Journal of Geographical Systems manuscript No.
(will be inserted by the editor)

Comparison of Approaches of Geographic Partitioning for Data Anonymization

William Lee Croft[1] · **Wei Shi**[2] · **Jörg-Rüdiger Sack**[1] · **Jean-Pierre Corriveau**[1]

Received: date / Accepted: date

Abstract Given the large volumes of detailed data now being collected, there is a high demand for the release of this data for research purposes. In particular, organizations are faced with the conflicting goals of a) releasing this data and b) protecting the privacy of the individuals to whom the data pertains. Especially, there is a conflict between the need to release precise geographic information (which is essential to many health care research fields such as spatial epidemiology) and the requirement to censor or generalize the same information for the sake of privacy protection. Ultimately, the challenge is to anonymize data in order to comply with government privacy policies while reducing the loss in geographic information as much as possible. In this paper, we present novel component approaches used to configure the Voronoi-Based Aggregation System (VBAS) as well as an in-depth comparison of their effectiveness. VBAS is a system which protects privacy by enforcing k-anonymity via the aggregation of regions of fine granularity into larger regions. We additionally discuss heuristics rooted in linear programming which we have also integrated in our system. Based on extensive comparisons, we highlight the strengths and weaknesses of the different approaches we tested. This enables us to make recommendations on how to satisfy user requirements via the selection of specific combinations of such approaches.

Keywords Geographic Partitioning · Data Anonymization · Health Care

Mathematics Subject Classification (2000) MSC 68U01

1: School of Computer Science, Carleton University
1125 Colonel By Drive, Ottawa, Canada

2: School of Information Technology, Carleton University
1125 Colonel By Drive, Ottawa, Canada

W. L. Croft **(Corresponding author)** - lee.croft@carleton.ca
W. Shi - weishi@cunet.carleton.ca
J.-R. Sack - sack@scs.carleton.ca
J.-P. Corriveau - jeanpier@scs.carleton.ca

1 Background

1.1 Introduction

Relevant and detailed data sets are critical for effective population-based research. As such, they are in high demand. However, since this data is of a sensitive nature, the privacy of individuals must be protected when data is released (Arzberger et al, 2004; Benitez and Malin, 2010; Emam et al, 2013; Lowrance, 2006; Samarati, 2001). Government policies place restrictions on how data can be released in order to ensure that privacy will be maintained. Thus, in order for a data set to be released, it must undergo a process of *anonymization* that brings it to a state in which the risk of disclosure of sensitive information is sufficiently low.

Although any directly identifying information can be trivially stripped from a data set, there is still a susceptibility to re-identification through techniques such as cross-referencing (Emam et al, 2013). Moreover, there will always be a trade-off between the level of protection that can be achieved on a data set and the resultant utility of the data (Gionis and Tassa, 2008). Furthermore, although it is desirable to minimize the loss of any type of information in the data set, in some cases the preservation of geographic information may be of particular importance. For example, studies concerned with the propagation of diseases across geographic areas require a high level of precision in the geographic information of the data set (Lyseen et al, 2014). In fact, any form of location-critical research such as spatial epidemiology requires high precision geographic information in order to be carried out (Rezaeian et al, 2007; Vora et al, 2008). Similarly, studies based on socioeconomic characteristics of areas may also call for regions of finer precision than those given by standard administrative boundaries. For example, some studies require regions at the neighborhood or community level (Thomas et al, 2008). More generally, in some cases, research questions require customizable boundaries for regions (Young et al, 2009). This can be problematic as the level of customizability may already be limited by the granularity of the geographic information in the original data set: either the user must take on the undesirable burden of defining all regional boundaries or the anonymization process must in some way handle this. In the absence of an anonymization process which can do this, an alternative is to provide finely grained regions.

The release of precise geographic details, however, greatly increases the risk of disclosure of sensitive information due to higher levels of distinctness in the records of the data set. This risk creates a barrier in the disclosure of essential geographic information. There is, therefore, a need for the careful consideration of the geographic information in a data set during anonymization if a high level of precision is to be retained. Since losses in geographic information may have negative effects on the ability to effectively analyze a data set (Olson et al, 2002), we postulate it is desirable to preserve as much geographic information as possible while maintaining an appropriate level of anonymity.

We present a set of approaches which can be supplied as different strategies for the components of the Voronoi-Based Aggregation System (VBAS) in order to provide different system configurations. VBAS is a system which anonymizes a data set through the application of geographic partitioning guided by the use of *Voronoi diagrams* (Aurenhammer and Klein, 2000). This system achieves a required level of anonymity (as specified by user input) in a data set through the

generalization (i.e., coarsening of the level of precision) of geographic attributes and the suppression of outlying records (i.e., records that violate the required level of anonymity). By aggregating regions, we avoid the need for the suppression of small regions that do not meet a required population size. This is desirable as the suppression of complete regions can lead to heavily censored data sets (Emam et al, 2009; Hawala, 2001). Furthermore, this form of aggregation enables us to maintain a higher degree of geographic precision than other methods such as cropping (Jung, H.-W. and Emam, K. E., 2014).

In a companion paper (Croft et al, 2016), we have selected a particular configuration of VBAS and have conducted a comparison of it with another system of geographic anonymization, GeoLeader, to show that VBAS is able to perform well in comparison to other systems in this field. Here, we study a number of different approaches that can be applied within VBAS in order to compare their performance with respect to the preservation of geographic precision, reduction of suppression and reduction of information loss. Through the use of appropriate approaches (to be discussed later in this paper), VBAS is able to provide a means to aggregate small regions of fine granularity that are geographically close to each other into larger regions that satisfy criteria for achieving a sufficient level of anonymity while maintaining a low level of geographic information loss. This technique may be used in combination with the generalization of other attributes (such as age and ethnicity) in order to trade off precision in other, less important attributes for an even higher level of geographic precision. In this work, we focus our efforts solely on the generalization of the geographic attribute.

1.2 Literature Review

In order to protect the privacy of the individuals in data sets, the data must often be de-identified before it can be considered safe for release. This process of de-identification is intended to protect against the risk of the data being re-identified and revealing sensitive information about specific individuals (Benitez and Malin, 2010; Emam et al, 2013; Lowrance, 2006; Samarati, 2001). Methods of de-identification generally involve the creation of *equivalence classes* with sufficient cardinalities in order to protect against re-identification. An equivalence class is a group of records which are indistinguishable from each other based on their *quasi-identifiers* (i.e., demographic-type attributes). Since a malicious party attempting re-identification will need this demographic information to perform cross-referencing, raising equivalence class cardinalities has the effect of lowering the malicious party's chances at making successful inferences (Benitez and Malin, 2010; Emam et al, 2013; Samarati, 2001).

