GeoAlign: Interpolating Aggregates over Unaligned Partitions

Jie Song
University of Michigan
Ann Arbor, Michigan
jiesongk@umich.edu

Danai Koutra
University of Michigan
Ann Arbor, Michigan
dkoutra@umich.edu

Murali Mani
University of Michigan, Flint
Flint, Michigan
mmani@umflint.edu

H. V. Jagadish
University of Michigan
Ann Arbor, Michigan
jag@umich.edu

ABSTRACT

Answering crucial socioeconomic questions often requires combining and comparing data across two or more independently collected data sets. However, these data sets are often reported as aggregates over data collection units, such as geographical units, which may differ across data sets. Examples of geographical units include county, zip code, school district, etc., and as such, they can be *incongruent*. To be able to compare these data, it is necessary to realign the aggregates from the source units to a set of target spatially congruent geographical units. Existing intelligent areal interpolation/realignment methods, however, make strong assumptions about the spatial properties of the attribute of interest based on domain knowledge of its distribution. A more practical approach is to use available reference data sources to aid in this alignment. The selection of the references is vital to the quality of prediction.

In this paper, we devise `GeoAlign`, a novel multi-reference crosswalk algorithm that *estimates* aggregates in desired target units. `GeoAlign` is adaptive to new attributes with need for neither distribution-related domain knowledge of the attribute of interest nor knowledge of its spatial properties in Geographic Information System (GIS). We show that `GeoAlign` can easily be extended to perform aggregate realignment in multi-dimensional space for general use. Experiments on real, public government datasets show that `GeoAlign` achieves equal or better accuracy in root mean square error (RMSE) than the leading state-of-the-art approach without sacrificing scalability and robustness.

1 INTRODUCTION

Data are often found in silos, created independently. For example, administrative agencies and governments collect a great deal of data about their domain, most of which are then published in aggregate form. The primary purpose of the data collection is administrative, and the choice of data representation and structure is made by each agency for its own purpose. These data can be invaluable for understanding many social issues, particularly in conjunction with other data sources. However, most administrative agencies are not concerned with interoperability with other agencies, therefore standardization is unlikely. On the other hand, agencies value the privacy of individual citizens, and do not want any benefits from public data release to hurt their primary administrative mission. Therefore, in many cases, they will release data only in aggregate form. Similar reasoning applies in many other contexts as well. For example, Google Trends data is aggregated by geographical unit and time period, to avoid disclosing information about individual queries.

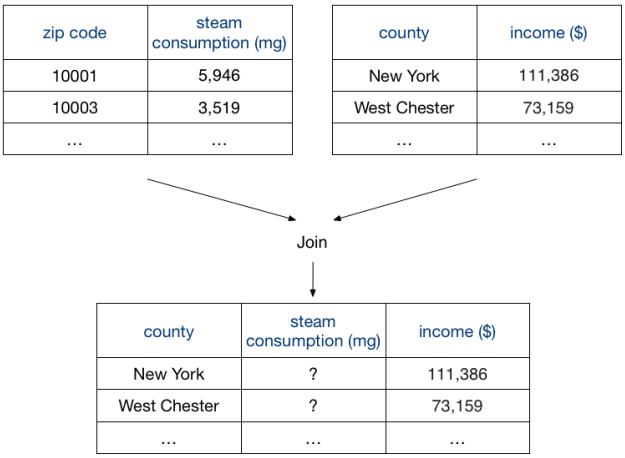


Figure 1: Join two tables for steam consumption (mg) and per capita income ($) in New York State together by county.

Data integration [25, 34] has been extensively studied, since there is often great benefit from joining multiple data sets. The bulk of the work on this topic addresses structural discrepancies, through schema mapping [2, 26, 42], and identification of individuals across data sets, through entity matching [21, 29]. One challenge not addressed in data integration is the case of data reported as aggregates over incompatible geographical/temporal units. This is a practical problem faced by government data center, NGOs, social scientists, and the general public when trying to related socioeconomic data to drive decision making processes, approximately 80% of which are related to a geographical location [14]. Even if the intention of joining such aggregated data based on their spatial or temporal properties seems to be the reasonable action of practice, these aggregates cannot easily be realigned accurately.

Motivating example. *Let us consider two tables shown in Figure 1 – one table has the steam consumption amount aggregated by zip code and the other has the per capita income reported by county. A sociologist wants to study the correlation of energy consumption with income in order to plan for future energy supply arrangement. Valuable insight could be obtained by joining these two tables. However, this is not straightforward since the data are reported on incompatible aggregate units, since one zip code may*

intersect several counties and one county may contain or overlap with multiple zip codes.

This challenge can be addressed by realigning one or both data sets to a common geographic type (target type) before performing the join. Let the intended target type be county, by which the per capita income is already reported. However, we only know the steam consumption amount by zip code, and have to estimate the number for each county. This estimate is obtained as a form of interpolation. Finding a good estimate of steam consumption per county is the challenge we need to address.

This problem of estimating aggregate values for geographic areas arises in many contexts, and has been extensively studied. *Areal Interpolation*, in Geographical Information Systems (GIS), is the process of aligning an attribute from one areal unit system (the source type of a set of polygons) to another spatially incongruent system (the target type of another set of polygons) [12, 22, 23, 31, 33]. It is more commonly known as *crosswalk*, or the *modifiable areal unit problem* in socioeconomic fields. If the attribute is uniformly distributed in space, then the interpolation can be performed in a straightforward way based on area. For example, if 70% of the area of a zip code lies in county A and 30% in county B, then we could estimate that 70% of the crimes reported in the zip code occurred in county A and the remaining 30% in B.

This uniform distribution assumption or homogeneity assumption rarely holds in practice. If we know something about the distribution, that can be taken into account in the interpolation. For example, if we know that more crimes occur in densely populated urban areas than in sparsely populated rural areas, we can take this into account. The mathematics can be tricky depending on exactly what we know about the distribution of the attribute of interest, so there has been a stream of research in the literature towards solving the problem based on different assumptions. We discuss this more in the related work.

In the data integration scenario, we often do not know much about an attribute of interest. Therefore, we may be unable to develop good rules for how it should be distributed. Even so, we can do better than make an unrealistic uniformity assumption, if we have access to additional data. In particular, if we can find a *reference* attribute, for which we know the detailed distribution, we can use it to perform a crosswalk from source units to target units of aggregation. For example, we may have detailed distribution available for population, with fine granularity aggregates giving us the population in every intersection of county and zip code. If we believe the crimes are distributed similarly to population (or at least more similarly to population than to area), then we can exploit our knowledge of population distribution to estimate the desired aggregates for number of crimes. In particular, consider a zip code with a population of 25,000 people. Suppose this zip code intersects two counties A and B, with the population in the intersections being 10,000 and 15,000 respectively. Suppose that we know there were 100 reported crimes in this zip code last year. We can estimate that 40 of these crimes occurred in county A and 60 occurred in county B, following the same ratio as the population. This approach makes no assumptions about the probability distribution of the reference attribute or the attribute of interest. It can work well if the attribute of interest is distributed similarly to the reference attribute. To the extent the distributions differ, the estimates will be off.

In this paper, our goal is to solve this data alignment problem through the use of more data. We often may have access to more than one candidate reference attribute, each with its own distribution. We may not have domain knowledge enough to understand which reference is most similar to our variable of interest. Even if we found the best reference, its distribution may still not be close enough. Is there some way we can combine the information in the multiple reference attributes to do better? And at the same time, more adaptively predicts the estimates to new attributes of interest than using a single reference.

In this paper, we develop `GeoAlign`, a technique that does just this. The idea is to weight their relative contributions to the final estimate so that the most similar reference attributes have the greatest impact on the estimate.

The intellectual contributions of the paper are as follows:

- We define the general aggregate interpolation problem over unaligned partitions in one or more dimensions, which is an important problem in data integration (§2).
- We propose `GeoAlign`, an adaptive multi-reference crosswalk algorithm that solves the areal interpolation problem by realigning aggregates from source units to target units by learning distribution similarities between the attribute of interest and the reference attributes (§3). We show that `GeoAlign` can be used not just in two-dimensional maps but also for spaces with arbitrary numbers of dimensions.
- We evaluate the performance of `GeoAlign` against real data from data.ny.gov and Esri data in 2-dimensional space. These experiments show that `GeoAlign` outperforms the state-of-the-art single reference crosswalk approach in accuracy (§4). It is, at the same time, efficient, scalable and robust to noisy references even when limited references are available.

We then survey related work in areal interpolation (§5) before we conclude with future work (§6).

2 PROBLEM STATEMENT

In this section, we first introduce the terms we use throughout this paper before we formally define the aggregate interpolation problem in multi-dimensional space. We then illustrate, with examples, the aggregate interpolation problem in 2-D and in other dimensions.

