Missing Locations, Missing Ties: Spatial Bias in Kinship Networks

Maryam Torkashvand [1]* and Caglar Koylu [1]

[1] School of Earth, Environment and Sustainability, University of Iowa, Iowa City, IA
* maryam-torkashvand@uiowa.edu

Keywords: family dispersion, family tree, imputation, missing data, spatial bias

Abstract

Linking individuals through kinship ties and locations, using spatiotemporal information in family trees, builds a spatial-social structure for examining demographic processes such as migration and family dispersion. Missing location attributes for available records are common in genealogies, which creates a discrepancy between the full set of social ties and the subset that can be spatially embedded. This study examines how incompleteness in birthplace attributes for US-born sibling sets affects family spatial dispersion and the direction of this effect under different treatments. We compute a spatial dispersion index (SDI) for sibship subnetworks embedded in their birth counties, under two filtering strategies and two imputation models. Across every decade from 1790 to 1910, restricting sibling sets to complete cases reduces the median SDI by 6 to 13 km. Conversely, imputing missing locations based on family spatial and temporal footprints increases SDI while keeping the long-term dispersion trends observed in incomplete data.

Introduction

Population-scale family trees are built from multi-source genealogical records (Kaplanis et al., 2018; Koylu et al., 2021), and carry both record missingness and attribute missingness for an observed record. The representativeness of genealogical datasets assesses the missingness from the first type of missingness and has been well documented through benchmarking aggregated distribution of external population (Li et al., 2013; Zhang & Zhu, 2018), linkage to census records (Koylu et al., 2025), and demographic microsimulation (Calderón-Bernal et al., 2025). Incompleteness arising from attribute missingness, however, is a different cartographic and analytical problem (Robinson, 2019). In a spatial-social network, aspatial ties connect places through their members' locations (Andris, 2016; Ye & Andris, 2021), and absence of location attribute from the observed network introduces new biases into the underlying spatial structure.

Colasurdo and Omenti (2024) showed that attribute missingness for an individual in family trees correlates with the attribute completeness of their kin, particularly parents and siblings. This suggests that location attribute missingness is not missing completely at random (MCAR), and introduces biases into spatial estimates beyond simply blurring them (Bennett et al., 1984; Rubin, 1976; Sutton et al., 2026). Imputation strategies predict the patterns of complete data under spatial and demographic assumptions (Curriero et al., 2010; Sutton et al., 2026; Zahmatkesh & Zech, 2026). However, the magnitude and direction of the changes these methods introduce to specific metrics in spatial-social networks remain understudied. We examine the pattern and non-

randomness of missing birthplace information in a population-scale family tree from 1790 to 1910. Defining a spatial dispersion index (SDI) for sibling subnetworks, we ask whether birthplace missingness is independent of the spatial dispersion of the affected families, and whether applied curation strategies shift the measured dispersion significantly enough to change conclusions about family dispersion in the nineteenth-century United States.

Data and Methods

We extract sibships from a family tree of approximately 50 million individuals connected through parent-child and marriage links, assembled by cleaning and integrating crowdsourced genealogical profiles (Koylu et al., 2021). The birthplaces of 17 million US-born individuals were geocoded to the finest available spatial resolution, with 73% to the county level, 26% to the state level, and 1% unresolved.

A sibship set S has n = |S| members, where $L = \{l_1, l_2, \ldots, l_m\}$ denotes the set of unique birth locations within S, where $m \leq n$. Each location l_i represents a node in the sibship subgraph, and the node weight c_i is the number of siblings in S born at location l_i, so that $\sum_{i=1}^{m} c_i = n$. Edge between (l_i, l_j) is weighted by the share of sibling pairs spanning them:

$$w_{ij} = \frac{c_i \cdot c_j}{\binom{n}{2}} = \frac{c_i \cdot c_j}{\frac{n(n-1)}{2}}, \quad i \neq j$$

where the denominator is the total number of pairs in the set, so that a large family does not dominate a small one. Each sibship forms a weighted undirected subgraph $G_S = (L_S, W_S)$ (Figure 1). Aggregating over the collection $\mathcal{S}$, with

$$C_i = \sum_{S \in \mathcal{S}} c_i^{(S)} \text{ and } W_{ij} = \sum_{S \in \mathcal{S}} w_{ij}^{(S)},$$

yields a spatial interaction network $G = (L, W)$ in which nodes are geographic units and edge weights record how often two units are connected through sibling births.