Generalization and *suppression* are techniques which are commonly employed to raise equivalence class cardinalities. Generalization coarsens the response categories of the quasi-identifiers in order to cause more records to fall into the same equivalence classes (Mohammed et al, 2009; Sweeney, 2002). Suppression completely removes records from the data set. This is done in order to avoid producing data sets with outlying records leading to the need for severe generalization in order to reach sufficient equivalence class cardinalities (Mohammed et al, 2009; Sweeney, 2002). Through the combination of these two techniques, a guarantee can be provided on the resultant data set in the form of *k-anonymity* (Mohammed et al,

2009; Sweeney, 2002). Given a user specified value for k as input, k-anonymity is the guarantee that every resultant equivalence class has a cardinality of at least k. By selecting an appropriate k value and ensuring that all equivalence classes meet this requirement, it becomes much more difficult for a party to re-identify the data as each record has a small group of other records from which they are indistinguishable.

Within the context of data with geographic information, an alternative to generalization and suppression is perturbation-based geomasking. This involves the application of geographic perturbations in order to produce a data set with the geographic information modified such that it can no longer be effectively used for re-identification (Armstrong et al, 1999; Bridwell, 2007; Clifton and Gehrke, 2013). While such geomasking approaches have been shown to produce data sets with high levels of utility for spatial analysis (Armstrong et al, 1999), they do not allow for the consideration of other non-geographic attributes and cannot be used with the traditional equivalence class-based measurement of k-anonymity. We therefore investigate alternative geographic-based methods for anonymity.

One strategy that can be applied to anonymize a data set is to focus on the population sizes of the geographic regions into which the data records are grouped. Since the reduction of distinctness is one method that can be used to protect privacy, records can be grouped together into larger regions in order to achieve this. When the set of geographic attributes (hereafter referred to as the geographic identifier) of a data set is fine-grained, records will be grouped into very small regions, preventing the creation of equivalence classes of high cardinalities since the geographic identifier is part of the quasi-identifiers. The coarsening of this geographic granularity therefore enables the cardinalities to become higher. This is essentially a form of generalization applied only to the geographic identifier.

The main difficulties with generalization of the geographic identifier are determining what an appropriate population size is and determining where region borders should be placed in order to produce an appropriate regionalization. In addition to the geographic identifier being two-dimensional, the distribution of the demographic population density across the geographic space is non-homogeneous. The creation of regions which will lead to appropriate equivalence classes is therefore not a simple task. Some basic approaches involving the use of existing regionalizations and static population cut-offs sizes have been proposed (Hawala, 2001; Greenberg and Voshell, 1990; Jung, H.-W. and Emam, K. E., 2014), however these approaches do not sufficiently address these issues and thus result in a greater loss of geographic precision than necessary (Boulos et al, 2006; Emam et al, 2009).

For a more complete review and discussion of these topics, we refer to the literature review in our companion paper (Croft et al, 2016).

With respect to determining an appropriate population size for the regions, a method of dynamically computing this value based on the input data set has been proposed (Emam et al, 2009). By analyzing the quasi-identifiers in the input data set, an approximation of a desirable population cut-off size is computed, thus reducing the loss of geographic precision that occurs when a static cut-off size which is larger than necessary is employed.

For the issue of determining the borders of the regions, once possible approach is to use linear programming to determine regions which should be aggregated. If the goal is to create regions which have a population above a certain level, this can be enforced through constraints. Various heuristics exist for this type of

problem such as the "location-allocation" heuristic (Goodchild and Massam, 1969) and the "transportation-location" heuristic (Cooper, 1972). In this problem, the input is a set of fixed destination locations, each with supply requirements, and a number of facilities, each with supply levels, to be placed. The goal is to select locations for the facilities and determine an allocation of supplies from facilities to destinations such that all constraints are satisfied (namely, each destination receives its required amount of supplies and each facility does not ship out more than its supply level) while an objective function related to transportation costs is minimized. This problem can be seen as a generalization of the transportation problem (Ford and Fulkerson, 1956) in which the facility locations are already determined. We will discuss those alternative heuristics in some detail (see Section 2.3.2) and analyze/compare them with our methods.

1.3 Problem Statement

The problem of anonymization of data can be seen as a problem with one or more hard constraints and various objective functions to be optimized. The constraints correspond to any desired guarantees with respect to anonymity. In our work, we employ a constraint that ensures that the resultant data set is k-anonymous. The objective functions correspond to the desirable traits of the anonymized data. In a typical setting for anonymization, these usually correspond to functions that measure levels of suppression and information loss. In our work, geographic precision is also an important factor, thus we consider as well an objective function that measures the compactness of aggregated regions.

When faced with multiple objective functions, one must find a way to address all of them. Possible strategies to achieve this include multi-objective optimization (e.g., as described in (Zhou et al, 2011)) or the amalgamation of these objective functions into a single function (e.g., via a weighted sum). Interdependencies between these functions and thus trade-offs are unavoidable. Such trade-offs are even more complex to understand if the importance of each objective is user-specific. Consequently, trying to obtain a solution akin to a Nash equilibrium for such objective functions is extremely challenging. Instead, in this paper, we study different approaches of anonymization in order to compare their performance with respect to various objectives. Though our main concern is the compactness of regions, we also consider the suppression of records, information loss and running time. The comparison of these approaches may guide a user in selecting the most suitable approach for their individual requirements. Furthermore, observations made on the relationships between the objectives can provide insight for further studies into methods to effectively optimize the various objectives according to the user requirements.

1.4 Contributions

In this paper, we present a number of new approaches for the components of VBAS. The anonymization strategy supported by VBAS consists of a sequence of 4 tasks, as explained in the next section, each task corresponding to one of the 4 main system components. In order to achieve each of the first 3 of these tasks, the

user chooses from several available approaches, each supported by a distinct actual module (which is easily substitutable with any of the other modules realizing this task). A domain expert may also easily develop, integrate and test new approaches for each of these tasks. Thus, different configurations of actual modules for the components of VBAS can be readily explored. Such configurability allows the end-user to fine-tune aggregation towards their requirements. It also enables domain experts to compare the advantages and limitations not only of different approaches to a task, but also of different sets of approaches supporting the anonymization strategy of VBAS.

Additionally, we have adapted linear programming algorithms to compare with two of our system components. We have implemented VBAS as a desktop application along with all algorithms discussed in this paper in order to test and compare all combinations of the proposed approaches.

Such an extensive evaluation leads to the identification and better understanding of trade-offs between different objectives. In turn, this allows us to recommend which approach to use for each of the first 3 of tasks of VBAS given a set of user requirements. Ultimately, by considering the most promising combinations of approaches, we can further assist the user by making recommendations as to which complete system configuration (i.e., set of approaches) to use given a set of user requirements. Drawing on these recommendations, we believe even a user with minimal knowledge of the anonymization process will be able to use VBAS in an effective manner.