2.1 Preliminaries

In Geometry, an n-dimensional universe $\Omega \subset \mathbb{R}^n$ can be partitioned into some *unit system* γ^y composed of a set of *units* $U^y = \{u_1^y, u_2^y, ...\}$, where $\forall_{u_i^y \in U^y} u_i^y \subset \mathbb{R}^n$. Units in U^y satisfy

$$\forall_{u_i^y, u_j^y \in U^y, i \neq j} u_i^y \cap u_j^y = \emptyset, \tag{1}$$

that is any pair of units in U^y is disjoint with each other since they have no spatial overlap in n dimensions. Suppose that an attribute of interest α_x exists, then we denote its *aggregate vector* as $a_x^y = [a_x^y[1], a_x^y[2], ..., a_x^y[|U^y|]]$ such that $a_x^y[i]$ is the aggregate of α_x in the ith unit of U^y.

As an example in 2-D space, in the universe of New York State Ω, county partitions compose a unit system γ^y. They share no areal intersection such that they are spatially incongruent with each other. Steam consumption, which is the attribute of interest α_x, has its data in Figure 1 collected from such a set of county units U^y. Another possible unit system is zip code partitions. We can view the steam consumption column in the table as its aggregate vector a_x^y for the county unit system. Each entry of

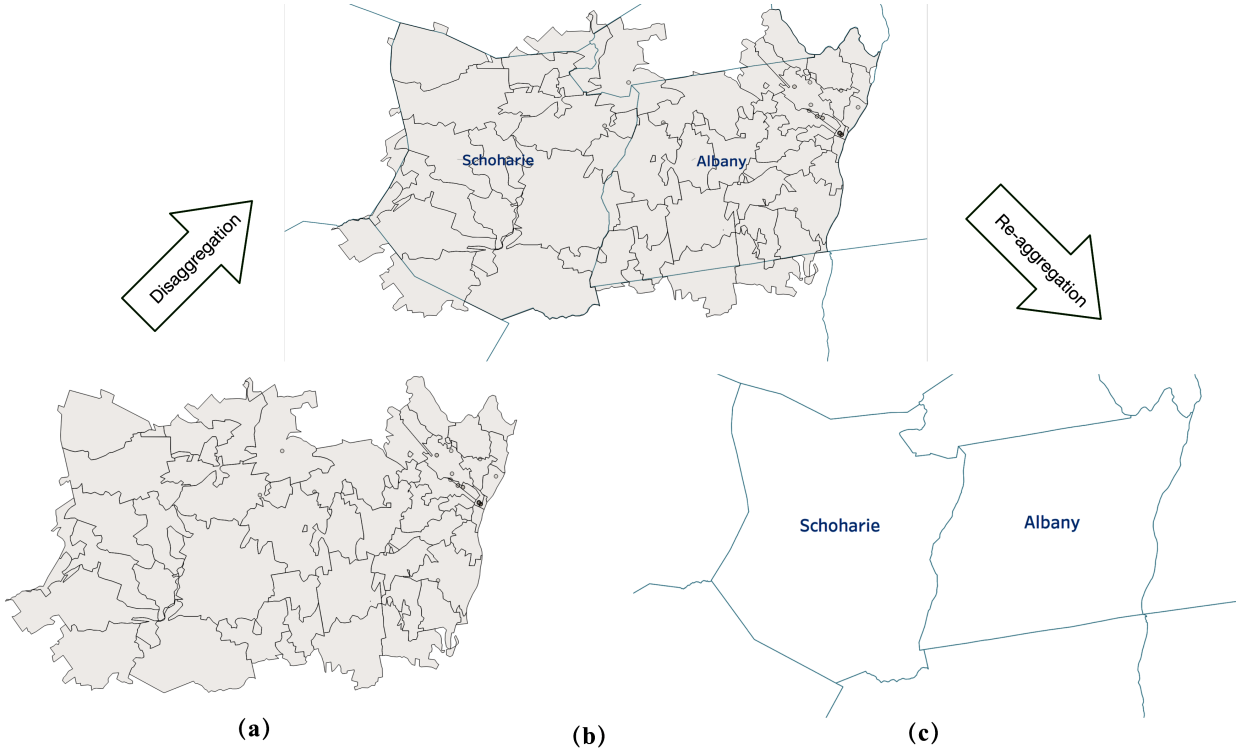


Figure 2: Examples of units in the partial map of New York State for aggregate interpolation: (a) zip code units (source units), (b) zip code and county intersection units and (c) county units (target units).

the vector represents the amount of steam consumption in some county.

2.2 The Aggregate Interpolation Problem

We define the following terms for the aggregate interpolation problem in $\mathbb{R}^n$:

- $U^s = \{u_1^s, u_2^s, ...\}$, *source units* of the *source unit system* γ^s in the universe Ω.
- $U^t = \{u_1^t, u_2^t, ...\}$, *target units* of the *target unit system* γ^t in the same universe.
- $a_o^s = [a_o^s[1], a_o^s[2], ..., a_o^s[|U^s|]]$, aggregate vector of the objective attribute α_o in source units. $a_o^s[i]$, the i^{th} aggregate of a_o^s, is collected from source unit u_i^s.
- $a_o^t = [a_o^t[1], a_o^t[1], ..., a_o^s[|U^t|]]$, aggregate vector of the objective attribute α_o in target units. $a_o^t[j]$, the j^{th} aggregate of a_o^t, is collected from target unit u_j^t.

Given U^s, U^t and a_o^s, *aggregate interpolation* approximates a_o^t as $\hat{a}_o^t = [\hat{a}_o^t[1], \hat{a}_o^t[2], ..., \hat{a}_o^t[|U^t|]]$.

Aggregate Interpolation Problem in 2-D When it comes to a 2-dimensional space $\mathbb{R}^2$, units are *simple polygons* consisting of straight, non-intersecting edges forming a closed path by pairwise join. A unit in 2-dimensional space can be denoted by

$$u_i = (V_{u_i}, E_{u_i}) \text{ where } |V_{u_i}| = |E_{u_i}| = n_i, \tag{2}$$

where V_{u_i} is a set of vertices in $\mathbb{R}^2$ and E_{u_i} is a set of edges connecting the vertices in V_{u_i} such that every vertex is shared by exactly two edges. Then, u_i is the closed area formed by connecting n_i vertices in V_{u_i} by n_i edges in E_{u_i}.

This problem is referred to as the *areal interpolation* problem in the GIS community. The 2-dimensional space is the map; and the unit system, also recognized as feature layer in GIS, is composed of partitions delimited by boundaries of some geographic type. Some of the most widely used geographic types in demographic data are county, zip code, and more. For instance, as shown in Figure 2, U^s is the feature layer for zip code in (a); U^t is the other feature layer for counties in (c). Given the aggregates of steam consumption in zip codes a_o^s shown in Figure 1 from the motivating example, the aggregate interpolation problem in 2-D approximates the steam consumption in counties, $\hat{a}_o^t$.

Aggregate Interpolation Problem in other dimensions In the 1-dimension setting of the problem, units are *intervals* or *line segments* between two points such that

$$u_i = [u_{i_1}, u_{i_2}], \tag{3}$$

where u_{i_1} and u_{i_2} are two points on the real line $\mathbb{R}$. We may illustrate the problem as interpolation of population histogram aggregates for two sets of age intervals as depicted in Figure 3. In this case, we can treat the set of narrow bins of age in (a) as U^s, the set of wide bins of age in (b) as U^t, for the same range of age as the universe of interest Ω. Given the population histogram for narrow age bins, a_o^s, the aggregate interpolation problem in 1-D predicts the population histogram for wide age bins $\hat{a}_o^t$.

Unit system overlapping also exist in 3-D or higher dimensions. One example is 3-D GIS data, such as the distribution of disease, evaluated for cubic units of different size scales. Another example is the data collected for 4-D space (3D) and time systems, such as environmental exposures, crosswalked to another system incongruent in both space and time units. For both cases, areal interpolation is the bridge to map the data across unit systems to enable side-by-side comparison with data from incompatible units.

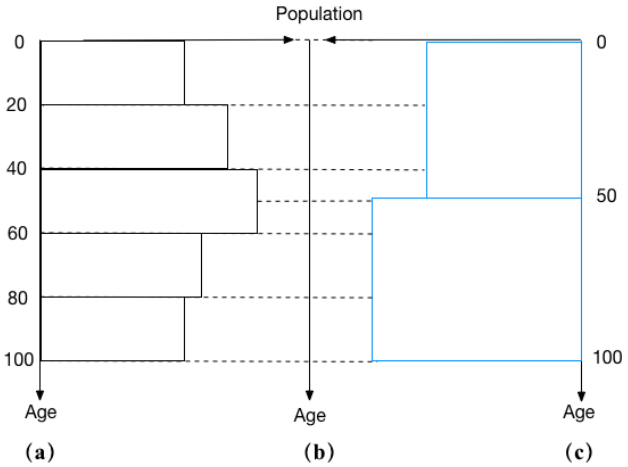


Figure 3: Realign population histogram in two sets of age intervals by transforming aggregates from (a) narrow bins to (c) wide bins. The dotted lines separate the age range into a set of tentative intersection units as in (b).

3 AGGREGATE INTERPOLATION BY GEOALIGN

In this section, we first introduce some additional definitions and notations used throughout the rest of the paper and a general two-step solution solving the aggregate interpolation algorithm. We then lay the groundwork for the assumptions made by GeoAlign before exploring the details of the algorithm.