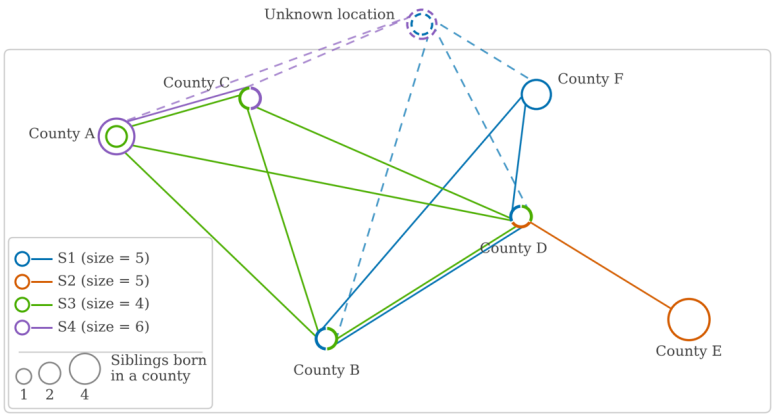


Figure 1. Sibling sets with missing locations as weighted spatial subgraphs.

Within this network, the SDI of a sibship is the expected distance between any two randomly drawn siblings,

$$DI_S = \frac{2}{n(n-1)} \sum_{p<q} d_{pq}$$

where d_{pq} is the distance between the birthplaces of members p and q. The index is normalized by sibship size and equals zero when all members share the same birth county. For the SDI, distance is calculated as geodesic distance between county centroids. To measure dispersion in attribute space, we also define an attribute dispersion index (ADI), where distance is the dissimilarity between county attributes. Finally, to prevent double-counting siblings, each sibship is anchored to the census decade *y* during which the eldest sibling was 30±5 years old.

We measure and compare the index under four curation strategies:

(1) Available-case filtering (S1): Excludes individuals without a county-level birthplace and calculates the index using the remaining members.
(2) Complete-case filtering (S2): Excludes the entire sibship if any member is missing a county-level birthplace.
(3) Co-county imputation (IMP): For members with state-level but no county-level birthplace, imputes the missing location as the modal birth county of their state-born co-siblings. This assumes localized geographic clustering within the state.
(4) Random-forest multiple imputation (RF): Imputes the missing county from a predictive distribution, following the classification variant of the MICE random-forest mechanism (van Buuren & Groothuis-Oudshoorn, 2011). Predictors include the birth counties of co-siblings and of the sibling nearest in birth year, the parents' birthplaces, the birth decade, and sibship size (Shah et al., 2014). To account for imputation uncertainty, we calculate the SDI across M = 20 completed datasets and pool the estimates according to Rubin's rules (Little & Rubin, 1987).

Results and Discussion

The first test uses observed data without imputation, comparing sibships with fully resolved county-level birthplaces against those missing at least one birth county. Across every decade, incomplete sibships are 12.4 to 23.4 percentage points more likely to be dispersed (the risk difference). Among dispersed sibships, the median SDI is 26.5 to 55.7 km higher for incomplete sets, indicating that the complete case loses highly mobile families. The comparison shows birthplace missingness is not MCAR with respect to the SDI. We also found that the overall missingness rate is a poor predictor of this bias. The rank correlation between the missingness rate and the median dispersion gap is 0.17, and its correlation with the risk difference is −0.39. The missingness rate is therefore a weak guide to the size of the bias, and it relates to the two measures in opposite directions. Where missingness is high, complete and incomplete sibships differ less in whether they are dispersed, but somewhat more in how far the dispersed ones are spread.

The two imputation methods are applied to the available-case population (S1), which consists of 143,342 sibships. All scenario comparisons are conducted on this consistent

population baseline, with the exception of the complete-case filtering scenario (S2), which restricts the analytical sample to 108,510 fully attributed sibships.