2 Methods

VBAS is designed to anonymize a data set by performing aggregation on an initial regionalization of fine granularity such that the aggregated regions will have sufficient levels of anonymity. To group the initial regions, we construct a Voronoi diagram on top of them. The initial regions are then grouped together based on the Voronoi regions in which they fall. Due to the nature of this grouping process, initial regions of roughly globular or non-elongated shape will be the most beneficial in the creation of aggregated regions of high compactness and geographic precision. This process avoids the need for suppression of small regions and does not require any type of predefined generalization hierarchy. A series of screenshots depicting the process run by VBAS can be seen in Appendix A - Figures 1-4. For a complete description of the VBAS aggregation methodology, we refer to the companion paper (Croft et al, 2016). Here, we focus on the individual approaches used within VBAS to achieve the overall process.

In order for the groupings of initial regions to produce an aggregated regionalization of desirable qualities, it is essential to carefully select the number of sites for the Voronoi diagram, as well as their locations. The complete process is therefore broken up into four main components:

- Approximation of the number of sites
- Selection of the site locations
- Aggregation of the regions
- Evaluation of the aggregation

Each component may be supplied with any approach that is able to complete the component's task. The system is set up in this way to allow for an ease of configuration through the selection of different component approaches. This serves as a benefit both when selecting appropriate approaches for anonymization in practice as well as for testing different approaches and their combinations. In the following subsections, we provide a detailed description of the proposed approaches which can be adopted for the components.

2.1 Site Number Approximation

The task of the first component is to select an appropriate number of sites to be used for the Voronoi diagram. Since each site will produce a single Voronoi region, the number of sites can be thought of as the number of aggregated regions that will be created. This number must be carefully selected. An approximation that is too high will result in a large number of aggregated regions, leaving the records spread too thin and resulting in levels of anonymity that remain too low. Alternatively, if the approximation is too low, there will be very few aggregated regions and their levels of anonymity will be greater than necessary, resulting in a greater loss of geographic precision than necessary.

We present here two different approaches for the site number approximation, both of which are derived from different models for an approximation of a dynamic Geographic Area Population Size (GAPS) cutoff (Emam et al, 2009) for the input data set. The dynamic GAPS cutoff models are intended to serve as a method to compute the required population cutoff size for any given data set based on its quasi-identifiers. This avoids the need to study each data set individually in order to manually determine the cutoff size.

The dynamic GAPS method has one model to compute the cutoff size based on the entropy of the data set and an alternative model to compute the cutoff size based on a max combinations value calculated from the quasi-identifiers (as explained below). The models used here are shown in Table 1.

Table 1 Regional GAPS Cutoff Models (Emam et al, 2009)

Region	Entropy Model	MaxCombs Model
Western Canada	$1588(Entropy^{0.42})$	$1588(MaxCombs^{0.42})$
Central Canada	$1436(Entropy^{0.43})$	$1436(MaxCombs^{0.43})$
Eastern Canada	$1978(Entropy^{0.304})$	$1978(MaxCombs^{0.304})$

The entries show the formulae used for each of the GAPS models for the 3 regions of Canada that were studied.

We have adapted the two dynamic GAPS cutoff models into site number approximation approaches by using the cutoff size as an approximation of a desirable average population for the aggregated regions. By dividing the total population of the data set by the cutoff size, we are able to make an approximation of the number of aggregated regions needed to achieve this cutoff as the average population. Since the number of aggregated regions is equal to the number of Voronoi sites, this serves as the approximation for the number of sites to place.

Entropy Approach The entropy model requires the entropy of the input data set to first be computed using the calculation shown in Equation 1. This value can then

be plugged into the entropy model as shown in Equation 2 in order to compute the cutoff size. Finally, we use this value to approximate the number of Voronoi sites in Equation 3.

Let: L be the size of the largest equivalence class
t_k be the number of equivalence classes of size k
N be the total number of records in the data set

$$Entropy = -\sum_{k=1}^{L} t_k \left(\frac{k}{N}\right) \left(\log \frac{k}{N}\right) \tag{1}$$

$$Cutoff = e^{B_0} \left(Entropy^{B_1}\right) \tag{2}$$

$$Sites = \frac{N}{Cutoff} \tag{3}$$

Max Combinations Approach The approach using the max combinations model is very similar to that of the entropy model; the only difference is in the computation of the max combinations value. This value is the total number of equivalence classes in the data set and is calculated as the product of the numbers of response categories for each quasi-identifier as shown in Equation 4. Once this value is calculated, it can be used to approximate the number of sites with Equations 2 and 3 by substituting the entropy value with the max combinations value.

Let: Q be the set of quasi-identifiers in the data set
$|q|$ be the number of response categories in a quasi-identifier q

$$MaxCombs = \prod_{q \in Q} |q| \tag{4}$$

2.2 Site Location Selection

Once the number of Voronoi sites has been selected, the next task is to select the locations at which to place them. The selection of these locations also has a large influence on the levels of anonymity in the aggregated regions as well as the amount of information that is lost during aggregation. It is easy to see that a dense cluster of sites placed in a region of very low population density would result in aggregated regions with very low populations, causing very low levels of anonymity. Additionally, the locations of the sites with respect to each other determine the shape and size of the Voronoi regions. These properties of the regions determine the level of precision in the geographic information that is released.

For this component, we provide two different approaches that can be applied.

2.2.1 Balanced Density

The goal of the balanced density approach is to divide the plane into a set of cells such that the number of cells is equal to the number of sites to be placed and each cell has roughly the same population within it. Each cell will then be assigned a

single site to be placed at the median of the initial region points that fall within the cell. For ease of organization, the cells are grouped together into rows such that all cells in a row have the same upper and lower boundaries (those of the row) and occupy the entire space covered by the row. The cells are given roughly the same populations in order to make the distribution of the Voronoi sites similar to the distribution of the population.

The approach works by making an initial approximation for an appropriate number of rows which can later be adjusted if necessary. The boundaries of the rows are then determined based on the global population distribution in order to produce rows which each have roughly the same population size. Based on the number of rows, an approximation of the required number of cells per row is then made in order to produce a total number of cells equal to the number of Voronoi sites. Since the number of Voronoi sites over the number of rows is unlikely to be an integer, rows are not required to each have the same number of cells. Some rows may have one less cell or one more cell than the approximated number. As this approach was applied in the companion paper (Croft et al, 2016), we refer to that paper for the fully detailed description.