3.1 GeoAlign preliminaries

Before introducing the general steps to solve the aggregate interpolation problem, we further define the set of *intersection units* for the intersection unit system γ^{st} as $U^{st} = \{u_1^{st}, u_2^{st}, ...\}$, where $\forall_{u_k^{st} \in U^{st}}, u_k^{st} \subset \mathbb{R}^n$. Each intersection unit is a subregion within some source unit and some target unit, that is

$$\forall u_k^{st} \in U^{st}, \exists u_i^s \in U^s \wedge u_j^t \in U^t, u_k^{st} \subseteq u_i^s \text{ and } u_k^{st} \subseteq u_j^t. \quad (4)$$

It can be thus deduced that $|U^{st}| \geq max(|U^s|, |U^t|)$.

The aggregate vector of the intersection units for some attribute α_x is denoted as $a_x^{st} = [a_x^{st}[1], a_x^{st}[2], ..., a_x^{st}[|U^{st}|]$.

In the simplest case, the intersection units are the n-dimensional spatial intersections of source and target units. For instance, for the areal interpolation problem in Figure 2, U^{st} is the set of intersection areas between zip codes and counties in (b); and for the histogram realignment problem in Figure 3, U^{st} is the set of age intersection intervals between source and target bins. More fine-grained partitions of intersection units may be introduced if necessary when disparate spatial properties of the attribute in these partitions are introduced by auxiliary data.

Assume that the probability density function of attribute α_x for γ^{st} is a piecewise function, denoted as

$$f_x^{st}(z) = \begin{cases} f_x^{st}[1](z) & , z \subset u_1^{st} \\ f_x^{st}[2](z) & , z \subset u_2^{st} \\ ... & \\ f_x^{st}[|U^{st}|](z) & , z \subset u_{|U^{st}|}^{st} \end{cases} \quad (5)$$

is known, then its aggregate in the source units and target units follows:

$$a_i^s = \sum_{\forall u_k^{st} \in U^{st}, u_k^{st} \subseteq u_i^s} a_k^{st} = \sum_{\forall u_k^{st} \in U^{st}, u_k^{st} \subseteq u_i^s} \int_{z \subset u_k^{st}} f_x^{st}[k](z)dz, \quad (6)$$

and similarly

$$a_j^t = \sum_{\forall u_k^{st} \in U^{st}, u_k^{st} \subseteq u_j^t} a_k^{st} = \sum_{\forall u_k^{st} \in U^{st}, u_k^{st} \subseteq u_j^t} \int_{z \subset u_k^{st}} f_x^{st}[k](z)dz. \quad (7)$$

Alternatively speaking, the aggregate in each source/target unit is equivalent to the sum of aggregates of all intersection units within it.

Two-step Approximation. We use a two-step solution to solve the aggregate interpolation problem for objective attribute α_o. In our solution, we first compute the approximate a_o^{st} (a_o^{st} is the aggregate vector for the intersection units). We then aggregate these approximate intersection unit aggregates to determine the approximate target unit aggregates. The two steps in our solution are described below:

(1) **Disaggregation**: Split the aggregates in each source unit to its intersection units. Mathematically speaking,

$$\hat{a}_o^{st}[k] = \mathcal{B}(a_o^s[i], ...), \text{ s.t. } u_k^{st} \subseteq u_i^s, \quad (8)$$

where the disaggregation function $\mathcal{B}(a_o^s[i], ...)$ computes the approximated $\hat{a}_o^{st}[k]$ of $a_o^{st}[k]$. Note that ... denotes the ancillary data that contribute to the approximation. Some of the most commonly used ancillary data are shape files of u_i^s and u_k^{st}, etc. More advanced approximation function may use external ancillary data. For instance, the distribution of a reference attribute that is positively related to the distribution of α_o.

(2) **Re-aggregation**: Aggregate the approximated intersection unit aggregates for the target unit they reside in, or equivalently

$$\hat{a}_o^t[j] = \sum_{\forall u_k^{st} \in U^{st}, u_k^{st} \subseteq u_j^t} \hat{a}_o^{st}[k]. \quad (9)$$

General Solution Properties. Regardless of the types of ancillary data available, some constraints are widely adopted in the existing two-step approximation solutions. We name two of them here.

One of these constraints is the *volume preserving* property [31, 46]. This property ensures that every source aggregate is preserved by the total of approximated aggregates in its intersection units, or

$$a_o^s[i] = \sum_{\forall u_k^{st} \in U^{st}, u_k^{st} \subseteq u_i^s} \hat{a}_o^{st}[k]. \quad (10)$$

The property is improving the estimation in that greater fidelity is given to the approximation in the intersection units, which propogates to a more accurate estimation in target units. It has been shown experimentally that methods following the volume preserving property make comparatively better predictions [31].

Homogeneity is also often used to compensate for the absence of information. Mathematically, for some attribute α_x, its probability density function in a given unit is constant. In other words,

its aggregate on any sub-unit of the given unit is proportional to the area of the sub-unit. However, the assumption of homogeneity is rarely met in the real world [49].

3.2 GeoAlign Assumptions

We often have access to multiple reference attributes, no one of which perfectly matches the objective attribute we wish to estimate. It would appear advantageous for us to use all of them instead of using a single reference attribute as the current extensive approaches described above. To this end, we propose GeoAlign, an aggregate interpolation algorithm that realigns aggregated data by learning from a combination of reference attributes to best predict the actual aggregates of the objective attribute in target units. GeoAlign leverages the advantages of extensive approaches and is, at the same time, robust to various objective attributes.

An intuitive idea could be to model the objective attribute aggregates as a function of multiple reference attributes aggregates in source units, evaluate coefficients with estimation methods and substitute reference attributes in target units for prediction. However, this is not applicable for the aggregate interpolation algorithm since training samples (objective attribute aggregates in source units) and test samples (objective attribute aggregates in target units) are not randomly drawn from the same population and the test samples are constrained by the training samples.

To address the linkage between two sets of samples and to account for the scale variations of reference attributes, in GeoAlign, the realignment of the objective attribute is related to that of the reference attributes through a statistical model for re-aggregation. In order to make the problem tractable, we assume that different attributes are independent across source units, and that every attribute is correlated in its distribution between source and target units. We will loose the independence assumption of references later as shown in experiments in §4.4.2.

3.3 Disaggregation Matrix

Since we study the partition of aggregates in intersection units, in the disaggregation step, $\mathcal{B}(a_o^s[i], ...)$ can be reformulated as

$$\hat{a}_o^{st}[k] = \frac{\omega_o^{st}[k]}{\omega_o^s[i]} a_o^s[i]$$
$$\text{subject to} \sum_{\forall u_k^{st} \in U^{st}, u_k^{st} \subseteq u_i^s} \omega_o^{st}[k] = \omega_o^s[i], \tag{11}$$

where $\frac{\omega_o^{st}[k]}{\omega_o^s[i]}$ is the share of aggregate in the k-th intersection unit ($\omega_o^{st}[k]$) over that in the i-th source unit ($\omega_o^s[i]$) it resides in. Intuitively, the re-aggregation step sums up the weighted share of all intersection units in all source units that overlap with the target unit. Alternatively speaking,

$$\hat{a}_o^t[j] = \sum_{\forall u_i^s, u_i^s \cap u_j^t \neq \emptyset} \frac{\sum_{\forall u_k^{st} \subseteq u_i^s \cap u_j^t} \omega_o^{st}[k]}{\omega_o^s[i]} a_o^s[i]. \tag{12}$$

Rather than approximating a_o^{st} in the disaggregation step, we can instead infer $\frac{\omega_o^{st}[k]}{\omega_o^s[i]}$, $\omega_o^{st}[k]$ or $\sum_{\forall u_k^{st} \subseteq u_i^s \cap u_j^t} \omega_o^{st}[k]$. This choice often depends on the type of ancillary data available. The most widely used ancillary data is the true disaggregation of a reference attribute between source and target units. For instance, for the population reference mentioned in the introduction, the population aggregates in intersection units of counties and zip codes. We denote the *disaggregation matrix* of some attribute α_x between two unit systems γ^{y_1} and γ^{y_2} as $DM_x^{y_1,y_2}$, where $DM_x^{y_1,y_2}[i,j]$ is its aggregate in the intersection area of $u_i^{y_1}$ and $u_j^{y_2}$. For γ^s and γ^t,

$$DM_x^{s,t}[i,j] = \sum_{\forall u_k^{st} \subseteq u_i^s \cap u_j^t} a_x^{st}[k] \tag{13}$$

The disaggregation matrix of the reference attribute between source and target units is often wrapped up in a crosswalk relationship file. When the disaggregation matrix of only one reference attribute α_r is available, we can substitute $\sum_{\forall u_k^{st} \subseteq u_i^s \cap u_j^t} \omega_o^{st}[k]$ for $DM_r^{s,t}$ to complete the approximation of the objective attribute in target units. This type of method is named as the *dasymetric method* [32, 33, 48]. A special case of it is the areal weighting method [30], using the disaggregation matrix of area as the reference. Dasymetric methods are widely employed in socioeconomic data realignment by general users [10].

Since we only consider the disaggregation matrix between source and target units, from now on, we use DM_x for $DM_x^{s,t}$.

