imputation exploiting spatial autocorrelation can contribute to estimating missing values in a weights surface for location modeling and subsequently improve solutions for spatial optimization, specifically *p*-median problems using a spatially imputed weights surface. This paper examines two spatial imputation methods, ordinary co-kriging and Moran eigenvector spatial filtering. Their results are compared with conventional linear regression, essentially Expectation-Maximization algorithm results for independent observations of Gaussian random variable cases. Simulation experiments show that spatial imputation produces better results for georeferenced data than simply ignoring any missing values and non-spatial imputation, and appropriately imputed values can enhance spatial optimization solutions, regardless of the number of medians, *p*.

1. Introduction

Geospatial modeling often encounters missing values in georeferenced data, ranging anywhere from a few to a considerable number of values. Such an incomplete dataset not only increases uncertainty about the dataset itself (e.g., its probability basis most likely experiences corruption) but also impacts any of its data analysis results (e.g., it effectively experiences a sample size reduction). A naive and simple treatment, for example, ignoring the missing values, can result in a sizeable reduction in the number of observations, causing reliability and precision issues in data analysis. Hence, accurately imputing missing values with quality imputations becomes a critical undertaking in empirical work.

Popular statistical imputation methods, such as Expectation-Maximization (EM), have been utilized in geographical analyses. Griffith (2010) shows a simpler specification of EM in the context of linear regression (LR). However, for spatial datasets, as the history of spatial interpolation and kriging exemplify, spatial information [i.e., spatial autocorrelation (SA)] can help fill missing gaps with more accurate values (i.e., scientifically based educated guesses) and result in improved imputation accuracy and precision in most empirical studies. For example, Griffith and Liau (2021) show that spatial imputation methods produce more accurate imputation results in their application for imputing population counts in Puerto Rico.

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

https://doi.org/10.1016/j.spasta.2024.100810
Received 1 September 2023; Received in revised form 30 December 2023; Accepted 2 January 2024
Available online 4 January 2024

Although the impact of imputation utilization has been widely recognized in spatial data analysis, this issue has been under-investigated in the context of location modeling. Only a few studies discuss the impact of imputations in the context of location-allocation (L-A) solutions, focusing on missing values in the employed weights surface. An L-A problem aims to identify optimal locations of facilities that minimize travel costs between them and demand points (Church et al., 2018). Its description and characterization relate to the following three components: facilities, locations, and demand magnitudes (Scaparra and Scutella, 2001). When a cost is set to the distance between a facility and a demand point, the optimal solution of an L-A problem minimizes the sum of weighted distances from each demand point to its allocated facility. Here, weights reflect the geographically varying sizes or importance levels of demands (Azarmand and Neishabouri, 2009). Meanwhile, varying weights can form a spatial pattern with their geographic distribution across demand locations (i.e., they are spatially autocorrelated); this is typical of georeferenced data.

Robin (1988) argues that a proper strategy is necessary to deal with missing weights in an L-A problem. Griffith (2003) shows that ignoring these missing values by assigning zero weights to their affiliated demand points results in an exclusion of those demand points from the optimization exercise, and, consequently, can substantially impact an L-A solution. Only slightly better would be their inclusion with conventional EM imputations, which would like spatial pattern, risking a substitution of predicted values that become local outliers. Griffith (1997; 2003) reports improved solutions for 1- and 2-median L-A problems using spatial imputation for missing weight surface values. Also, Griffith et al. (2022) further discuss how imputation using spatial statistics can contribute to enhancing L-A solutions. One remaining problem is to address this issue for $p > 2$.

This paper explores how spatial imputation, which exploits SA in geospatial datasets, can contribute to estimating missing values in a weighted surface and improve solutions to spatial optimization problems, specifically, p-median problems for large values of p. It compares the performances of three imputation methods: LR, ordinary co-kriging (OCK), and Moran eigenvector spatial filtering (MESF). Specifically, this paper investigates the following three research questions using simulation experiments: (1) do spatial imputation techniques produce better substitute values than non-spatial methods, (2) do spatial optimization solutions improve with imputed weights surfaces, and (3) do different imputation methods result in substantially different L-A solutions?

2. Literature review

L-A problems were first proposed by Cooper (1963) and have been further developed via various model specifications and algorithms. Although it is specified with Euclidean distances in a two-dimensional (2D) space, it applies to a network space or other hybrid geographic landscapes. Any correctly specified L-A model's exact solution is obtainable (although it might not be unique), given enough time and computer memory. Still, heuristic and meta-heuristic methods, such as simulated annealing, Tabu search, genetic algorithm, variable neighborhood search, and ant colony algorithm, are popularly utilized for large problems (Azarmand and Neishabouri, 2009) because of these time and/or memory restrictions. These strategies are commonly employed for planning purposes to locate facilities such as multilocational healthcare outlets, hubs in intermodal logistic networks, and systems of disaster evacuation shelters. For example, Mestre et al. (2015) propose two L-A models for reorganizing hospital networking systems to improve geographical access and minimize costs. Ishfaq and Sox (2011) present a model based on the multiple-allocation p-hub median approach, highlighting the impact of the intermodal connectivity cost at a hub. Zhao et al. (2017) introduce an integrated L-A model for emergency shelters considering the coverage radius of each shelter (capacity) and its cost.

Spatial imputation utilizes spatial information latent in data. Specifically, SA can provide attribute values synchronizing tendency among observations, which can help improve accuracy in imputation results. It has been used in a wide range of incomplete data applications, such as those in spatial health, crop yield, and climate recordings. For example, Baker et al. (2014) employ the spatial imputation method for health survey data using spatial as well as non-spatial correlation among covariates. Lokupitiya et al. (2006) estimate missing crop yield data, considering that these agricultural data tend to be spatially autocorrelated. Qin et al. (2021) focus on the intrinsic properties of climate data, which show a strong seasonality and spatial correlation suitable for aiding the imputing of missing values. These examples empirically emphasize that spatial imputation methods are more effective than their non-spatial counterparts for geographical datasets when SA is present.

Popular spatial imputation methods include kriging, a spatialized EM algorithm, nearest neighbor averaging, mean substitution, regression prediction, spline interpolation, and sundry spectral approaches. Griffith and Liau (2021) conducted a comparative study to investigate the performances of spatial imputation methods using simulation experiments, specifically, the three spatial statistics methods of kriging, spatial autoregression, and MESF. They report that kriging is best when the percentages of missing values are extremely high (e.g., 75 % or greater). Amitha et al. (2021) find that stochastic regression imputation results introduce optimal replacement values if a variable with missing values has either a high or a low level of SA. On the one hand, when the data have a high SA level, stochastic regression results using the spatial lag, spatial Durbin, and spatial Durbin error models are highly accurate. On the other hand, when the level of SA is low, regression imputation with the spatial lag X model also performs well.

Although spatial imputation has been utilized in the literature, its impact on the L-A problem has not been extensively and systematically investigated. Griffith (1997; 2003) explores the impact of missing georeferenced data imputations on the single-facility and 2-median L-A solution cases. These studies show that, for missing geographic data, spatial statistical models are more effective than ignoring missing values when solving p-median problems. However, because they are limited to small size problems with $p = 1$ or 2, they encourage investigations for large p values.