Figure 2 compares the distribution of the index across the two filtering rules and the co-county imputation (IMP), showing that these strategies do not affect the shape of the distribution. However, in all thirteen decades, the available-case (S1) SDI median exceeds the complete-case (S2) SDI median by 6.2 to 13.3 km, and its dispersed share of sibships is 2.6 to 4.6 percentage points higher. This gap can be explained by an inherent bias in the family trees. Geographically stable families are more often documented completely in genealogies, meaning S2 inadvertently filters for stability. IMP fills missing values without adding dispersion because it can only place a member where the family is already located. As a result, it leaves the dispersed share identical to S1 but lowers the median SDI by 4.8 to 7.7 km, since each imputed location adds pairs at zero distance.

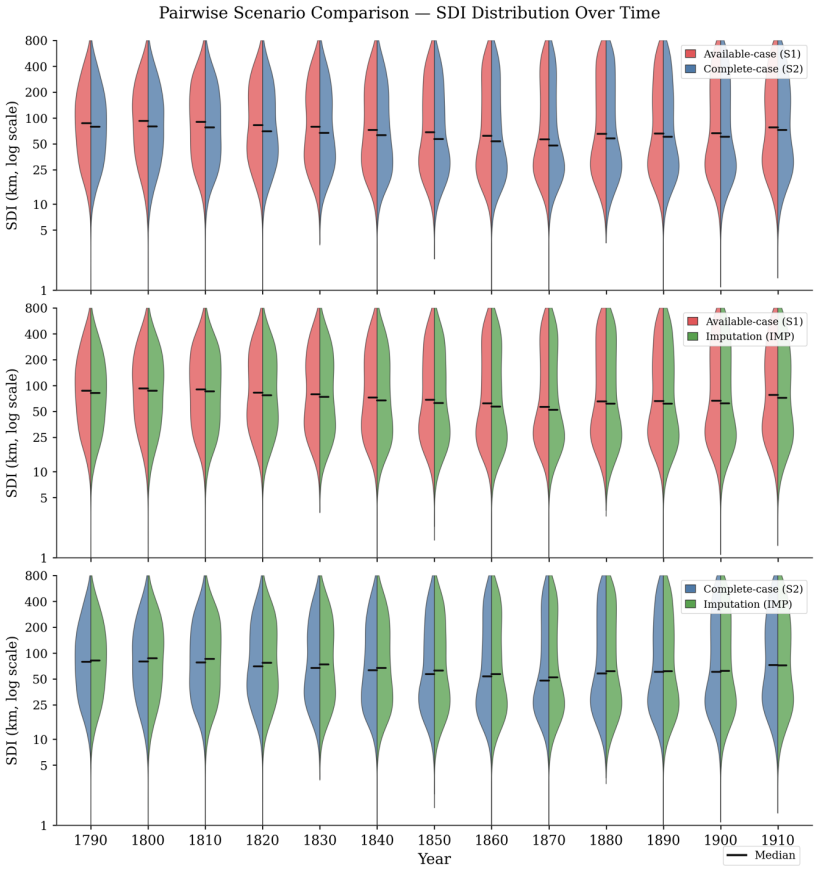


Figure 2. SDI distribution among dispersed sets over time under available-case filtering (S1), complete-case filtering (S2) and co-county imputation (IMP). Black rules mark medians.

RF imputation reproduces the same century-long shape of sibship dispersion as S1 (Figure 3). Both strategies show a steady clustering of sibling birthplaces starting in 1800, followed by renewed dispersion after 1880. However, under S1, the decline in the median SDI begins earlier and exhibits a steeper slope compared to the RF imputation, where the SDI varies less across decades and consistently remains above the S1 median.

The bar chart in Figure 3 shows the proportion of available family-based predictors for individuals with a missing birth county. County-level evidence is the most frequent predictor, but its availability decreases over time. An accuracy test on 30% of the data with observed birth counties shows that the RF model correctly imputes the county from an available sibling's birth county in 81% of cases.