Table 1: Notations in §2 and 3

Notation	Description
Ω	an n-dimensional universe of interest
γ^y	a unit system in Ω, for example γ^s at source level
$U^y = \{u_1^y, u_2^y, ...\}$	the set of units in γ^y
α_o	the objective attribute
$A_r = \{\alpha_{r_1}, \alpha_{r_2}, ...\}$	the set of reference attributes
$\alpha_x \in \alpha_o \cup A_r$	an attribute of interest
$a_x^y = [a_x^y[1], a_x^y[2], ...a_x^y[\lvert U^y \rvert]]$	the aggregate vector of α_x in units of U^y
f_x^y	the probability density function of α_x for γ^y
$\mathcal{B}(a_o^s[i], ...)$	the disaggregation function
ω_x^y	the weighted share vector of α_x for γ^y
$a_x'^y$	the normalized a_x^y
$DM_x^{y_1,y_2}$	the dimension matrix of α_x, where $DM_x^{y_1,y_2}[i,j]$ is the aggregate of α_x in the intersection of $u_i^{y_1}$ and $u_j^{y_2}$
$\beta = [\beta_1, \beta_2, ...\beta_{\lvert A_r \rvert}]$	weights computed from Equation (15)

3.4 GeoAlign Algorithm

In the real world, the disaggregation matrix of more than one references attributes is often available. GeoAlign is a volume-preseving method that leverages the distribution similarity of the objective attribute with reference attributes at the source level and predicts the dimension matrix of the objective as a weighted combination of the dimension matrices of the references. We will first extend some of the notations in Section 2, and then describe our proposed algorithm in detail.

Notation. Let $A_r = \{\alpha_{r_1}, \alpha_{r_2}, ...\}$ be the set of reference attributes available. The aggregate vectors of these reference attributes in source units are represented as $a_{r_1}^s, a_{r_2}^s, \ldots, a_{r_{|A_r|}}^s$, where $a_{r_k}^s = [a_{r_k}^s[1], a_{r_k}^s[2], ..., a_{r_k}^s[|U^s|]]$ for the kth reference attribute. Similarly, the aggregate vectors of these reference attributes in target units are represented as $a_{r_1}^t, a_{r_2}^t, ..., a_{r_{|A_r|}}^t$, where $a_{r_k}^t = [a_{r_k}^t[1], a_{r_k}^t[2], ..., a_{r_k}^t[|U^t|]]$.

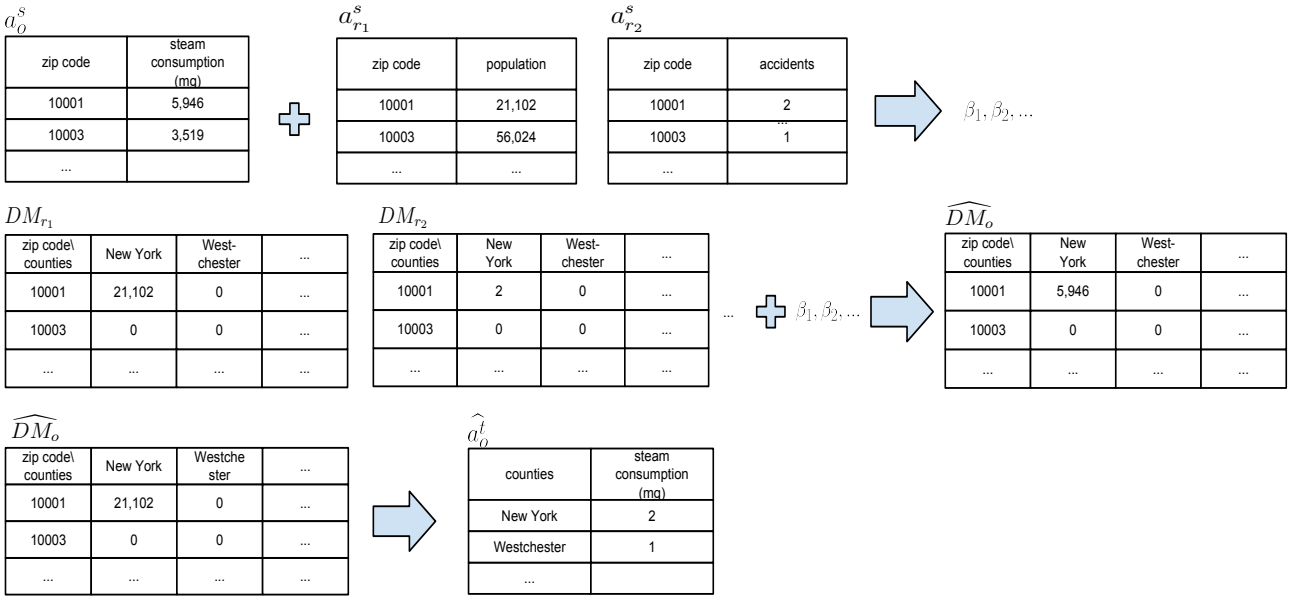


Figure 4: GeoAlign interpolation for the objective steam consumption data in Figure 1 from zip codes to counties using two reference attributes: population and accidents, in three steps: weight learning, disaggregation and re-aggreagtion.

We assume that the ancillary data available is the disaggregation matrix of all the reference attributes. We denote the disaggregation matrix of the kth reference attribute as DM_{r_k}.

To avoid variation in scale, we normalize the objective attribute and the references at the source level, adjusting their values measured on different scales to a notionally common scale. This is reasonable in two ways. First, GeoAlign is dependent on the distribution similarity between the objective attribute and the references across source units rather than their actual value similarity. Second, GeoAlign jointly considers the similarity of the objective with multiple references. The magnitude of the references should not be a contributing factor.

The normalized $a_{r_k}^s$ is denoted by $a'^s_{r_k}$ for $k = 1, 2, ..., |A_r|$, and is computed as $a'^s_{r_k} = a^s_{r_k} / \max_{i, i \le |U^s|} a^s_{r_k}[i], a^s_{r_k}[i] \ge 0$.

The aggregate vector of the objective attribute in source units a_o^s is also normalized similarly, and is denoted as a'^s_o.

GeoAlign Steps In the disaggregation step, GeoAlign computes $\widehat{DM_o}$, which is the estimated weighted dimension matrix of the objective attribute. Our intention is to best predict $\widehat{DM_o}$, and at the same time, preserve its volume preserving property. We propose

$$\widehat{DM_o}[i,j] = \begin{cases} \frac{\sum_{k=1}^{|A_r|} \beta_k \times DM_{r_k}[i,j]}{\sum_{k=1}^{|A_r|} \beta_k \times a^s_{r_k}[i]} \cdot a^s_o[i], & \sum_{k=1}^{|A_r|} a^s_{r_k}[i] \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where $\beta = [\beta_1, \beta_2, ..., \beta_{|\mathbf{A_r}|}]$ is the learned weight vector and $\sum_{i=1}^{|A_r|} \beta_i = 1$.

Our preliminary experiments lay the ground work of our assumption such that the higher the similarity between two attributes at the source level, the more likely their distribution in the intersection level are similar. We can thus express the objective attribute as linearly associated with the reference attributes for both aggregate vector in source units and disaggregation matrix. The weights are obtained by solving a constrained linear least squares programming problem with objective function:

$$\begin{aligned} &\min_{\beta} \frac{1}{2} ||\mathbf{A}\beta - \mathbf{b}||^2 \\ &\text{subject to } \sum_{k=1}^{|A_r|} \beta_k = 1 \\ &\text{where } \beta_k \ge 0, \text{ for } k = 1, 2, ..., |A_r| \end{aligned} \tag{15}$$

where $\mathbf{A}$ is the column-wise concatenation of $a'^s_{r_k}$ for $k = 1, 2, ..., |A_r|$ and $\mathbf{b}$ is a'^s_o. Instead of computing $\widehat{DM_o}$ by directly applying the weights to DM_{r_k}s, we adapt it to the scale of reference attributes and insert back the weights to Eq. (14) to get an adjusted $\widehat{DM_o}$.

The approximated disaggregation matrix of the objective attribute satisfies the volume preserving property such that

$$\widehat{DM_o}[i,j] \ge 0 \quad \text{and} \quad \sum_{j=1}^{|U^t|} \widehat{DM_o}[i,j] \approx a^s_o[i]. \tag{16}$$

The estimated aggregates of the objective attribute in target units are computed in the reaggregation step as

$$\widehat{a}^t_o[j] = \sum_{i=1}^{|U^s|} \widehat{DM_o}[i,j] \tag{17}$$

Following the pseudocode in Algorithm 1, we further illustrate the algorithm by the motivating example in Figure 1, with the steps depicted in Figure 4. Assume that GeoAlign is crosswalking the steam consumption objective from zip codes to counties. Moreover, assume that the aggregate vectors, $a^s_{r_1}$ and $a^s_{r_2}$, and the disaggregation matrices, DM_{r_1} and DM_{r_2}, for two reference attributes, population and accidents, are readily available. Maximizing the distribution similarity across units between the normalized objective, a'^s_o, and the normalized references, $a'^s_{r_1}$ and $a'^s_{r_2}$, the objective attribute is first optimized as a weighted combination of the references at the source level (zip code level). The weights, β_1 and β_2, are then reassigned to the disaggregation matrices of the references DM_{r_1} and DM_{r_2}, and adjusted to predict an approximated disaggregation of the objective $\widehat{DM_o}$. The approximated disaggregation matrix is eventually re-aggregated

Algorithm 1: GeoAlign

Input: aggregate vectors of reference attributes in source units $a^s_{r_1}, a^s_{r_2}, ..., a^s_{r_{|A_r|}}$; corresponding disaggregation matrices $DM_{r_1}, DM_{r_2}, ..., DM_{r_{|A_r|}}$; and the aggregate vector of the objective attribute in source units a^s_o.