3. Methodology

This paper investigates the impacts of spatial imputation on L-A solutions, focusing on the p-median problem where p is large. The

formulation of this problem is with a weights surface; it can be solved with the proprietary computer software IBM ILOG CPlex. A simulation experiment was conducted to evaluate the quality of p-median solutions with, first, missing and, second, imputed weights. The three aforementioned imputation methods were utilized and compared. To reiterate, the two spatial imputations are OCK and MESF; the other is LR, a non-spatial method.

3.1. The p-median problem

The problem aims to locate p facilities in space to minimize the sum of the weighted distances between each of n demand points and their respective single allocated facility (Church et al., 2018; Daskin and Maass, 2015; ReVelle and Swain, 1970). It can be formulated mathematically as follows:

$$minimize \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij} w_i \sqrt{\left(u_i - U_j\right)^2 + \left(v_i - V_j\right)^2} \tag{1}$$

subject to:

$$\sum_{j=1}^{n} x_{ij} = 1 \quad \forall i \tag{2}$$

$$\sum_{j=1}^{n} x_{jj} = p \tag{3}$$

$$x_{ij} - x_{jj} \leq 0 \ \forall i,j \ and \ i \neq j \tag{4}$$

$$x_{ij} = \{0,1\} \ \forall i,j \tag{5}$$

The subscript notation i and j are the indices for respectively naming demand points and facilities, n denotes the number of demand points, x_{ij} is a decision variable (1 if demand i is allocated to facility j, or 0 otherwise), w_i denotes a weight quantifying demand at point i, $(u_i,\ v_i)$ is the Cartesian coordinates of demand point i, $(U_j,\ V_j)$ is the coordinates of facility location j, $\sqrt{(u_i - U_j)^2 + (v_i - V_j)^2}$ is Euclidean distance between demand point i and facility location j, and p is the number of facilities to be located. The objective function (1) minimizes the total demand-weighted distance. Constraint (2) ensures that each demand point i is assigned to a facility j. Constraint (3) ensures that exactly p facilities are required to be sited. Constraint (4) states that demand point i can be assigned to only one facility j that has been sited. Integer and binary restrictions are shown in constraint (5).

The weights, w_i, is the target of imputation because its missing values impact p-median solutions. For example, if missing values are replaced with zero, those corresponding demand points become ignored in a solution. In the simulation experiments for this paper, a portion of weights will be suppressed and imputed with the three different imputation methods.

3.2. Non-spatial and spatial methods

The EM algorithm furnishes one of the most popular imputation methods. Griffith (2010) demonstrates that it can be formulated with LR for Gaussian distributions. This specification introduces dummy variables for records with missing values and inserts zero in the response variable for those records. The coefficients of the dummy variables are EM-imputed values. This formulation can be expressed as follows:

$$\begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} = \begin{pmatrix} \mathbf{1}_o \\ \mathbf{1}_m \end{pmatrix} \boldsymbol{\beta}_0 + \begin{pmatrix} \mathbf{X}_o \\ \mathbf{X}_m \end{pmatrix} \boldsymbol{\beta}_X + \begin{pmatrix} \mathbf{0}_o \\ -\mathbf{I}_m \end{pmatrix} (\mathbf{Y}_m) \ + \boldsymbol{\varepsilon}. \tag{6}$$

Here, the subscript o indicates records with observed values, whereas the subscript m indicates missing value records and the response variable, and covariates are sorted in the order of observed and missing values for convenience. $\mathbf{Y}_o$ is the n_o-by-1 response variable vector, where n_o is the number of records with observed values; that is, $n_o = n - n_m$ where n is the total number of records, and n_m is the number of records with missing values. $\mathbf{0}_m$ is the n_m-by-1 vector of zeros. Meanwhile, $\mathbf{1}_o$ indicates the part of the intercept term vector of ones for records with observed response values, and $\mathbf{1}_m$ indicates the part of that vector for records with missing response values. Similarly, $\mathbf{X}_o$ is the matrix of covariate values for the records with observed response values, and $\mathbf{X}_m$ is the matrix of covariate values for the records with missing response values. $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_X$ denote regression coefficients for the intercept and p covariates, respectively. $\mathbf{0}_o$ is an n_o-by-n_m matrix of zeros, and $\mathbf{I}_m$ is an n_m-by-n_m identity matrix. $\mathbf{Y}_m$ is the n_m-by-1 vector of coefficients for the dummy variables that are imputed values for missing response values. $\boldsymbol{\varepsilon}$ is an n-by-1 vector of independent and identically distributed (i.e., iid) normal random errors (Griffith et al., 2022).

This LR specification extends to account for SA among observations using MESF, which introduces eigenvectors extracted from a transformed spatial weights matrix $\mathbf{C}$ as additional covariates into a regression model (Griffith et al., 2019; Griffith and Chun 2021). These eigenvectors, extracted from **MCM,** where $\mathbf{M} = \mathbf{I} - \mathbf{1} \cdot \mathbf{1}^T/n$, portray uncorrelated and orthogonal spatial map patterns, collectively representing a range of SA from extreme positive to extreme negative. A set of k judiciously selected eigenvectors, $\mathbf{E}_k$,