2.2.2 Anonymity-Driven Clustering

The Anonymity-Driven Clustering (ADC) approach selects site locations as the resultant locations of cluster centers after running a process of iterative cluster optimization based on the framework of the k-means algorithm. In order to adapt this to create clusters that suit our needs, it was necessary to design clustering criteria based on levels of anonymity. As such, the following modifications were made to the algorithm:

1. An objective function that measures levels of anonymity is employed.
2. The optimization step has been redesigned to improve anonymity.
3. The convergence criteria has been modified to accommodate these changes.

The initial region points are provided as the input point set for the algorithm. The Voronoi site locations taken as the output of the algorithm are determined by the locations of the cluster centers at the time of convergence. The clusters, as determined by k-means membership (where each point belongs to the nearest cluster center), are particularly useful in this context as membership is determined by the Voronoi diagram in the same way. This means that when providing the final cluster centers as sites to the Voronoi diagram algorithm, the points in each cluster are exactly the points that will be grouped together by the Voronoi region that pertains to the site that was the center of the cluster. In other words, each cluster of points accurately represents an aggregated region. This fact allows for the ability to evaluate at any time the quality of the aggregation represented by the current clusters.

It should be noted that the selection of the initial cluster centers has an impact on the quality of the results produced. While the clustering can be run by selecting the initial centers at random, it is recommended to use another site location approach as a seeding method for the initial centers.

Anonymity-Based Objective Function

The objective function must evaluate the quality of the aggregation in order to improve it during the clustering process. To do this, we consider the levels of anonymity in each cluster as well as the equivalence classes sitting at the lowest level of anonymity.

Although the global anonymity is determined by the lowest anonymity across these regions, an increase in local anonymity can still reduce the amount of suppression that is later needed. By taking the sum of the local levels of anonymity, we have a function that changes monotonically as the levels of anonymity improve.

This sum can be used as the dominating factor for the objective function, however most moves are not likely to improve the level of anonymity of a whole region. They are more likely to improve the anonymity of one or more equivalence classes. Furthermore, some moves may decrease the anonymity of other equivalence classes. We need to be able to track these changes as well. To do this, we use an objective function which is dominated by local levels of anonymity but is also influenced by the ratio of equivalence classes sitting at the lowest level of anonymity as shown in Equation 5.

Let: A be the set of aggregated regions
a_a be the level of anonymity of an aggregated region a
E be the set of all equivalence classes
E_α be the set of all equivalence classes at the lowest level of anonymity

$$\sum_{a \in A} a_a + \left(1 - \frac{|E_\alpha|}{|E|}\right) \tag{5}$$

Optimization Step

The goal of the optimization step is to improve levels of anonymity by adjusting cluster centers. Since the lowest cardinality equivalence classes are a bottleneck to the overall anonymity, we only relocate the cluster centers sitting at the lowest level of anonymity.

The cardinality of an equivalence class must be increased by taking members from its neighboring regions. To do this, we search a neighborhood, defined as the set of adjacent aggregated regions. The site location of each neighbor is paired with the squared number of equivalence class members within that neighbor. We also pair the site location of the region to be improved with a weight of 10. The site of the region will then be adjusted to the median of the locations weighted by their paired values. The squared weights ensure a stronger pull towards areas of higher density and the weight of 10 is used to ensure that the center does not stray overly far from its original location.

Prior to committing the change for the new cluster center location, a check is performed to verify that the objective function value will actually increase. This is done in order to provide the guarantee that each step of optimization that is committed will improve the objective function value.

Convergence

The final consideration for the algorithm is its convergence criteria. There are two scenarios in which optimization will cease. The first occurs if all clusters have reached a sufficient level of anonymity (the user specified value for k-anonymity). The other scenario occurs if the optimization has reached a round where the objective function value no longer increases. If this occurs, the configuration will be reverted to that of the previous round and the process will end.

2.3 Construction of Geographic Aggregation

2.3.1 Basic Voronoi Aggregation

The basic approach to the construction of the aggregation is the approach that was originally used in VBAS. This simply consists in taking the site locations, as determined by the previous component, and providing them as the input sites to construct a Voronoi diagram (Aurenhammer and Klein, 2000). With the diagram constructed, each initial region point must be categorized based on the Voronoi region in which it falls. Point location can be run efficiently for these points since the Voronoi diagram is a planar subdivision. The resultant Voronoi groupings of initial region points represent the initial regions that will be aggregated together.

Since the aggregation of the regions is driven by approximations of desirable populations in the aggregated regions based on the quasi-identifiers of the data set, it is not guaranteed that the resultant regions will be k-anonymous. In order to verify the anonymity of the aggregated regions, we must determine their equivalence classes. These equivalence classes are based on the members of the equivalence classes in the initial regions being merged together. In order to ensure k-anonymity at this point, any resultant equivalence classes that do not have a cardinality greater than or equal to k will have all of their records suppressed.

2.3.2 Iterative Voronoi Aggregation

We can make an iterative version of our basic Voronoi approach by using a loop in which the groupings of initial regions are created according to the Voronoi partitioning and then each Voronoi site is adjusted to the median of the initial regions in its Voronoi region. This process is repeated until no site shifts to a new location. This is, in fact, the k-means algorithm using the initial Voronoi sites as the seeds for the cluster centers. The benefit of using this iterative approach is that the sum of the distances between initial regions and Voronoi sites can be further reduced, leading to more compact aggregated regions.

2.3.3 Site Optimization

Taking the idea of optimization further, we can allow for other types of moves to be made. One way to do this is to allow for these moves to modify the number of sites that are being used. A simple way to improve the level of compactness is to add more sites. Since all initial regions are allocated to the nearest site, it is easy to see that any initial regions allocated to a newly added site must be nearer to that site than the site of their prior allocation. However, the addition of new sites also has the effect of spreading the population more thinly across the aggregated regions which will lead to higher levels of suppression. The site optimization approach both adds new sites and removes other sites in an attempt to balance these levels.

We can decide when sites should be added or removed based on the population levels of the current sites. Ideally, each aggregated region should have roughly the same population level. We can calculate this level as the global population over the number of sites. If a site in the current allocation has less than half of this population level, we remove the site altogether. If a site has more than a

25% surplus in its population level, we split the site in two. This split can be achieved by removing the original site and placing two new sites on either side of the original location, infinitesimally close to that point. These moves by themselves will not achieve balanced population levels, however they can be combined with the iterative site relocation described in Section 2.3.2 to produce an aggregation with balanced population levels.