Output: estimated aggregates of the objective attribute in target units $\hat{a}^t_o$

1 **Step 1. Weight Learning:** Compute weights, β, by solving the least squares problem in Equation (15)

2 **Step 2. Disaggregation:** Compute the estimated weighted disaggregation matrix of the objective attribute, $\widehat{DM_o}$, using Equation (14)

3 **Step 3. Re-aggregation:** Re-aggregate to estimate the aggregates of the objective attribute in target units, $\hat{a}^t_o$, using Equation (17)

to derive an approximate of the objective at the target county level ($\hat{a}^t_o$).

It can be easily shown that GeoAlign is applicable to any dimension since the algorithm involves no dimension dependent information or computation. Rather, the only information needed is the true partition of reference attributes in source and target intersection units regardless of dimension or dimension-related information, such as spatial correlation for geospatial data. Alternatively, if true partition of references in finer granularity is available, the data can be aggregated to the level of source and target intersection as a reference attribute.

4 EXPERIMENTAL EVALUATION

We evaluated the feasibility of the GeoAlign algorithm from two crucial aspects: whether the algorithm can correctly complete the realignment task (effectiveness), and whether the runtime of the algorithm is fast enough (efficiency). Additionally, we consider runtime scalability when larger datasets are involved and the robustness of the algorithm when low quality or limited reference attributes present.

We compare the performance of GeoAlign with that of areal weighting method [31] and dasymetric method [32, 33, 48] that utilizes three reference attributes separately.

4.1 Experimental Setup

We developed the GeoAlign algorithm in Python. All experiments were performed on a 2.3 GHz Intel Core i7 with 8 GB memory and a 7200 rpm SATA disk.

We evaluated GeoAlign for 2-D areal interpolation. We used county and zip code as the two geographic types of interest, and focused on data from two different universes, New York State and the United States. Most of the New York State data ere collected from data.ny.gov, populated in tabular form. Three population level demographic datasets have been used as reference data for the single crosswalk algorithm, namely the population data from United States Census Bureau [4], the aggregated USPS residential address data and the aggregated USPS business address data [41]. In addition, we also selected five large individual level datasets (The New York State Restaurants dataset is generated by selecting unique restaurants in the Food Service Inspections dataset) with geographic information and aggregated their number of records for the intersection area of the two geographic types to form their disaggregation matrices [5–8]. Thus we obtained a total of eight reference datasets with accurate distributions by zip code and by county, and their disaggregation matrices from zip codes to counties.

Besides the three population level Census data, which cover the entire nation including New York State, other data for the United States were collected from Esri, where the Maps and Data group provides publicly available geocoded GIS data. Six individual level GIS data [15–20] were aggregated based on their geospatial information for zip code and county levels and their intersections using ArcGIS Pro [27]. We also computed the area of units at these three levels, which is later used as the reference attribute by the areal weighting method, yielding 10 datasets in total for the universe of the United States.

There are more datasets with attributes for which the aggregate vectors are available for both zip code and county for New York State or for the United States. However, we did not use them as reference attributes due to two reasons. First, it was not clear whether these aggregates are accurate or approximate. In §4.4.1, we further discuss the impact of the reference approximates on the prediction. The other reason is that several attributes do not have their disaggregation matrices publicly accessible and such attributes cannot be used as reference attributes. In case of limited reference attributes, we show in §4.4.2 that GeoAlign makes reasonable predictions even when the references are poorly selected.

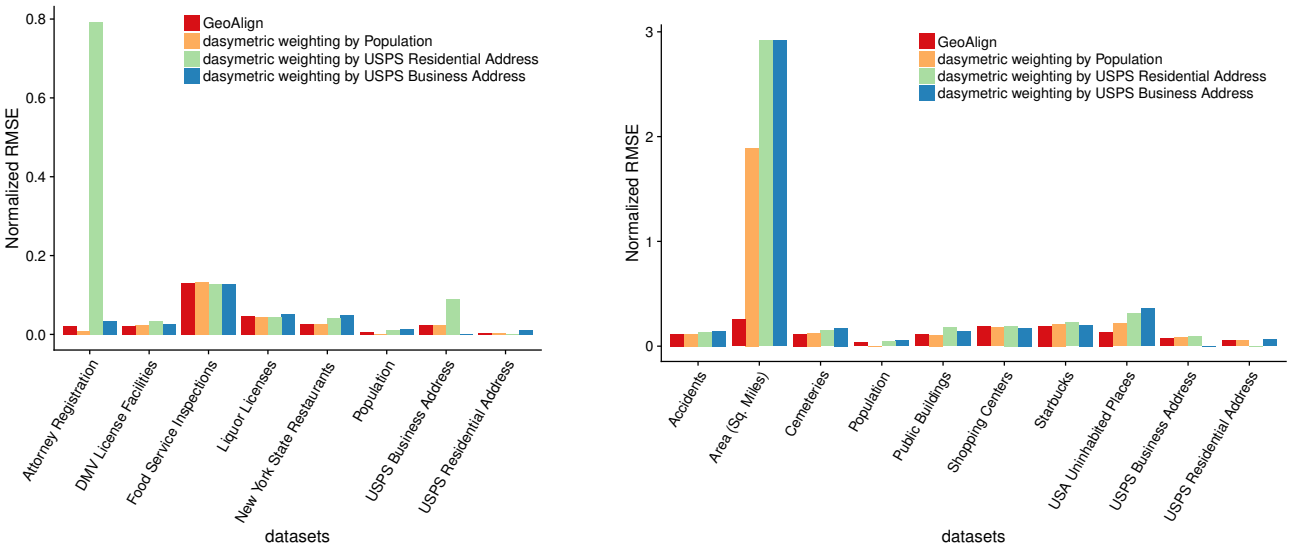
Since the number of datasets with accurate disaggregation matrix is limited, we adopted the cross-validation evaluation method that deals with the problem well. We conducted two series of experiments, one for each universe. More specifically, for each universe, we picked one of the datasets as the test dataset, in turn, and used the remaining datasets to develop crosswalks in GeoAlign whose combined weighted performance is then evaluated for the test dataset. The performance of GeoAlign is compared with the base-line single reference crosswalk method that redistributes by a disaggregation matrix of some known attribute. More specifically, GeoAlign is compared with the areal weighting method and the dasymetric algorithm referencing the three population level datasets. Note that when one of the population reference datasets or the area dataset is used as the test dataset, the performance of both methods referencing this dataset is not evaluated.

4.2 GeoAlign Effectiveness

To evaluate the effectiveness of GeoAlign, we adopted root mean square error (RMSE) as the evaluation criterion that computes the deviation of estimated aggregates from true aggregates of the attribute in counties. To ease the comparison across datasets of heterogeneous scales, in Figure 5, we show the RMSE normalized by the mean of the measured data (NRMSE).

The NRMSE of GeoAlign is compared with that of the dasymetric method using three population level datasets and the areal weighting methods for both New York States (Figure 5a) and the United States (Figure 5b), using eight and ten datasets respectively. The performance of areal weighting method is not shown in the figure since it makes poor predictions for all test datasets: over 15 times of the NRMSE of GeoAlign for New York State experiments and over 50 times of the NRMSE of GeoAlign for the United States experiments.

The NRMSE of GeoAlign is less then 0.13 for New York State experiments and less than 0.26 for the United States experiments.



(a) New York State **(b) the United States**

Figure 5: `GeoAlign` prediction performance (NRMSE) compared with dasymetric methods. Since a better prediction yields a lower NRMSE, `GeoAlign` is making comparable or better predictions than the dasymetric methods for tests in New York State and the Unite States.

Though three dasymetric methods have comparable error on most datasets, for these datasets, `GeoAlign` is making equal or better predictions. It should also be noted that no one of these three methods is predicting uniformly well for all datasets as `GeoAlign` does, in whichever universe. For instance, the dasymetric method referencing the population data presents much higher error than the other methods when predicting for attorney registration and USPS Business Address counts for counties in New York State; all three dasymetric methods fail in accuracy for both area and USA uninhabited places datasets in the United States.

Except the USPS business address dataset, the rest three are individual level datasets with limited number of observational units that are sparsely distributed in the universe. Also, they do not align well with demographic attributes as those in the areal weighting and dasymetric methods. We observe that `GeoAlign` accounts for sparsity and heterogeneous distributions with flexibility.