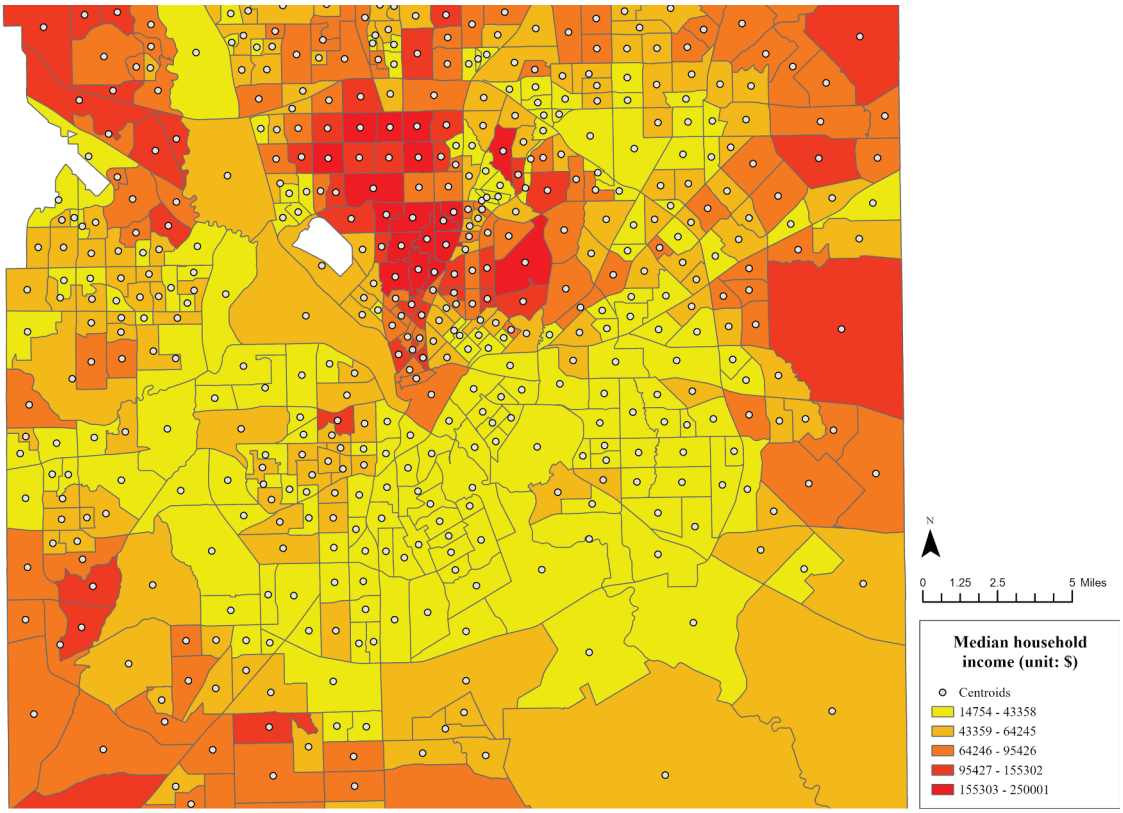


Fig. 1. Median household income (unit: US dollar) for the census tracts in Dallas County, Texas.

Table 1
The RMSE and MAE values of the OCK, MESF, and LR.

	method	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
RMSE	OCK	5.193	7.580	9.417	10.948	12.346	13.721	14.987	16.401	17.620	18.838
	MESF	4.319	6.422	8.036	9.432	10.672	11.950	13.124	14.355	15.634	16.830
	LR	5.294	7.777	9.682	11.059	12.447	13.684	14.800	16.039	16.937	17.759
MAE	OCK	0.734	1.499	2.252	3.021	3.825	4.635	5.467	6.389	7.269	8.232
	MESF	0.628	1.297	1.965	2.652	3.382	4.117	4.890	5.709	6.582	7.491
	LR	0.758	1.548	2.325	3.067	3.856	4.618	5.381	6.213	6.964	7.725

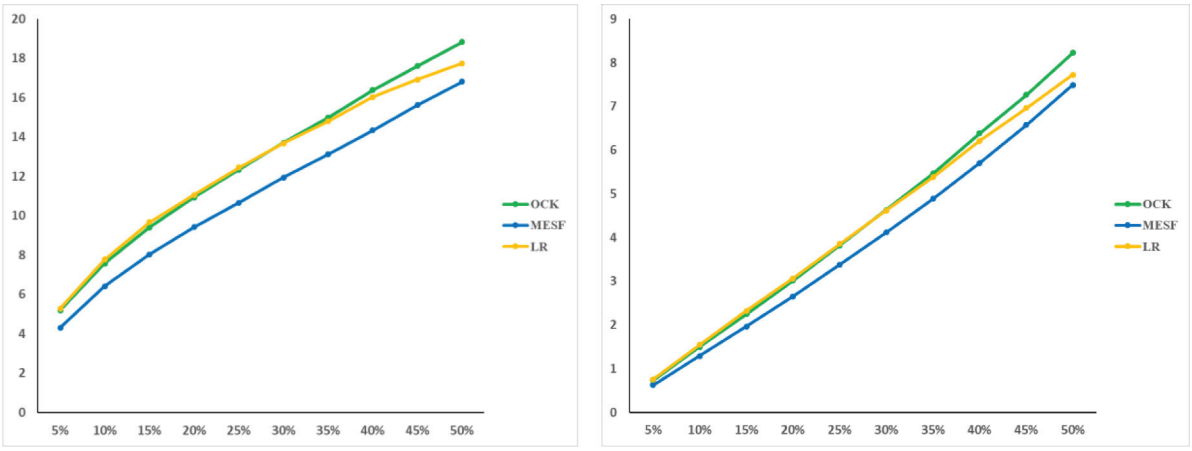


Fig. 2. The RMSE and MAE values for the three imputation methods.

captures spatially autocorrelation components that are not explained in a regression model. Chun et al. (2016) discuss how these k eigenvectors can be selected from the full set of n eigenvectors (k is much smaller than n). The MESF-based imputation can be expressed as follows (Griffith et al., 2022):

$$\begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} = \begin{pmatrix} \mathbf{1}_o \\ \mathbf{1}_m \end{pmatrix} \boldsymbol{\beta}_0 + \begin{pmatrix} \mathbf{X}_o \\ \mathbf{X}_m \end{pmatrix} \boldsymbol{\beta}_X + \begin{pmatrix} \mathbf{0}_o \\ -\mathbf{I}_m \end{pmatrix} (\mathbf{Y}_m) + \sum_{j=1}^{k} \begin{pmatrix} \mathbf{E}_{o,\,j} \\ \mathbf{E}_{m,\,j} \end{pmatrix} \boldsymbol{\beta}_{\mathbf{E}_k} + \boldsymbol{\varepsilon}, \tag{7}$$

where $\mathbf{E}_{o,j}$ is the part of eigenvector j associated with the observed values, and $\mathbf{E}_{m,j}$ is the part of eigenvector j associated with the missing values. $\boldsymbol{\beta}_{\mathbf{E}_k}$ is the k-by-1 regression coefficient vector for the eigenvector $\mathbf{E}_k$ used to construct an eigenvector spatial filter.

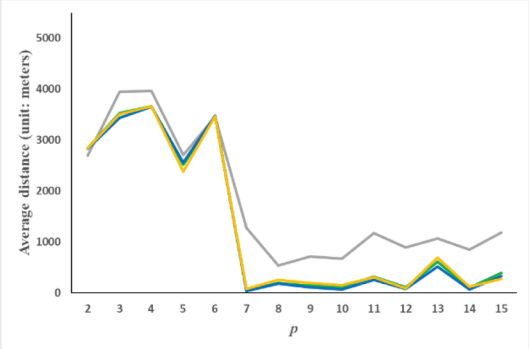
OCK also is utilized for imputation purposes, and its performance is then compared with those of the other two methods. Kriging is a well-known technique for spatial interpolation that predicts unobserved values of a random spatial process. It imputes missing values by exploiting redundant information and exploding georeferenced datasets (Griffith and Liau, 2021). Co-kriging is the bivariate extension of kriging, allowing the use of a secondary data source covariate to complement observed primary data. The effectiveness of co-kriging relies on the pattern of missing data and the strength of the relationships between the response variable and included covariate. OCK is a multivariate extension of ordinary kriging, known as the best linear unbiased estimator (Bae et al., 2018).