The site optimization algorithm starts by calculating the ideal population level. Next, we perform site relocation and then enter into an iterative phase consisting of site removal across all current sites followed by site relocation. This phase ends when an iteration is reached where no sites are removed. This is done to ensure that we initially eliminate any regions with very low populations since the next and final phase converges based on compactness and could have left some of these regions in existence. Finally, we iteratively perform both site removal and site splitting across all of the sites followed by site relocation. Since site removal and site splitting make changes that compete with each other, we use compactness as the criterion for convergence by stopping when a round no longer improves the compactness any further. We measure the compactness here as the sum of the distances between initial regions and the sites to which they are allocated. When a round is reached where the compactness is found to deteriorate rather than improve, we revert back to the configuration of the previous iteration and use this as the final aggregation.

2.4 Linear Programming Heuristics

An alternative for a portion of the aggregation process is a partitioning defined by a linear programming solution. The use of linear programming heuristics alone cannot fully address the problem of aggregation in this setting as they do not handle the approximation of the number of regions or the suppression of records and they also require the configuration of constraint settings. They can, however, potentially replace two components of our system: site location selection and aggregation construction. To do this, existing facility-location type linear programming heuristics can be adapted to this setting through the proper configuration of their constraint settings. Thus, we include this technique in our testing to compare its performance as well. The initial regions from our setting correspond to the destinations given as input and the number of facilities is determined by the number of sites calculated in the first component. By ensuring that each destination is supplied by exactly one facility, the allocation of destinations to facilities defines the groups of initial regions which are to be aggregated together.

The objective function of a linear programming heuristic can be specified as the sum of the distances between destinations and the facilities that are providing supplies to them. This corresponds to the goal of creating compact aggregated regions. Constraints can be specified such that the aggregated regions will have roughly equal populations in order to reduce the level of suppression that will be required. This can be done by setting the demand of each destination to 1 and setting the supply level of each facility to the ceiling of the number of destinations over the number of sites. While a solution can be found in this way, the constraints may be too restrictive, thus leading to an aggregation of poor quality. By increasing supply levels, alternate decisions can be made by the heuristic which favour the

objective function over the consideration of equal population levels. As it is not clear what an appropriate choice of supply level is, we employ tests under five different supply level settings: the base level multiplied by factors of 1, 1.2, 2, and 5, and an unconstrained version. These settings were chosen in order to observe the influence both of constraints which are loosened slightly as well as those which are loosened significantly.

2.4.1 The Alternating Location-Allocation Heuristic

The alternating location-allocation heuristic (Goodchild and Massam, 1969) works by alternating between two steps: allocation of supplies and optimization of facility locations. A transportation problem solver is used to compute an allocation of supplies from facilities to destinations which minimizes an objective function under a set of constraints. Facility locations are then optimized by shifting them to the median of the destinations which they supply to further reduce the objective function value. Different methods of seeding may be used for the initial selection of facility locations. We have chosen to employ the seeding method used for the transportation-location heuristic (Cooper, 1972) which is referred to as the alternate location-allocation method. Convergence is reached when no facility locations shift at the end of an iteration. At this point, we can group initial regions (represented by destinations) according to the facility which supplies them in order to obtain the partitioning for the aggregated regions. Equivalence classes can then be updated and suppression can be applied in the same fashion as with the Voronoi-based aggregation.

2.4.2 Transportation-Location Heuristic

There are a number of other heuristics which can be applied to the location-allocation problem. We chose the transportation-location heuristic as detailed in (Cooper, 1972). This heuristic involves an initial selection of facility locations and an unconstrained destination allocation. The allocation is unconstrained with respect to the facility supply levels. This is followed by an iterative process in which a destination is re-allocated from an overburdened facility to another facility with surplus and then a sub-iterative process is executed. In the sub-iterative process, facility locations are adjusted and destinations are reallocated to facilities to which they are nearer than their currently allocated facility if there is any surplus of supply at the nearer facility. A full description of the algorithm can be seen in the referenced paper (Ibid.).

2.5 Evaluation of Aggregation

Once the regions have been aggregated, we must measure the quality of the aggregation that has been produced. The measures listed here are the same as those used in the original VBAS paper (Croft et al, 2016). Although the measurements applied here may be broken up into groups of related measurements in order to form different approaches for this component, we provide a single approach here that contains all of the relevant measurements for the comparisons used in this

paper in order to facilitate the testing. In this section, we briefly describe each of these measurements.

- **Suppression:** The measurement of suppression is simply used to observe the quality of the aggregation based on how many records were suppressed. If a large number of records were suppressed then it likely indicates a poor aggregation as this means that the equivalence classes of the aggregated regions did not have sufficient cardinalities. Thus, lower levels of suppression are preferable. To better enable comparison between data sets of varying sizes, we take this measurement as the percentage of the original records that are suppressed.

- **Compactness:** The compactness of the final regions can be used as a measure for the level of geographic precision. More compact regions are desirable as this would provide greater geographic detail for researchers. This measurement is taken as the sum of distances between the initial region points and the site of their aggregated region.

- **Discernibility:** We employ the discernibility (Bayardo and Agrawa, 2005) information loss metric in order to determine how much geographic information has been lost by checking for overburdened equivalence classes. Higher values indicate a greater amount of lost information, thus, lower values are preferable.

- **Non-Uniform Entropy:** We also employ the non-uniform entropy (Gionis and Tassa, 2008; Mohammed et al, 2009) information loss metric to measure the loss in geographic information based on the probability of correctly guessing the original geographic region of a record given its aggregated region. As with discernibility, higher values indicate a greater amount of information loss.

- **Running Time:** The final measurement is simply a measure of how long the whole process of aggregation takes from start to finish. This is used to determine how the use of the various approaches will affect the time taken to achieve anonymity.

2.6 Summary of Approaches

As a summary for the approaches described throughout this section, we provide a quick reference to the system components and approaches in Table 2.

3 Comparative Evaluation

In order to test and compare different component approaches, we have generated test data sets using publicly available data sets from Statistics Canada. With these testing sets, we have run various scenarios to observe the effectiveness of the approaches. All tests were run on a machine using 16 GB of RAM and a 4.01 GHz processor.

Construction of the Voronoi diagram and point location within the Voronoi diagram are handled by code from CGAL (CGAL, 1995). An open source transportation solver (Morreau, J.-P., 2009) is used for the location-allocation heuristic.