4.3 `GeoAlign` Efficiency and Scalability

We evaluated the efficiency of `GeoAlign` in terms of algorithm runtime. Apart from the horizontal efficiency comparison across cross-validated tests for a given universe, we also considered the scalability of `GeoAlign` runtime. This is realized by comparing `GeoAlign` efficiency vertically across the universes of different scales.

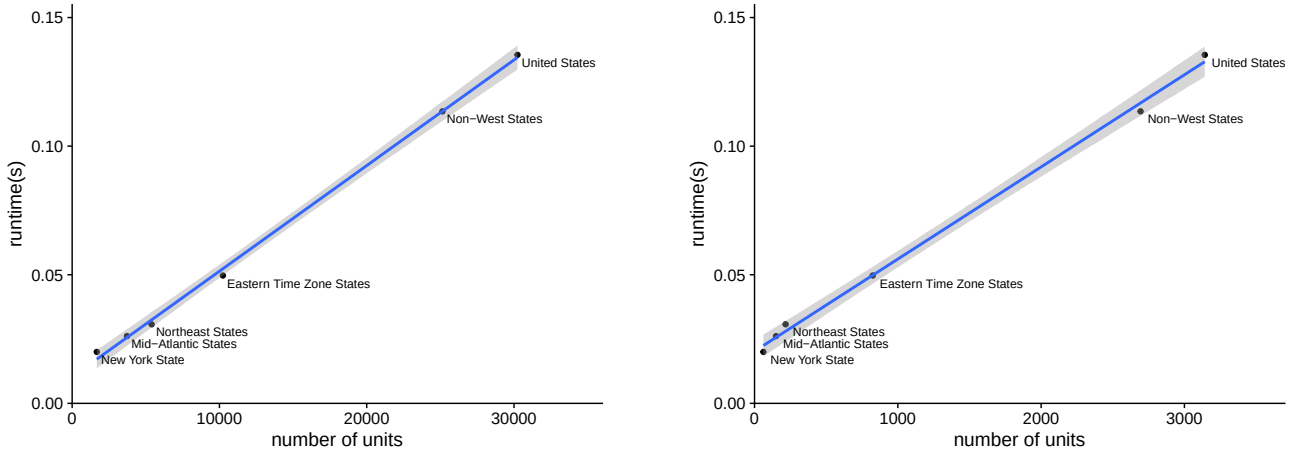
In addition to New York State and the United States, new universes were selected as a set of states whose boundaries are congruent with any other state in the universe. The selection is a greedy process that ensures the states in a universe are tightly connected from a geospatial perspective. These four new universes include Mid-Atlantic division and Northeast region defined by Census Bureau, states contained entirely in the Eastern Time Zone and all states excluding the ones in the Census West Region (non-West). They form a spatial coverage hierarchy preventing the inter-state influence of randomly selected universes.

Moreover, for factor control purpose, instead of collecting more datasets for new universes, for each universe, we subset the ten datasets covering the United States, keeping the entries collected from units within the universe as inputs.

To avoid random error, we averaged the runtime across ten trials for the cross-validated experiments in each universe.

Experimental results show that `GeoAlign` runtime is stable across experiments for the same universe. This is consistent with our claim that the complexity of `GeoAlign` is not related to the magnitude of the count data. The majority of the runtime, over 90%, is spent on computing the disaggregation matrix after the weights are estimated. Note that the aggregate vectors of the objective attribute in source geographic units has the same size for all the different datasets (the size is $|U^s|$). Similarly the aggregate vectors of the reference attributes in source geographic units are all of the same size (all of size $|U^s|$), the aggregate vectors of the reference attributes in target geographic units are all of the same size (all of size $|U^t|$). Further, all the disaggregation matrices are all of the same size as well. The reason for the minor difference in `GeoAlign` runtime for different datasets is because of the difference in the number of non-zero entries in the disaggregation matrix, which is stored as sparse matrix, of reference attributes. For the disaggregation matrix, sparse datasets, such as cemeteries, have less non-zero entries, while dense datasets, such as population, have more non-zero entries. Matrix operations involving sparse matrices are influenced by this factor in SciPy package.

As for cross-universe comparison, we ploted `GeoAlign` runtime versus the number of zip codes (source units) and the number of counties (target units) in Figure 6. These two plots show that `GeoAlign` is fast: it runs for less than 0.15 second even for crosswalk between 30238 zip codes and 3142 counties in the United States universe. They also prove the linear relationship between `GeoAlign` runtime with the number of units in source and target levels since the dominating disaggregation matrix construction operation is linearly related to these two factors.



(a) Zip Code Level

(b) County Level

Figure 6: `GeoAlign` runtime scales linearly with respect to the number of units in source level and target level

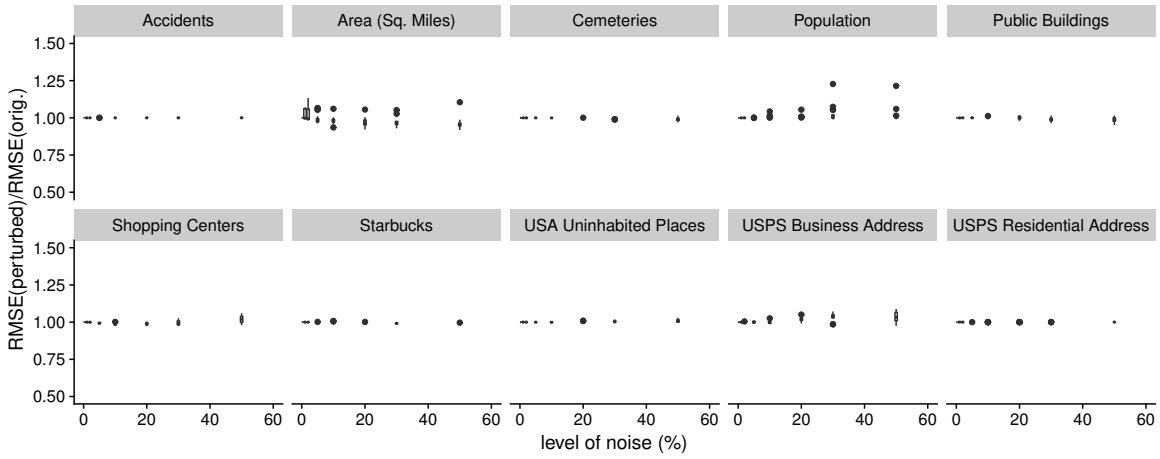


Figure 7: When noises are introduced in references, the prediction deviation is evaluated as the ratio of the RMSE using the perturbed references to the RMSE using the original references. The closer the ratio is to 1, the more invariant `GeoAlign` is to reference noises. For up to 50% level of noise, most experiments have the prediction deviation around 1 indicating the robustness of `GeoAlign` to noisy references.

4.4 `GeoAlign` Robustness

As mentioned earlier in §4.1, during the reference attribute collection process, we encountered two difficulties: the undetermined accuracy of reference attributes at the source level, and the limited availability of datasets with disaggregation matrix. We conducted two series of experiments evaluating the robustness of `GeoAlign` with respect to these two problems respectively.

4.4.1 Inaccurate Reference Attributes. Public aggregated data can be derived in multiple ways. They can be aggregates of individual level data, approximates derived from some crosswalk algorithm, etc. Without the raw data and the transformation information available, the accuracy of these aggregates are unknown. It is thus hard to determine whether the data can be used as references.

To quantitatively evaluate the influence of the accuracy of reference attributes on `GeoAlign`, we artifially introduced "noise" to the reference attributes. We define *noise* as the deviation from the actual value. Noise is measured by "levels" such that a $x\%$ level of noise for y is $\pm x * y/100$. The noise-polluted y is thus $(1 + x/100) * y$. For each of the ten cross-validated experiments in United States, we synthetically generated noisy reference attributes at the source level with 1%, 2%, 5%, 10%, 20%, 30% and 50% degrees of noises for all references. Each experiment is replicated 20 times to account for random error due to randomness of positive or negative noises. We quantify the prediction deviation as the ratio of the RMSE using perturbed noisy reference attributes to the RMSE using the original reference attributes. The closer the prediction deviation is to 1, the smaller the impact of the noises is. `GeoAlign` is making better prediction with the perturbed reference attributes if the ratio is higher than 1; whereas a less than 1 ratio indicates worse prediction with the perturbed reference attributes.

In Figure 7, we show the box plot of the prediction deviation with respect to different levels of noise. The prediction performance of `GeoAlign` is stable across experiments. For each experiment, `GeoAlign` is making robust predictions for all levels of noise. Though for the area and population datasets, higher levels

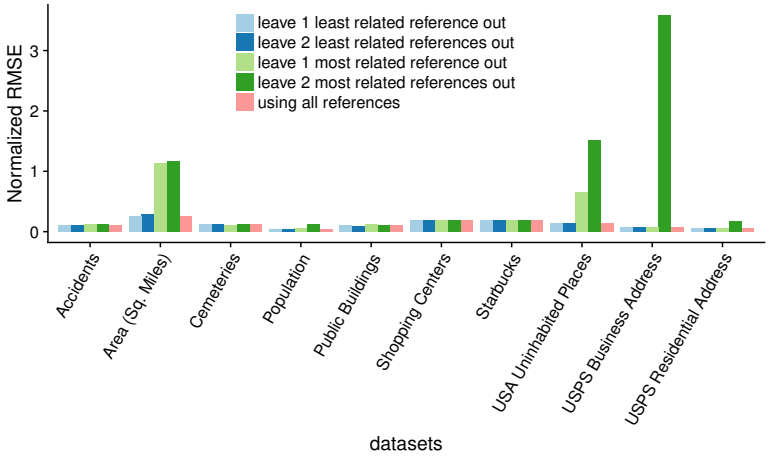


Figure 8: `GeoAlign` is robust to the choice of reference attributes. Though extra reference attributes do not create any loss, reference attributes with higher correlation with the objective are preferred.

of noise resulted in higher prediction error, the mean prediction deviation for these levels is still small (less than 1.1).