3.3. Simulation experiments

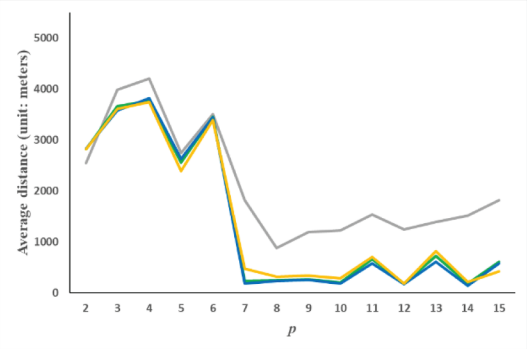
In this study, the demographic dataset based on the 2018 United States American Community Survey (ACS) provided by Maptitude Mapping Software is employed for the specific simulation experiments that explore the impacts of imputations on p-median problem solutions. This p-median problem is specified with the centroids of the 529 census tracts in Dallas County, Texas (see Fig. 1) as the demand points. Using the median household income of the tracts as weights, the locations for p facilities are identified for these 529 centroids. These p-median problems are solved with the IBM ILOG CPlex version 20.1.0.0.

The simulation experiment was conducted with a range of missing value percentages, from 5% to 50%, in 5% increments. Randomly selected values were suppressed and then imputed with the three methods: LR, MESF, and OCK. Population densities for age 25 or older were used as a covariate for the imputations. Next, the p-median problems were solved for various p values, from $p = 2$ to $p = 15$. The imputation computations and solving of the p-median problem were replicated 1000 times. In addition, these same p-median problems were solved by replacing their suppressed values with zero. This replacement is equivalent to ignoring the missing values in a p-median problem. Accordingly, its results are expected to be worse than those with the three imputation approaches, serving as a worst-case scenario. Expected deviations should be somewhat mitigated on average by the spatial randomness of the suppressions.

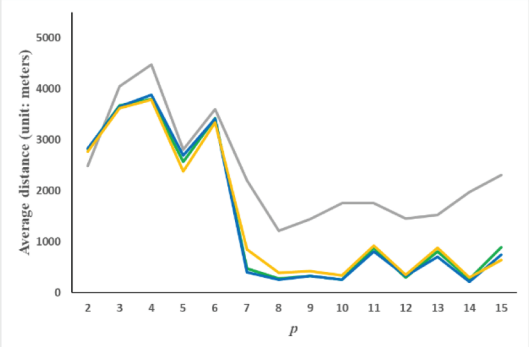
a) 5% missing values



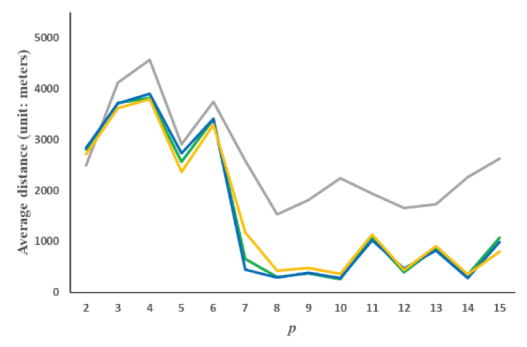
b) 10% missing values



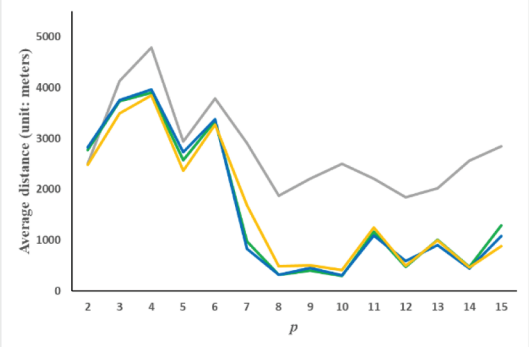
c) 15% missing values



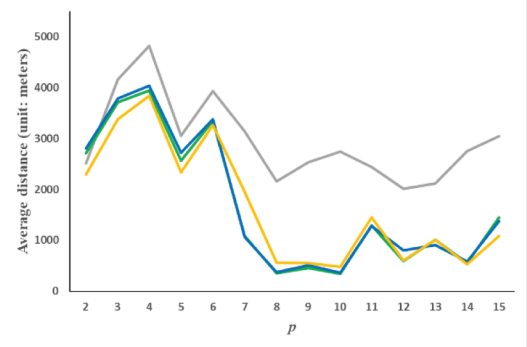
d) 20% missing values



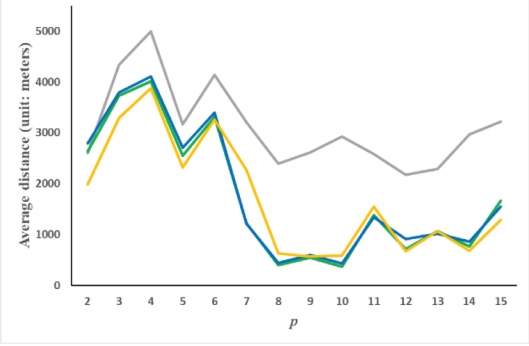
e) 25% missing values



f) 30% missing values



g) 35% missing values



h) 40% missing values

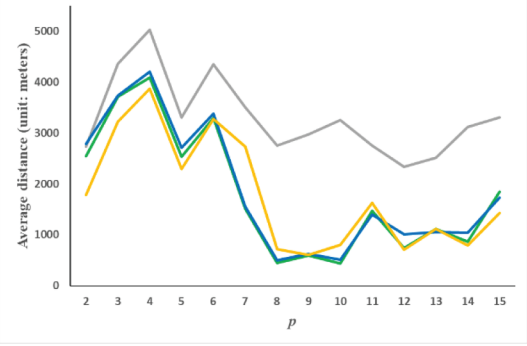


Fig. 3. The average distances between the complete and estimated dataset optimal p solution points using zeros in missing values, OCK, MESF, and LR.

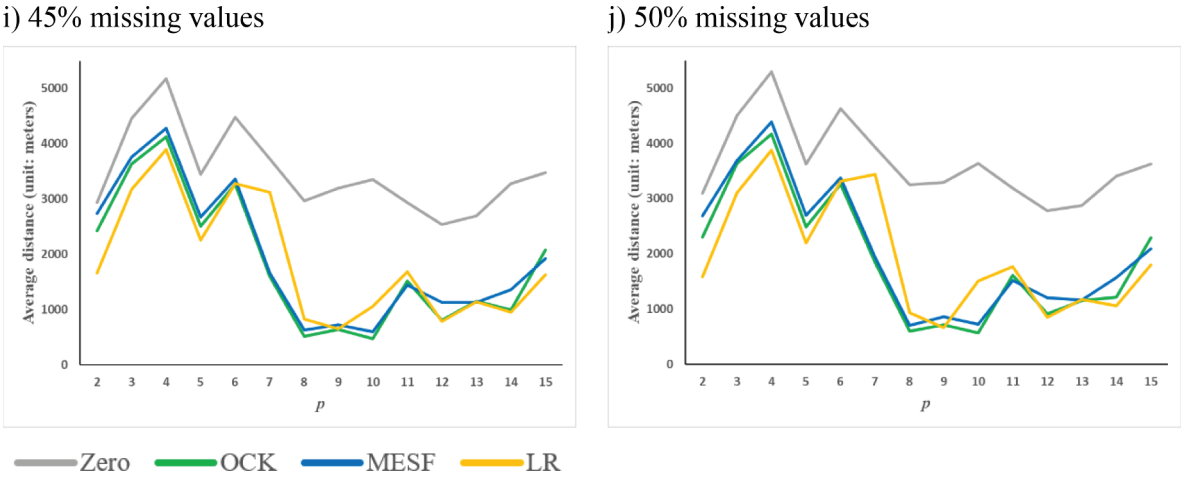


Fig. 3. (*continued*).