Table 2 VBA System Components and Evaluation Metrics

VBAS Components (C_i)	**Approaches** (C_iA_j) **or Measurements** (C_iM_k)
(C_1) Approximation of the number of sites	(C_1A_1) Entropy
	(C_1A_2) Max Combinations
(C_2) Selection of the site locations	(C_2A_1) Balanced Density
	(C_2A_2) Anonymity-Driven Clustering
(C_3) Aggregation of the regions	(C_3A_1) Basic Voronoi Aggregation
	(C_3A_2) Iterative Voronoi Aggregation
	(C_3A_3) Site Optimization Aggregation
	($C_{23}A_4$) The Alternating Location-Allocation Heuristic
	($C_{23}A_5$) Transportation-Location Heuristic
(C_4) Evaluation of the aggregation	(C_4M_1) Suppression
	(C_4M_2) Compactness
	(C_4M_3) Information Loss - Discernibility
	(C_4M_4) Information Loss - Non-Uniform Entropy
	(C_4M_5) Time

This code employs the "stepping-stone" approach (Charnes and Cooper, 1954). All other code was written by us in C++. Originally, CGAL was also employed for nearest and farthest neighbor queries in the transportation-location heuristic, however preliminary testing revealed that the overhead required for the construction of the relevant data structures led to longer running times than a linear time implementation of the queries.

3.1 Generation of Testing Data

The data sets that we have used to generate the testing data are the public use microdata file from the 2011 National Household Survey (NHS) (Statistics Canada, 2014) and the Canadian dissemination areas data set (Statistics Canada, 2015). As required by Statistics Canada's data use regulations, it is stated that the results or views expressed here are not those of Statistics Canada.

We have used the NHS data set, which contains respondent level information across a wide range of demographic attributes for a 2.7% sample of the Canadian population, to make approximations for the distributions of attribute values across the response categories of a selection of the demographic attributes. Since the NHS data set has geographic precision at the granularity of provinces and territories, the approximations were made for each of the selected attributes in each province and territory. Since a much finer degree of geographic precision is needed to conduct our tests, we have combined these approximations of the distributions with the dissemination areas data set to create our testing sets. Each dissemination area has a population in the range of 400 and 700 (Statistics Canada, 2015). We have therefore randomly selected a population within this range for each dissemination area as number of records to generate for the area. Each record generated in this way is given a value in each of the selected attributes by selecting from among the response categories of the attribute with a probability of selection in each category corresponding to the approximated distribution that was made for that attribute in the province or territory in which the dissemination area exists. Additionally, each record is given a geographic attribute indicating the dissemination area to which it pertains. This process of data generation produces a testing set with a population roughly equal to that of Canada and with a geographic precision at the level of dissemination areas. In order to produce testing sets for different regions to work with, this set for all of Canada is broken up into three subsets: Western Canada, Central Canada, and Eastern Canada. Since the system requires an input

file with information about the initial regionalization, the dissemination areas data set is also broken up into three corresponding subsets. With these subsets created, any pair of respondent data and the matching dissemination areas subset can be supplied as the input files to VBAS in order to run tests.

3.2 Testing Scenarios

In order to test the component approaches, four different selections of quasi-identifiers on which to achieve anonymity have been made and these selections have been run using the Eastern and Western region testing sets for a total of eight different test scenarios. Scenarios 1-4 represent the quasi-identifier selections in Eastern Canada and Scenarios 5-8 represent the same selections in Western Canada. We have employed these selections of quasi-identifiers in different regions to provide a range of different scenarios and to avoid any bias induced by the input data in a single scenario configuration. The combinations of quasi-identifiers that are used can be found in Appendix B. With a range of results from different scenarios, we can identify relationships between the different approaches in order to determine which of them consistently perform well for each measurement. The results of these tests have been recorded in all measurements indicated in the evaluation component.

For each scenario, all possible combinations of the component approaches have been tested. There are 12 possible combinations of our own approaches. Additionally, there are four more combinations which can be formed with the two linear programming heuristics using the two site number approximation approaches.

4 Results and Discussion

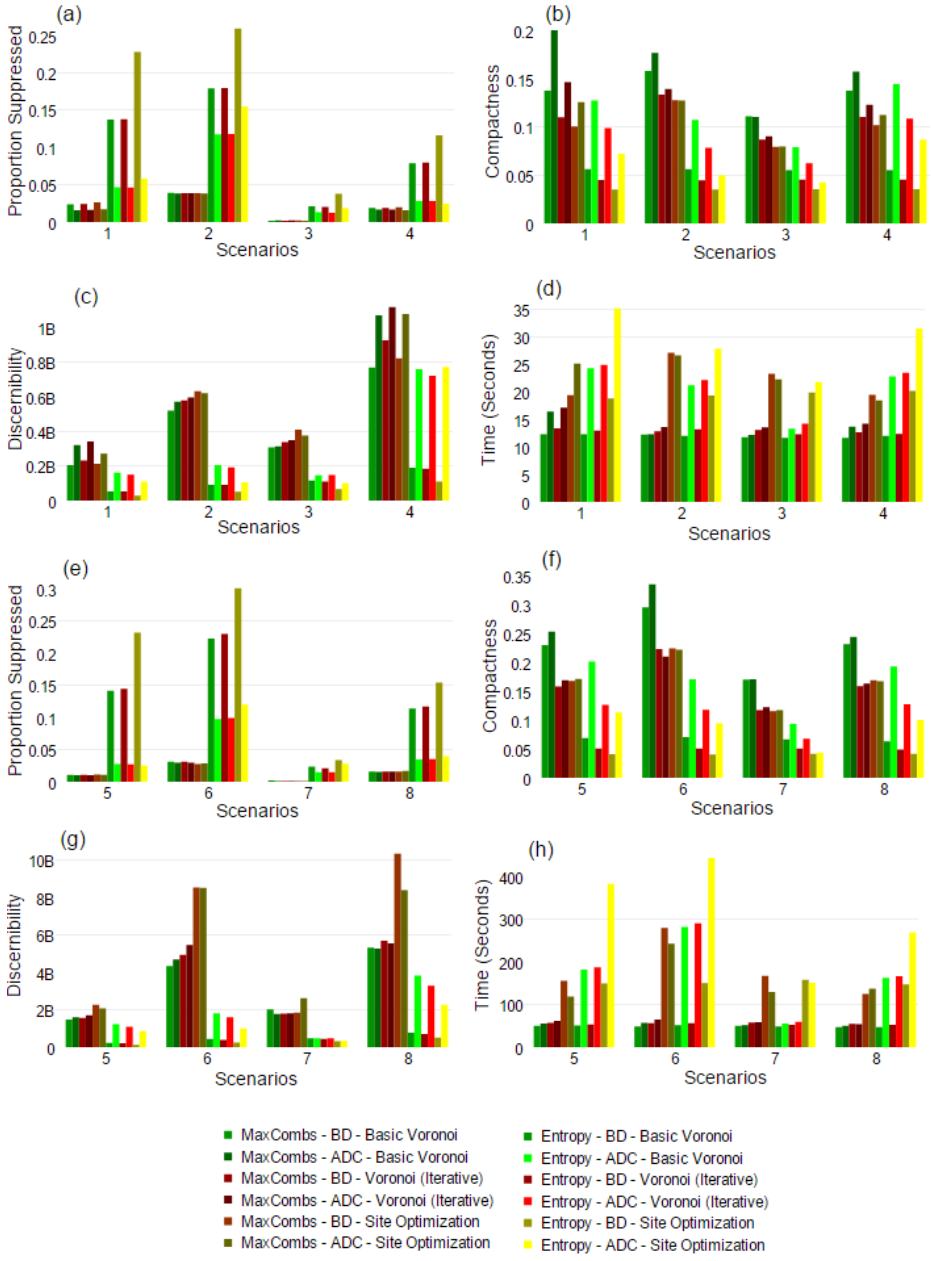
In this section, we discuss the results seen from the comparisons of the different approaches. Graphs showing these results can be found in Appendices C, D and E. The raw data from the test runs can be found in Appendices F, G and H. Due to highly similar findings from the results of both the discernibility and non-uniform entropy information loss metrics, we simply refer to these measures as information loss throughout the discussion and show only the discernibility measure in graph form for conciseness. The raw data for both measures can be found in the appendix.