4.4.2 Limited Reference Attributes. In general, we cannot predict how many reference attributes will be available. We may have very few, or we may have very many. In the process of reference attribute selection, there are two questions to consider: whether `GeoAlign` can make reasonable predictions with limited number of reference attributes, and how to select the reference attributes when more than one is available.

To answer these two questions, we chose multiple subsets of reference attributes among all reference attributes and repeated the cross-validated experiments for datasets in the United States. The subset of reference attributes were chosen based on their relationship with the target attribute of each test dataset. We adopted the leave-n-out metric such that $n = 1, 2$ for reference attributes with the highest (or lowest) correlation with the target attribute at the source level. The NRMSEs of these four series of experiments are compared with experiments using all reference attributes in Figure 8.

For 7 out of 10 tests, `GeoAlign` is making robust predictions regardless of the subset of reference attributes used. As for the series of experiments leaving 1 or 2 least target-attribute-related reference(s) out, the performance of `GeoAlign` is almost identical to using all reference attributes. This is in accordance with `GeoAlign`'s ability of assigning little weights to reference attributes loosely related to the target attribute.

Leaving out the most target-related attributes out can have an impact on accuracy. This does impact three of our attributes: area, USA uninhabited places and USPS business address datasets. None of the references are closely related to the area and the USA uninhabited places datasets at the source level (correlations less than 0.25). Apart from the two references left out, the rest of the references have even lower correlation with the target attribute (less than 0.2 and 0.05 respectively). According to the assumption basis of `GeoAlign`, the distribution of the target attributes is thus poorly related to the distribution of these attributes, leading to increased prediction error. We also found that leaving out the reference most related to the target attribute has almost no impact on the prediction for the USPS business address dataset; while leaving out top two such references dramatically worsens the situation. Further analysis reveals that these two references are highly correlated with each other at the source level ($\approx 96\%$), the weight assigned to the reference most related to the target attribute is reassigned to the other when the former is left out. This verifies that similar attributes at the source level are also similarly distributed in the intersection units, as the predicted disaggregation matrix of the target attribute is almost the same regardless of using the reference most related to the target attribute or not.

These experiments give us more insight into `GeoAlign` reference attribute selection. `GeoAlign` prefers reference attributes highly related to the target attribute at the source level. For reference attributes poorly related to the target variable, it is able to weigh their contributions accordingly. The reference attributes are not necessarily independent of each other and the reference attributes are not necessarily accurate at the source level. From the user's perspective, `GeoAlign` is able to make reasonable predictions by simply given all available reference attributes.

5 RELATED WORK

In the GIS community, spatial interpolation has advanced from isoline mapping in cartography to data realignment in different units or grids for multivariate analysis in geographic research [3, 31, 38]. *Realignment*, *crosswalk*, or *regridding*, is commonly used today as a preprocessing step before further data analysis in physics and socioeconomics to interpolate spatial or temporal data distribution from one grid to another [28]. Since these data are either point or areal based, two categories of methods are proposed for these two types respectively.

Areal interpolation is a subset of the spatial interpolation problem that realigns aggregates. Early methods built upon point-based interpolation, such as point-in-polygon method, do not follow the volume-preserving property such that reconstruction of exactly the original aggregates of each source unit with the transformed value of each target unit is not possible [31, 44]. It has been shown that these methods are not comparable in approximation efficiency with those that do have the property [31, 47]. Later methods thus introduce the property and turn over to the area-based areal interpolation instead [12]. These approaches depend highly on the spatial properties of the data collection area and thus different forms of ancillary data are introduced ever since.

Areal weighting method, one of the early area-based areal interpolation method, makes use of the area ancillary data available in the form of disaggregation partitions between source and target units [13, 36]. This method is widely available in GIS software for general users nowadays. However, it assumes even distribution within units (homogeneity) whereas this assumption hardly stands in reality. Areal weighting has been extended by referring to other single known reference attributes, called dasymetric weighting [1, 24, 37, 43]. These methods are restricted by the assumption of proportionality of the objective attribute to the single reference attribute. Hence the selection of the reference attribute is vital to the prediction accuracy and the methods are not adaptive to different objective attributes.

The regression methods are later introduced as extensions to the dasymetric methods allowing for multiple auxiliary variables. In general, the regression methods involve a regression of the source level data of the objective attribute on the values of the references in target units. For this track of methods, more advanced techniques such as EM algorithm, Monte Carlo simulation, smoothing techniques[9, 11, 31, 45, 46, 48], etc., are introduced later in the literature. However, they make different assumptions of density distribution within units, some of the mostly used ones are Poisson distribution and binomial distribution, and their performances are rather assumption dependent [30] and auxiliary variable dependent. Recently, more complicated regression models [35, 39, 40] are developed based on domain knowledge such as spatial correlation. However, they lack general applicability to heterogeneous target attributes and are hard to implement for practitioners.

These approaches can also be categorized as extensive or intensive approaches based on their approximation target. Extensive approaches approximate a_o^{st} while intensive ones approximate f_o^{st}. Most approaches for solving the areal interpolation problem are intensive approaches that build spatial statistical models for f_o^{st} in the disaggregation step. These approaches, mostly developed in 2-D space, can be extended to higher dimensions, though these extensions are typically non-trivial. Other major limitations of intensive approaches include narrow scope of application and low robustness to heterogeneous objective attributes.

Current intensive approaches for areal interpolation are not generally applicable for aggregate interpolation due to three main reasons. First, integration of f_o^{st} is computable in 2-D, however, it is computationally intensive in high dimensions with complex f_o^{st}. Second, shape files are indispensable for intensive approaches, and the probability density function for each intersection unit, $f_o^{st}[k]$, is associated with the shape files of source and/or intersection units. Further, attributes in plain tables without handy shape files of target units typically fail re-aggregation. Even if shape files are available, some of them constantly change over time, resulting in approximation inaccuracies. Last but not least, these approaches are not easily approachable for general users, especially those with little technical proficiency in mathematics, statistics and GIS. The $\hat{f}_o^{st}$ model is built upon the spatial knowledge of the objective attribute; however, this knowledge is not available for all users. Further, implementations of intensive approaches are not publicly available, making them even harder to use.

Another limitation of intensive approaches is that they are not adaptive to new attributes. $\hat{f}_o^{st}$ models are attribute dependent since the true f_o^{st} models for two attributes can be very different. Another point to note is that these approaches make many assumptions of $\hat{f}_o^{st}$. For instance, the distribution model of each intersection unit, the choice of parameters for these distributions and so on. Any change in these assumptions may dramatically influence the accuracy of approximation in some target unit. What is worse, there is no efficient verification of whether they are appropriate or not.

Extensive approaches are more generally applicable than the intensive ones: they can be easily extended to high dimensions, need no unit shape files, and are easy to implement. However, existing extensive approaches make use of a single reference attribute and are still limited in robustness. When the objective attribute and the reference attribute does not share similar spatial distribution, the approximated result can differ substantially from the true aggregates in target units. Further, since they use the same reference attribute irrespective of the objective attribute, they are not adaptive to different objective attributes with heterogeneous spatial distributions.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we formally define the problem of aggregate interpolation in multi-dimensional space and propose `GeoAlign`, an adaptive multi-reference algorithm that realigns aggregates better than state-of-the-art approaches for real socioeconomic datasets. Unlike existing areal interpolation algorithms, `GeoAlign` requires no knowledge of spatial properties or dasymetric maps of source and target units and is thus generally applicable for plain aggregate tables. Our experiments show that `GeoAlign` is making better predictions in a reasonably short time. Its runtime scales linearly with the number of units in source and target levels, and is robust to noisy references even when limited references are available.

A potential future direction is to extend this work into an automatic aggregate data integration system that joins multiple aggregate tables without user intervention.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grants ACI-1640575, IIS-1250880, and IIS-1743088.

REFERENCES

[1] Branislav Bajat, Nikola Krunić, and Milan Kilibarda. 2011. Dasymetric Mapping of spatial distribution of population in Timok Region. In *Proceedings of International conference Professional practice and education in geodesy and related fields, Klavodo-Djerdap, Serbia.*

[2] Zohra Bellahsene, Angela Bonifati, Erhard Rahm, and others. 2011. *Schema matching and Mapping*. Vol. 20. Springer.

[3] Peter A Burrough, Rachael McDonnell, Rachael A McDonnell, and Christopher D Lloyd. 2015. *Principles of geographical information systems*. Oxford University Press.