The performance of the imputations is measured with both root mean square error (RMSE) and mean absolute error (MAE). These quantities have been used frequently to evaluate prediction methods in many kinds of research (Armstrong and Collopy, 1992; Hodson, 2022). The following formulas calculate them:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(A_i - F_i)^2}{n}} \tag{8}$$

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |A_i - F_i| \tag{9}$$

A_i is the actual value, and F_i is the imputed values.

4. Results

Table 1 and Fig. 2 present the RMSE and MAE results, which are mean values across 1000 simulation replications. First, as expected, the RMSE and the MAE values tend to increase as missing percentages increase for all three methods. Among the three methods, the MESF-based approach produces the lowest errors for all missing percentages. Although OCK and LR produce comparable results, the OCK results are slightly better for small missing values percentages (5–25%), and the LR results are slightly better for large missing values percentages (30–50%). The gap between their RMSE and MAE values gets slightly larger for 35% or greater missing values percentages.

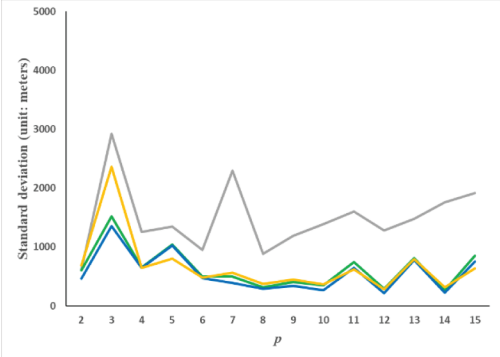
As for imputation accuracy, the MESF model produces better imputation results than the others. Unlike the findings that spatial imputation performs better than non-spatial imputation in the literature (e.g., Griffith and Liau, 2021), the RMSE and the MAE indices show a mixed result. LR performs better than OCK when the missing percentage is 30% or greater.

Fig. 3 shows the quality of the p-median results with imputations. Specifically, it presents mean distances of displacements for p facilities with imputations for missing values from the optimal facility locations without any missing values. First, the results show that the displacements for the optimal locations are greater when missing values are simply replaced with zero than with any of the imputations. The only exceptions appear for some $p = 2$ cases, for which the simple replacement with zero produces comparable results with the other imputation cases; this outcome could be because the missing map pattern is random, which helps minimize, on average, optimal location impacts for weights becoming zero. This finding indicates that the imputation of missing weights improves p-median solutions. Also, the gap between the replacement with zero and the other imputation cases tends to be large when the missing rates are high. The gap is generally larger, with a 50 % missing rate (Fig. 3e), than the other missing rates (Figs. 3a-d).

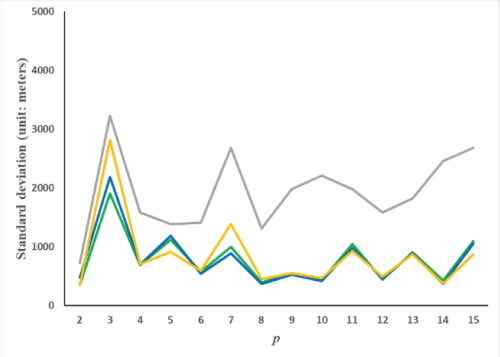
Second, the three imputation methods' average displacements with imputed values are comparable overall. But still, some differences are observable. For 5% missing values, when p is 6 or less, the displacements with LR are slightly smaller than for the other imputation approaches. In contrast, when p gets larger, the MESF results have a smaller value. The results for 10% missing values show the same pattern: the imputation with LR leads to a slightly better result for small p values (from two to six), whereas the MESF results have shorter displacements for large p values. The results for the other missing rates (15–35%) have the same pattern, although some fluctuations are detectable. A similar pattern persists for the 40–50% missing rates, although the results with LR tend to become a little more deviant.

Third, the impacts of missing rates on p-median solutions tend to be large for small p values. The mean displacements are larger for $p = 2$, 3, 4, 5, or 6 than for larger p values across all missing rates. The pattern is clear for 5 and 10 % missing rates. For these two, the