4.1 Comparison of Site Number Approximation Approaches

When comparing the two site number approximation approaches, the entropy approach consistently determined a greater number of sites to use than the max combinations approach. The effects of this are very clear from the results. The entropy approach caused a greater amount of suppression but produced better levels of compactness and information loss (Figure 1 a-c and e-g).

These findings are rather intuitive since the number of sites corresponds to the number of aggregated regions. A greater number of regions implies higher levels of geographic precision which accounts for the lower values of information loss

Fig. 1 Approach Combinations Comparison Graphs. Scenarios 1-4 represent the quasi-identifier selections in Eastern Canada and Scenarios 5-8 represent the same selections in Western Canada.



and compactness. Additionally, with a greater number of regions, the records are spread more thinly across them, causing more suppression.

In terms of running time, both approaches had very similar times except for when ADC was employed for site location selection (Figure 1 d and h). As ADC is more sensitive to the number of sites than other approaches, when entropy site number approximation is used, the running time became much higher than what was observed with the max combinations approach.

If a user should prioritize reduction of suppression over reduction of geographic information loss then the max combinations approach is preferable. For a user that prioritizes the reduction of the geographic information loss, the entropy approach is the preferable choice.

4.2 Comparison of Site Location Selection Approaches

While the balanced density approach aims to create regions with roughly equal populations levels, ADC prioritizes the reduction in suppression. This is most noticeable in cases where the entropy site number approximation was used as this causes a greater number of sites to be given as input. Similar to the differences noted in the site number approximation comparison, the reduction in suppression achieved by ADC generally led to a degradation in terms of compactness and information loss (Fig. 1a–c, e–g). These effects are quite prominent with the entropy site number approximation approach but much less so with the max combinations approach. In fact, in a few cases the opposite was observed, albeit by a very small margin. These outlying cases were mostly observed when the site optimization aggregation construction approach was employed. This is because the site optimization approach has the ability to modify the number of sites and thus counteracts some of the changes made by ADC.

For the comparison of running times, the balanced density approach was typically faster than ADC (Figure 1 d and h). This is because ADC performs an iterative optimization over the sites unlike the balanced density which immediately selects the final location for each site. It is also for this reason that the gap between the running times was much larger when the entropy site number approximation was used since this increased the number of rounds of optimization that ADC executed as well as the time spent on each round. Once again, when site optimization was employed for the aggregation construction, outlying cases were seen where ADC caused the running time to drop in comparison to balanced density, this time, typically when using the max combination site number approximation. This is due to the fact that the moves already made by ADC cause a reduction in the number of rounds needed by the site optimization.

We conclude that balanced density is preferable when the user prioritizes compactness and reduction of information loss, as well as faster running time. If a user wishes to reduce the level of suppression then ADC is a preferable choice.

4.3 Comparison of Aggregation Construction Approaches

When comparing the levels of suppression produced by the aggregation construction approaches (Figure 1 a and e), all three produce very similar results when

using the max combinations site number approximation approach. With the entropy approach, the site optimization has higher levels of suppression than the other two although this gap is reduced slightly when they are used with ADC for the site location selection.

In general, site optimization had the best performance for the level of compactness, followed by the iterative Voronoi approach and then the basic Voronoi approach (Figure 1 b and f). However, in a few of the cases when using max combinations, the iterative Voronoi approach did better in compactness by a small margin. In terms of information loss, the opposite trend can be seen with max combinations (Figure 1 e and g); The basic Voronoi approach performs the best, followed by the iterative Voronoi approach and then the site optimization. Since the iterative Voronoi and site optimization approaches make moves of optimization which prioritize compactness, it follows that they would show better levels of compactness and worse levels of information loss than the basic Voronoi approach. When using the entropy site number approximation, the same trend is seen with compactness, however, for information loss, the basic Voronoi and iterative Voronoi approaches produce very similar results while the site optimization performs better than them. It is likely that the site optimization approach was able to make better moves of optimization in the presence of a larger number of sites.

In terms of the running time (Figure 1 d and h), the basic Voronoi approach is consistently the fastest as it does not make any moves of optimization. The iterative Voronoi takes only slightly longer. The site optimization takes significantly longer as it has more room to make moves of optimization and considers more criteria than the iterative Voronoi approach.