[4] United States Census. 2010. 2010 Census Data [Data file]. Available from https://www.census.gov/2010census/data/. (2010). Accessed: 2014-08-14.

[5] NY Open Data. 2013. Facilities Licensed by the Department of Motor Vehicles (DMV License Facilities) [Data file]. Retrieved from https://data.ny.gov/Transportation/Facilities-Licensed-by-the-Department-of-Motor-Veh/nhjr-rpi2. (2013). Accessed: 2017-05-01.

[6] NY Open Data. 2013. Food Service Establishment Inspections: Beginning 2005 (ACTIVE) (Food Service Inspections) [Data file]. Retrieved from https://health.data.ny.gov/Health/Food-Service-Establishment-Inspections-Beginning-2/2hcc-shji. (2013). Accessed: 2014-08-14.

[7] NY Open Data. 2013. Liquor Authority Quarterly List of Active Licenses (Liquor Licenses) [Data file]. Retrieved from https://data.ny.gov/Economic-Development/Liquor-Authority-Quarterly-List-of-Active-Licenses/hrvs-fxs2. (2013). Accessed: 2014-08-14.

[8] NY Open Data. 2013. NYS Attorney Registrations (Attorney Registration) [Data file]. Retrieved from https://data.ny.gov/Transparency/NYS-Attorney-Registrations/eqw2-r5nb. (2013). Accessed: 2017-05-01.

[9] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* (1977), 1–38.

[10] Cory L Eicher and Cynthia A Brewer. 2001. Dasymetric Mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science* 28, 2 (2001), 125–138.

[11] Cory L. Eicher and Cynthia A. Brewer. 2001. Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science* 28, 2 (2001), 125–138.

[12] Robin Flowerdew and Mick Green. 1993. *Developments in areal interpolation methods and GIS.* Springer Berlin Heidelberg, Berlin, Heidelberg, 73–84.

[13] Robin Flowerdew, Mick Green, and S Fotheringham. 1994. Areal interpolation and types of data. *Spatial analysis and GIS* 121 (1994), 145.

[14] Carl Franklin. 1992. An Introduction to Geographic Information Systems: Linking Maps to Databases. *Database* 15, 2 (April 1992), 12–21.

[15] Esri ArcGIS Gallery. 2014. Starbucks [Map]. Retrieved from https://services.arcgis.com/nzS0F0zdNLvs7nc8/arcgis/rest/services/Starbucks/FeatureServer. (2014). Accessed: 2017-10-02.

[16] Esri ArcGIS Gallery. 2017. Accidents reported to National Highwat Traffic Safety Administration in 2011 (Accidents) [Map]. Retrieved from http://services.arcgis.com/0TU5BETrBlnIhvOx/ArcGIS/rest/services/NHTSAAccidents2011/FeatureServer. (2017). Accessed: 2017-10-02.

[17] Esri ArcGIS Gallery. 2017. US Shopping Centers 2015 (Shopping Centers) [Map]. Retrieved from https://services1.arcgis.com/6kyLQ3wRvoPKn52I/arcgis/rest/services/US_Shopping_Centers_2015/FeatureServer. (2017). Accessed: 2017-10-02.

[18] Esri ArcGIS Gallery. 2017. USA Cemeteries (Cemeteries) [Map]. Retrieved from http://www.arcgis.com/home/item.html?id=5b08fa8bb5a64ea7848dc5188e47994a. (2017). Accessed: 2017-10-02.

[19] Esri ArcGIS Gallery. 2017. USA Public Buildings (Public Buildings) [Map]. Retrieved from http://www.arcgis.com/home/item.html?id=d5d5b513a40145ffa60b67d9c7ab9680. (2017). Accessed: 2017-10-02.

[20] Esri ArcGIS Gallery. 2017. USA Uninhabited Places [Map]. Retrieved from http://www.arcgis.com/home/item.html?id=5f0a5776cbaf4b34b9600809bf791d69. (2017). Accessed: 2017-10-02.

[21] Lise Getoor and Ashwin Machanavajjhala. 2012. Entity resolution: theory, practice & open challenges. *Proceedings of the VLDB Endowment* 5, 12 (2012), 2018–2019.

[22] Michael F Goodchild, Luc Anselin, and Uwe Deichmann. 1993. A framework for the areal interpolation of socioeconomic data. *Environment and planning A* 25, 3 (1993), 383–397.

[23] Michael F Goodchild, Nina Siu Ngan Lam, and University of Western Ontario. Dept. of Geography. 1980. *Areal interpolation: a variant of the traditional spatial problem.* London, Ont.: Department of Geography, University of Western Ontario.

[24] I. Gregory. 2002. The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems* 26, 4 (7 2002), 293–314.

[25] Alon Halevy, Anand Rajaraman, and Joann Ordille. 2006. Data integration: the teenage years. In *Proceedings of the 32nd international conference on Very large data bases.* VLDB Endowment, 9–16.

[26] Mauricio A Hernández, Renée J Miller, and Laura M Haas. 2001. Clio: A semi-automatic tool for schema Mapping. *ACM SIGMOD Record* 30, 2 (2001), 607.

[27] Esri Inc. 2017. ArcGIS Pro 2.0 [Computer Software]. Available from https://pro.arcgis.com/en/pro-app/. (2017).

[28] Karen Kemp. 2008. *Encyclopedia of geographic information science.* Sage.

[29] Hanna Köpcke and Erhard Rahm. 2010. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering* 69, 2 (2010), 197–210.

[30] Nina Siu-ngan Lam. 1982. An evaluation of areal interpolation methods. In *Proceedings, Fifth International Symposium on Computer-Assisted Cartography (AutoCarto 5)*, Vol. 2. 471–479.

[31] Nina Siu-Ngan Lam. 1983. Spatial Interpolation Methods: A Review. *The American Cartographer* 10, 2 (1983), 129–150.

[32] Mitchel Langford. 2006. Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, environment and urban systems* 30, 2 (2006), 161–180.

[33] Mitchel Langford, David J Maguire, and David J Unwin. 1991. The areal interpolation problem: estimating population using remote sensing in a GIS framework. *Handling geographical information: Methodology and potential applications* (1991), 55–77.

[34] Maurizio Lenzerini. 2002. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.* ACM, 233–246.

[35] X. H. Liu, P. C. Kyriakidis, and M. F. Goodchild. 2008. Population-density Estimation Using Regression and Area-to-point Residual Kriging. *Int. J. Geogr. Inf. Sci.* 22, 4 (Jan. 2008), 431–447.

[36] John Markoff and Gilbert Shapiro. 1973. The Linkage of Data Describing Overlapping Geographical Units. *Historical Methods Newsletter* 7, 1 (1973), 34–46.

[37] Jeremy Mennis and Torrin Hultgren. 2006. Intelligent Dasymetric Mapping and Its Application to Areal Interpolation. *Cartography and Geographic Information Science* 33, 3 (2006), 179–194.

[38] Lubos Mitas and Helena Mitasova. 1999. Spatial interpolation. *Geographical information systems: principles, techniques, management and applications* 1 (1999), 481–492.

[39] Andrew S. Mugglin, Bradley P. Carlin, and Alan E. Gelfand. 2000. Fully Model-Based Approaches for Spatially Misaligned Data. *J. Amer. Statist. Assoc.* 95, 451 (2000), 877–887.

[40] Daisuke Murakami and Morito Tsutsumi. 2011. A new areal interpolation method based on spatial statistics. *Procedia-Social and Behavioral Sciences* 21 (2011), 230–239.

[41] Office of Policy Development and Research (PD & R). 2010. HUD USPS Zip Code Crosswalk Files [Data file]. Available from https://www.huduser.gov/portal/datasets/usps_crosswalk.html. (2010). Accessed: 2014-08-14.

[42] Erhard Rahm and Philip A. Bernstein. 2001. A survey of approaches to automatic schema matching. *The VLDB Journal* 10, 4 (2001), 334–350.

[43] Michael Reibel and Michael E Bufalino. 2005. Street-Weighted Interpolation Techniques for Demographic Count Estimation in Incompatible Zone Systems. *Environment and Planning A* 37, 1 (2005), 127–139.

[44] Yukio Sadahiro. 2000. Accuracy of Count Data Estimated by the Point-in-Polygon Method. *Geographical Analysis* 32, 1 (2000), 64–89.

[45] Guillermo Q. Tabios and Jose D. Salas. 1985. A comparative analysis of techniques for spatial interpolation of precipitation. *JAWRA Journal of the American Water Resources Association* 21, 3 (1985), 365–380.

[46] Waldo R Tobler. 1979. Smooth pycnophylactic interpolation for geographical regions. *J. Amer. Statist. Assoc.* 74, 367 (1979), 519–530.

[47] Paul R Voss, David Dryden Long, and Roger Bruce Hammer. *When census geography doesn't work: Using ancillary information to improve the spatial interpolation of demographic data.*

[48] John K Wright. 1936. A method of Mapping densities of population: With Cape Cod as an example. *Geographical Review* 26, 1 (1936), 103–110.

[49] Yichun Xie. 1995. The overlaid network algorithms for areal interpolation problem. *Computers, environment and urban systems* 19, 4 (1995), 287–306.