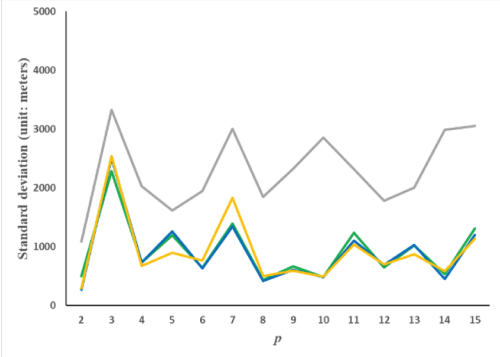
a) 5% missing values



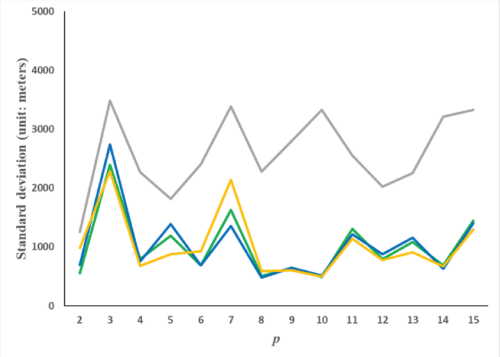
b) 10% missing values



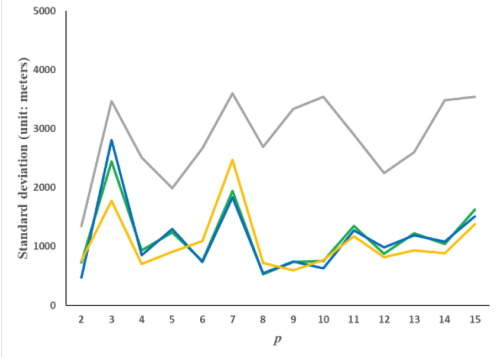
c) 15% missing values



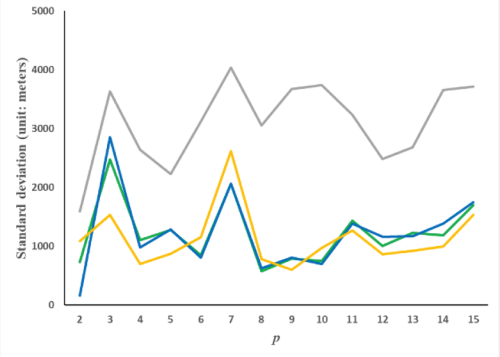
d) 20% missing values



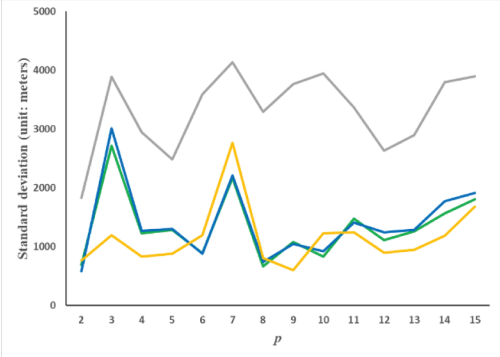
e) 25% missing values



f) 30% missing values



g) 35% missing values



h) 40% missing values

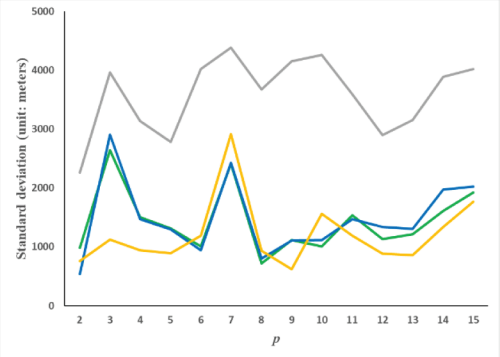


Fig. 4. The standard deviations of the *p*-median solutions around their geometric mean centers using zeros in missing values, OCK, MESF, and LR.

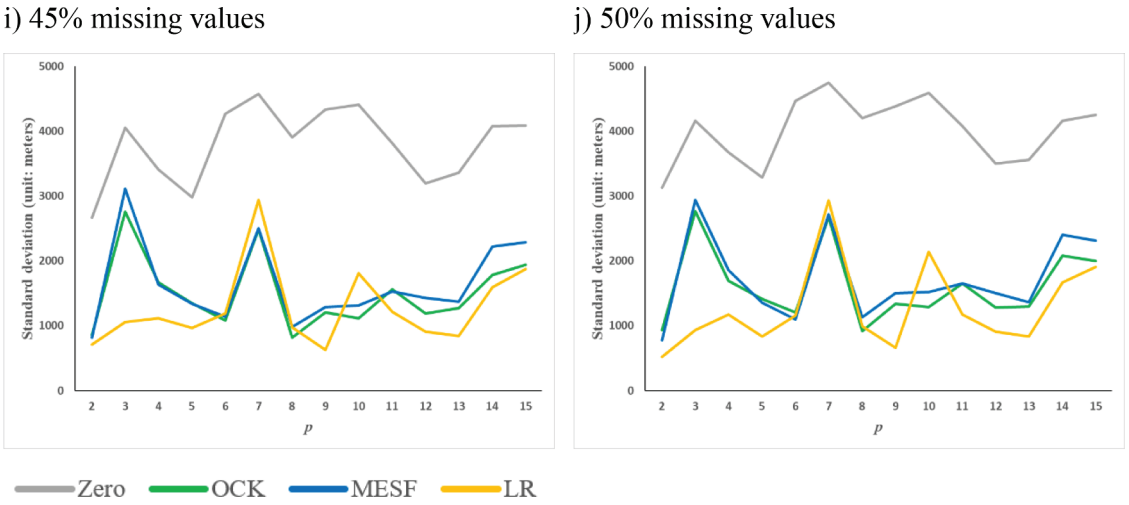


Fig. 4. (*continued*).

mean displacement distances for $p = 2, 3, 4, 5$, or 6 are much larger than the other p values. Differences between the small and large p groups diminish as missing rates get larger. The 40% and 50% results show a much flatter trendline slope.

The graphs in Fig. 4 show the p-median solutions' standard deviations around their geometric mean centers, calculated with 1000 solutions for each p cluster. This process was repeated for each imputation method and percentage of missing values. It presents how concentrated or dispersed solutions are, which may indicate the precision or consistency of the solutions, which tend to be more dispersed when missing values are replaced with zero than any of the corresponding three imputations. Overall, the p-median results with MESF and OCK are similar across all p values and missing rates, and the LR results deviate from their results. Specifically, the p-median solutions with LR results tend to be more concentrated than the other results when the missing rate is 25% or greater. This may be explained by the fact that imputation with the LR approach generally tends to be done with a global mean. Hence, when a missing rate is high, a large portion of values tend to be similar after the imputation, so the p-median solutions are more concentrated in space.

Fig. 5 presents the changes in the objective function values and fitted curves using nonlinear regression only for selected p values presented as a ratio to their optimal solutions with no missing values (red). As expected, the replacements with zero (grey) lead to a large deviation from optimality with no missing values. Because zero weights lead to the exclusion of their demand points from the model specification, their objective function values become smaller. This objective function decrease continues as more demand points have a zero weight (the missing rate increases). When the missing rate is 50%, the objective function value is less than 50% of that with no missing values. The objective function values are underestimated even with the three imputation approaches. However, their results are much closer to the ones with no missing values. Fig. 6 presents the three imputations only to clearly show differences among these approaches. This pattern is consistent for all p values. The MESF imputation results have closer objective function values to those for the complete data across all p values and missing rates. This finding indicates that the MESF approach produces the best results among the imputation approaches. Table 2 presents the pseudo-R^2 measure, indicating the goodness of fit for nonlinear regression models. All the values demonstrate a strong level of agreement, consistently maintaining a high agreement level of around 99%.

5. Conclusions and discussion

This paper investigates the impacts of demand point weights imputation on the quality of p-median solutions. Simulation experimental results show that spatial imputation approaches produce more accurate solutions than a non-spatial imputation approach (specifically, LR) for georeferenced data. It also shows that the MESF approach produces better results than OCK. It furnishes a convenient methodology with its flexible structure to incorporate spatial patterns into imputation.

Evidence accumulated for this paper also reveals that better imputation results contribute to increased accuracy of p-median solutions. That is, appropriately imputed values can help improve spatial optimization solutions. Specifically, the MESF approach consistently yields better results than the other two approaches, LR and OCK. That is, the extended expectation-maximization using MESF, which accounts for spatial autocorrelation, consistently lead to more accurate location-allocation solutions.[1] Generally, spatial optimization problem solutions obtained with spatial imputation values are better than those obtained with non-spatial methods. The objective function values using spatial imputation are closer to their optimal solution counterparts computed with a complete dataset than those obtained with non-spatial imputation. In addition, imputation-based p-median locations tend to be closer to their affiliated complete data optimal solution. Another interesting outcome is that the LR approach generally produces more accurate outcomes than

[1] Often MESF successfully adjusts for missing variables (Griffith and Chun, 2016).