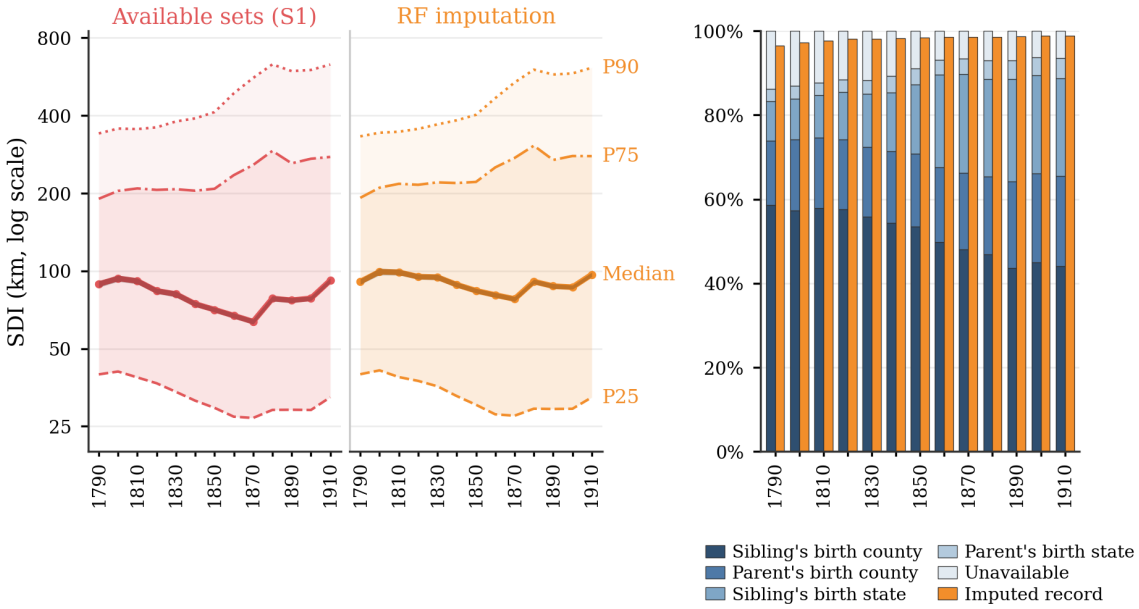


Figure 3. Left, quantiles of SDI among dispersed sets under available-case filtering and random-forest imputation, pooled by Rubin's rules over M = 20. Right, the family evidence available for each record missing a county, as a share of records per decade.

Maps of the median SDI in 1880 across the four strategies show a distinct 'frontier signature' characterized by high dispersion concentrations in the West and along the Mississippi River (Figure 4). However, the higher dispersion observed in the West and Midwest relative to the East is partially attributable to larger county administrative boundaries in those regions. Comparing the maps, the sibships excluded by S2 are concentrated primarily in the Midwest and West. Comparing S1 and RF maps shows that RF imputation increases both the measured dispersion and the total number of retained sibships in California, Washington, and East Texas, which explains the overall increase in SDI under the RF strategy.