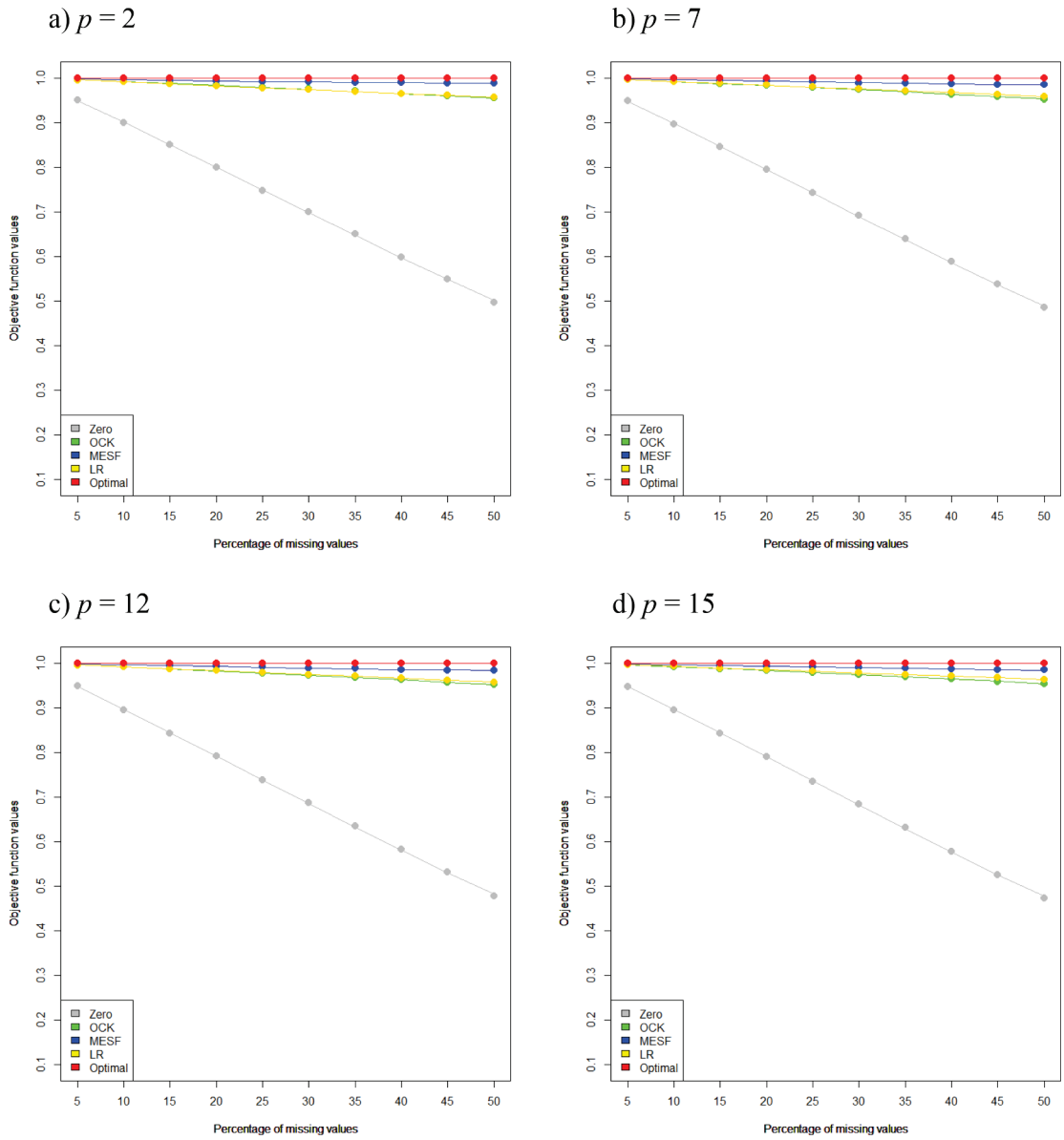


Fig. 5. The objective function values and fitted curves of p-median solutions with imputations for selected p values (including zero weights).

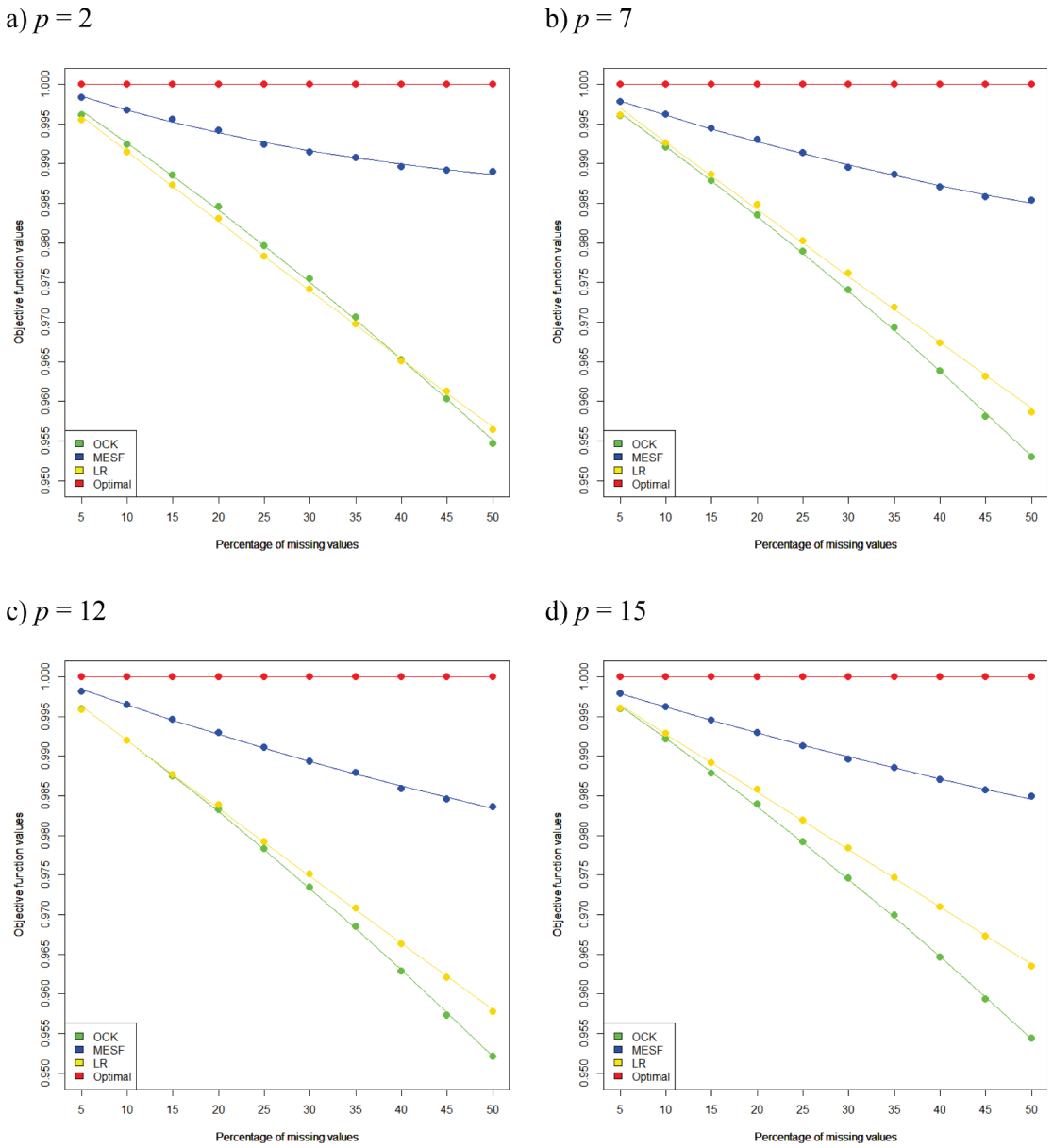


Fig. 6. The objective function values and fitted curves of p-median solutions with imputations for selected p values (excluding zero weights).