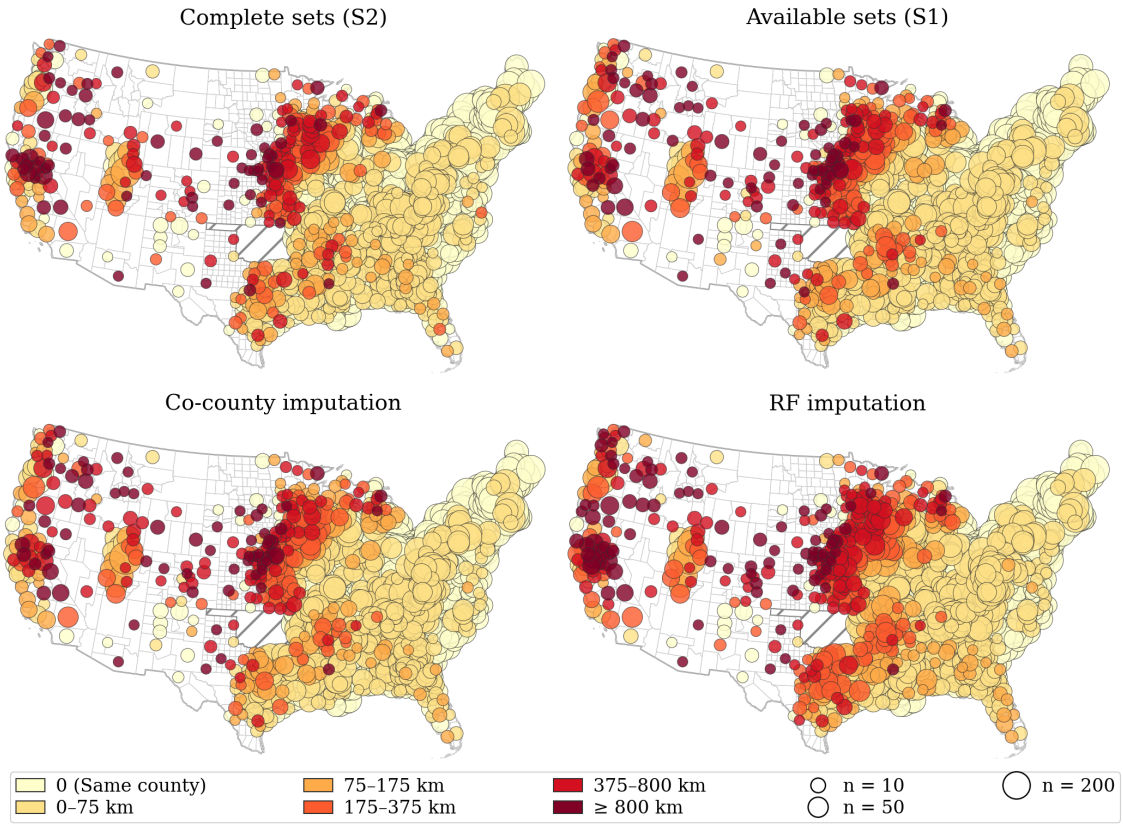


Figure 4. Median SDI in 1880 under the four treatments, plotted at the mean population centre of each set's located birth counties. Symbol size is the number of sets.

Finally, we replaced geographic distance with Gower dissimilarity based on county population density, improved farmland, and coastal attributes (Figure 5). Spearman rank correlations between SDI and ADI range from 0.94 to 0.98 across the decades with county attribute data (1850–1910) and across all four treatments. The non-MCAR pattern reported above therefore does not depend on measuring dispersion geographically.

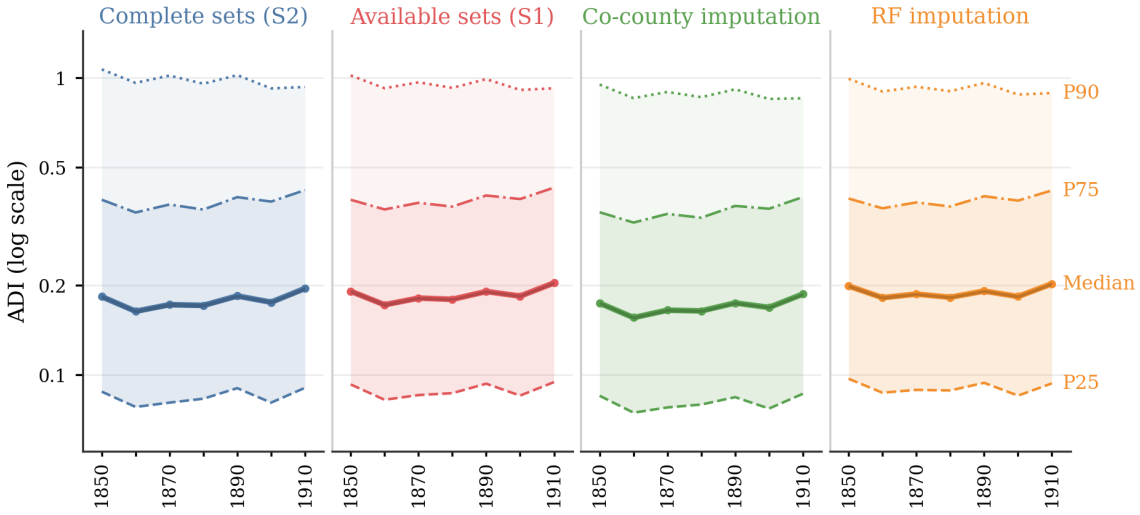


Figure 5. Quantiles of the attribute dispersion index (ADI) under the four treatments, 1850-1910.