Table 2
The pseudo-R^2 values of the goodness of fit for nonlinear regression models.

method	$p = 2$	$p = 7$	$p = 12$	$p = 15$
Zero	0.9999	0.9999	0.9999	0.9999
OCK	0.9997	0.9998	0.9999	0.9999
MESF	0.9969	0.9985	0.9992	0.9992
LR	0.9998	0.9994	0.9997	0.9998

the OCK approach when missing rates are high.

The research summarized in this paper warrants future investigations involving more extensive simulation experiments. Certainly, imputator formulations can be extended with a multivariate specification approach. In addition, experiments with other imputation methods, such as spatial autoregression, can broaden the comparative conclusions. Second, an experiment with irregular geographic landscape shapes would be insightful. In this paper, the study area is limited to a square. An investigation with irregular shapes would

help to generalize implications. Third, whereas the experiment for this paper used random sampling to choose observations to be suppressed as missing values, other sampling designs merit examination, such as geographically stratified random sampling to spread missing values more evenly over a landscape, missing values clusters, or attribute value magnitude dependent sampling. Spatial imputation can be affected when a majority of spatial neighbors of a missing value also have missing values because imputations become an intensive function of other imputations. Furthermore, spatial imputation may perform better when missing values are spatially dispersed rather than clustered. The question it spawns asks whether this property transfers to *p*-median solutions.

Acknowledgments

This research was supported by the U.S. National Science Foundation, 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

Amitha, P., Binu, V.S., Seena, B., 2021. Estimation of missing values in aggregate level spatial data. Clin. Epidemiol. Glob. Health 9, 304–309. https://doi.org/10.1016/j.cegh.2020.10.003.

Armstrong, J.S., Collopy, F., 1992. Error measures for generalizing about forecasting methods: empirical comparisons. Int. J. Forecast 8 (1), 69–80. https://doi.org/10.1016/0169-2070(92)90008-W.

Azarmand, Z., Neishabouri, E., 2009. Location allocation problem. Farahani R. Z., Hekmatfar, M. (Eds.). Facility Location: Concepts, Models, Algorithms and Case Studies. Springer, Heidelberg, pp. 93–109.

Bae, B., Kim, H., Lim, H., Liu, Y., Han, L.D., Freeze, P.B., 2018. Missing data imputation for traffic flow speed using spatio-temporal cokriging. Transp. Res. C Emerg. Technol. 88, 124–139. https://doi.org/10.1016/j.trc.2018.01.015.

Baker, J., White, N., Mengersen, K., 2014. Missing in space: an evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. Int. J. Health Geogr. 13, 1–13. https://doi.org/10.1186/1476-072X-13-47.

Chun, Y., Griffith, D.A., Lee, M., Sinha, P., 2016. Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. J. Geogr. Syst. 18, 67–85. https://doi.org/10.1007/s10109-015-0225-3.

Church, R.L., Murray, A., Church, R.L., Murray, A., 2018. Classic beginnings. Church, R. L., Murray, A. (Eds.). Location Covering Models: History, Applications and Advancements. Springer, Berlin, pp. 23–47.

Cooper, L., 1963. Location-allocation problems. Oper. Res. 11 (3), 331–343. https://doi.org/10.1287/opre.11.3.331.

Daskin, M.S., Maass, K.L., 2015. The *p*-median problem. Laporte, G., Nickel, S., Saldanha da Gama, F. (Eds.). Location Science. Springer International Publishing, pp. 21–45.

Griffith, D.A., 1997. Using estimated missing spatial data in obtaining single facility location-allocation solutions. L'Espace Géogr. 26 (2), 173–182.

Griffith, D.A., 2003. Using estimated missing spatial data with the 2-median model. Ann. Oper. Res. 122 (1), 233–247. https://doi.org/10.1023/A:1026106825798.

Griffith D.A., 2010. Some simplifications for the expectation-maximization (EM) algorithm: the linear regression model case. InterStat, 23.

Griffith, D.A., Chun, Y., 2016. Evaluating eigenvector spatial filter corrections for omitted georeferenced variables. Econometrics 4 (2), 29. https://doi.org/10.3390/econometrics4020029.

Griffith, D.A., Chun, Y., 2021. Spatial autocorrelation and Moran eigenvector spatial filtering. Fischer, M., Nijkamp, P. (Eds.). Handbook of Regional Science, 2nd ed. Springer-Verlag, Berlin, pp. 1863–1894.

Griffith, D.A., Chun, Y., Kim, H., 2022. Spatial autocorrelation informed approaches to solving location–allocation problems. Spat Stat. 50, 100612. https://doi.org/10.1016/j.spasta.2022.100612.

Griffith, D.A., Chun, Y., Li, B., 2019. Spatial Regression Analysis Using Eigenvector Spatial Filtering. Academic press.

Griffith, D.A., Liau, Y.T., 2021. Imputed spatial data: cautions arising from response and covariate imputation measurement error. Spat Stat. 42, 100419. https://doi.org/10.1016/j.spasta.2020.100419.

Hodson, T.O., 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci. Model. Dev. 15 (14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022.

Ishfaq, R., Sox, C.R., 2011. Hub location–allocation in intermodal logistic networks. Eur. J. Oper. Res. 210 (2), 213–230. https://doi.org/10.1016/j.ejor.2010.09.017.

Lokupitiya, R.S., Lokupitiya, E., Paustian, K., 2006. Comparison of missing value imputation methods for crop yield data. Environmetrics Off. J. Int. Environmetrics Soc. 17 (4), 339–349. https://doi.org/10.1002/env.773.

Mestre, A.M., Oliveira, M.D., Barbosa-Póvoa, A.P., 2015. Location–allocation approaches for hospital network planning under uncertainty. Eur. J. Oper. Res. 240 (3), 791–806. https://doi.org/10.1016/j.ejor.2014.07.024.

Qin, Y., Ren, G., Zhang, P., Wu, L., Wen, K., 2021. An imputation method for the climatic data with strong seasonality and spatial correlation. Theor. Appl. Climatol. 144, 203–213. https://doi.org/10.1007/s00704-021-03537-9.

ReVelle, C.S., Swain, R.W., 1970. Central facilities location. Geogr. Anal. 2 (1), 30–42. https://doi.org/10.1111/j.1538-4632.1970.tb00142.x.

Rubin, D.B., 1988. An overview of multiple imputation. In: Proceedings of the Survey Research Methods Section of the American Statistical Association, 79, p. 84.

Scaparra, M.P., Scutella, M.G., 2001. Facilities, locations, customers: building blocks of location models. In: A survey, Technical Report del Dipartimento di Informatica, No TR-01-182001. IT: Università di Pisa, Pisa.

Zhao, L., Li, H., Sun, Y., Huang, R., Hu, Q., Wang, J., Gao, F., 2017. Planning emergency shelters for urban disaster resilience: an integrated location-allocation modeling approach. Sustainability 9 (11), 2098. https://doi.org/10.3390/su9112098.