Conclusion and Future Work

Birthplace missingness in family trees is not MCAR, and the choice of treatment strategy changes spatial measures in spatial-social networks. Excluding incomplete families mainly removes movers, which risks underestimating true migration patterns and family dispersion. Co-county imputation adds no new spatial links and pulls the network toward the family's mean population center, whereas drawing locations from a distribution conditioned on kin evidence increases both the length of the spatial edges and raises the dispersed share

Measuring distance between county centroids hides movement within a county and compresses the index where counties are small. The 81% imputation accuracy is measured on records with an observed county, so it is an upper bound for the records we actually impute. The tree also over-represents well-documented families, so record missingness compounds the attribute missingness studied here.

Future work will aggregate the sibship subnetworks into a location matrix of sibling co-occurrence, with an explicit node for unknown locations, to examine the effect of location missingness on network-level metrics. Initial maps suggest that the imputed locations concentrate on the frontier. Regressing the measured dispersion gap on county characteristics, such as frontier status, urban composition, and demographic structure, would allow us to attribute these gaps to interpretable historical sources.

References

Andris, C. (2016). Integrating social network data into GISystems. *International Journal of Geographical Information Science*, *30*(10), 2009-2031.

Bennett, R. J., Haining, R. P., & Griffith, D. A. (1984). The Problem of Missing Data on Spatial Surfaces. *Annals of the Association of American Geographers*, *74*(1), 138-156. https://doi.org/https://doi.org/10.1111/j.1467-8306.1984.tb01440.x

Calderón-Bernal, L. P., Alburez-Gutierrez, D., & Zagheni, E. (2025). Analysing Biases in Genealogies Using Demographic Microsimulation. *European Journal of Population*, *41*(1), 34. https://doi.org/10.1007/s10680-025-09756-4

Colasurdo, A., & Omenti, R. (2024). Using online genealogical data for demographic research: An empirical examination of the FamiLinx database. *Demographic Research*, *51*(41), 1299-1350. https://doi.org/10.4054/DemRes.2024.51.41

Curriero, F. C., Kulldorff, M., Boscoe, F. P., & Klassen, A. C. (2010). Using imputation to provide location information for nongeocoded addresses. *PloS one*, *5*(2), e8998. https://doi.org/https://doi.org/10.1371/journal.pone.0008998

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D. G., Price, A. L., & Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, *360*(6385), 171-175. https://doi.org/doi:10.1126/science.aam9309

Koylu, C., Guo, D., Huang, Y., Kasakoff, A., & Grieve, J. (2021). Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S. *International Journal of Geographical Information Science*, *35*(12), 2380-2423. https://doi.org/10.1080/13658816.2020.1821885

Koylu, C., Helgertz, J., Roberts, E., & Kasakoff, A. B. (2025). *Linking Population-Scale Family Tree Data with the 1880 U.S. Census Records: An Initial Assessment of Representativeness* https://ssha2025.ssha.org/abstracts/250333

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61-77. https://doi.org/10.1080/15230406.2013.777139

Little, R., & Rubin, D. (1987). Multiple imputation for nonresponse in surveys. *Wiley*, *10*, 9780470316696.

Robinson, A. C. (2019). Representing the Presence of Absence in Cartography. *Annals of the American Association of Geographers*, *109*(1), 286-300. https://doi.org/10.1080/24694452.2018.1473754

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*, *179*(6), 764-774. https://doi.org/10.1093/aje/kwt312

Sutton, D., Basiri, A., Li, Z., & Brunsdon, C. (2026). Missing Data Can Be a Geographic Phenomenon. *Annals of the American Association of Geographers*, 1-19. https://doi.org/10.1080/24694452.2026.2640220

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1 - 67. https://doi.org/10.18637/jss.v045.i03

Ye, X., & Andris, C. (2021). Spatial social networks in geographic information science. *International Journal of Geographical Information Science*, *35*(12), 2375-2379. https://doi.org/10.1080/13658816.2021.2001722

Zahmatkesh, S., & Zech, P. (2026). Spatio-Temporal Missing Data Imputation: A Systematic Literature Review with a Focus on Statistical and Machine Learning-Based Approaches. *ACM Comput. Surv.*, *58*(10), Article 267. https://doi.org/10.1145/3797903

Zhang, G., & Zhu, A. X. (2018). The representativeness and spatial bias of volunteered geographic information: a review. *Annals of GIS*, *24*(3), 151-162. https://doi.org/10.1080/19475683.2018.1501607