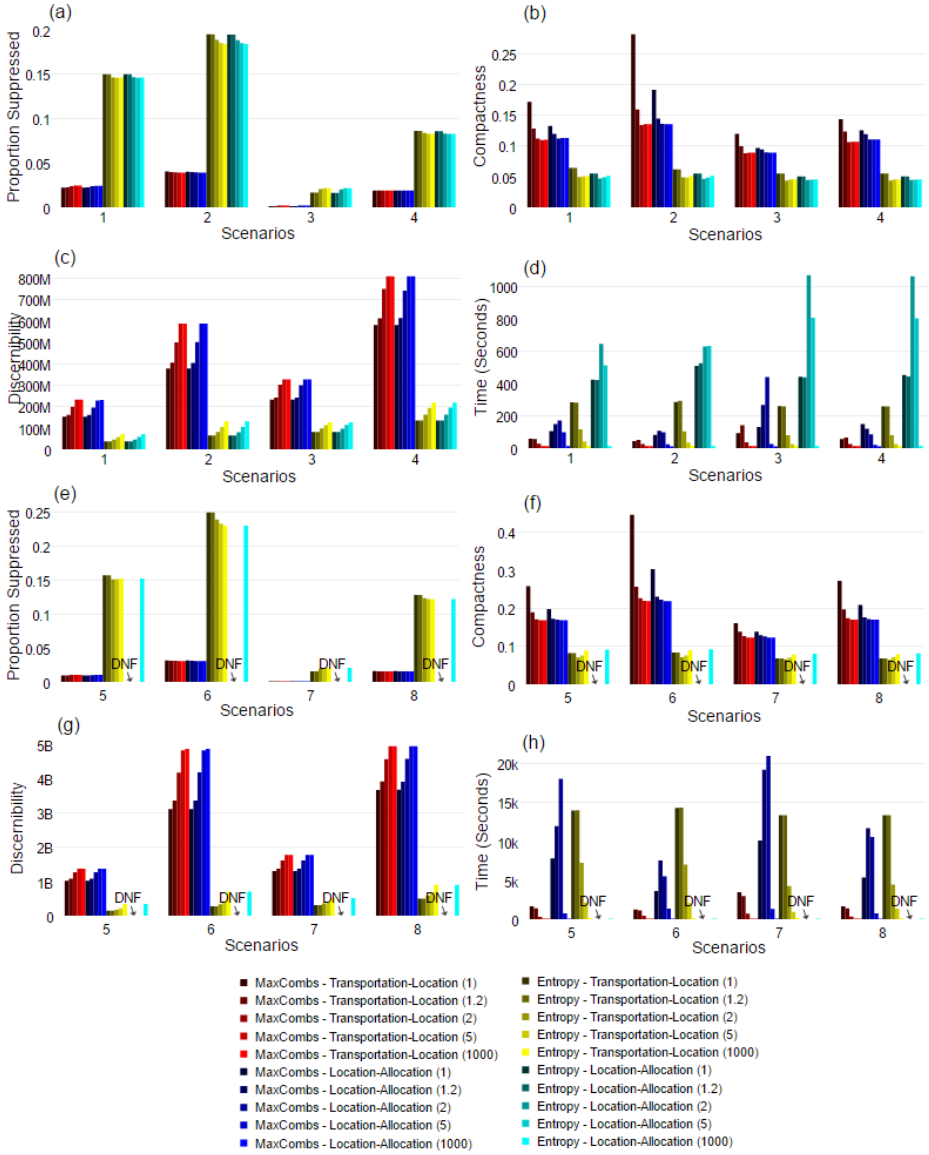
If the user prioritizes compactness, the site optimization approach would be the best choice. The iterative Voronoi approach, however, may also be a good choice if the user wants faster running times while still achieving a good level of compactness. Performance in the other measurements between these two approaches varies depending on which approaches the user selects for the other components.

4.4 Comparison of Linear Programming Heuristics

Here, we compare all combinations of the two linear programming heuristics with each other under their five different constraint settings using the two site number approximation approaches. Tests for the location-allocation heuristic under constraint settings 1, 1.2, 2 and 5 with the entropy site number approximation approach in Western Canada have been omitted as each of these test runs took over 10 hours to complete.

First, as a general comparison between the two different heuristics, it can be seen that they are both very similar in terms of their measurements of suppression and information loss (Figure 2 a, c, e and g). The two approaches, under the same constraint setting, produce nearly identical results for these measurements. However, significant differences are seen when observing compactness and running time (Figure 2 b, d, f and h). With respect to compactness, the transportation-location heuristic produces worse levels of compactness than those of the location-allocation heuristic when the constraints are very tight (lower multipliers). Yet, as the constraints are loosened, the two approaches produce values that are nearer together

Fig. 2 Linear Programming Comparison Graphs.



The DNF tag indicates cases where the test did not finish within 10 hours

and at multipliers of 2 and up, they are almost the same. In terms of running time, the location-allocation heuristic is many times slower than the transportation-location heuristic under all constraint settings except for when they are unconstrained.

Next, we consider the effects of the constraint selection for these approaches. Though we hypothesized that tighter constraints should produce lower levels of suppression, this was not always the case (Figure 2 a and e). This may be due

to the fact that there is not a direct relationship between population level and the level of suppression. The relationship between the constraint settings and the levels of compactness is fairly clear (Figure 2 b and f). As expected, in most cases, loosening the constraints caused the level of compactness to improve. This is due to the fact that the algorithms have a greater ability to make moves of optimization for the objective function since they are not as restricted in the required population levels. The levels of information loss increase as the constraint settings are loosened (Figure 2 c and g). This is to be expected as less balanced population levels will lead to a higher degree of information loss. In terms of the running times, it seems to be the case that there is a peak in the times near the middle of the constraint settings (Figure 2 d and h). This is likely due to two competing factors. At tighter constraint settings, there may be a need for more rounds of optimization in order to satisfy these settings and minimize the objective function. On the other hand, loosening the constraint settings has the effect of allowing for more moves of optimization to be made which could also lead to a greater number of iterations.

We recommend the selection of the transportation-location heuristic over the location-allocation heuristic due to its much shorter running times. Similarly, the very high running times of the transportation-location heuristic at tight constraint settings make the selection of a low constraint multiplier undesirable. As seen from the great increase in time from the Eastern to Western Canada scenarios, it is clear that the heuristic does not scale well at tight constraint settings. The findings here suggest that since the use of constraints does not seem very beneficial, the use of linear programming heuristics may not be very appropriate when applied in this manner. The unconstrained versions of these heuristics essentially amount to an approach which consists of seeding followed by iterative optimization. If a user prioritizes low information loss then they may select a linear programming heuristic at tight constraint setting, however the disadvantages in the other measurements do not make this a very judicious choice.

4.5 Comparison of Recommended Combinations

To provide a summary of some of the results from the previous comparisons, we have listed the best approaches for each of the measures in Table 3. Next, we present specific combinations of approaches that we have selected as user recommendations. The selected combinations are listed in Table 4. Note that in the cell for information loss on the aggregation construction component, there are three approaches listed. The two linear programming heuristics were tied with each other when kept at a factor of 1 for the constraint setting. When using the max combinations site number approximation, the linear programming heuristics performed best, however, when using the entropy site number approximation, the site optimization approach produced better results.

For ease of reading, each combination will henceforth be referred to by its ID as specified in the table. Combination 1 provides the best level of suppression in all but one of the scenarios, however its margin of superiority is quite small (Figure 3 a and e). As a result, Combinations 2 and 3 are strong contenders even when a user prioritizes suppression. Depending on the scenario, they both display slight to moderate improvements over Combination 1 in terms of compactness, information loss and running time (Figure 3 b-d and f-h). Between these two, Combination

Table 3 Summary of the approaches that performed the best in each of the measures.

	Suppression	Compactness	Information Loss	Running Time
Site Number Approximation	Max Combinations	Entropy	Entropy	Max Combinations
Site Location Selection	ADC	Balanced Density	Balanced Density	Balanced Density
Aggregation Construction	Inconclusive	Site Optimization	Location-Allocation (1) / Transportation-Location (1) / Site Optimization	Basic Voronoi

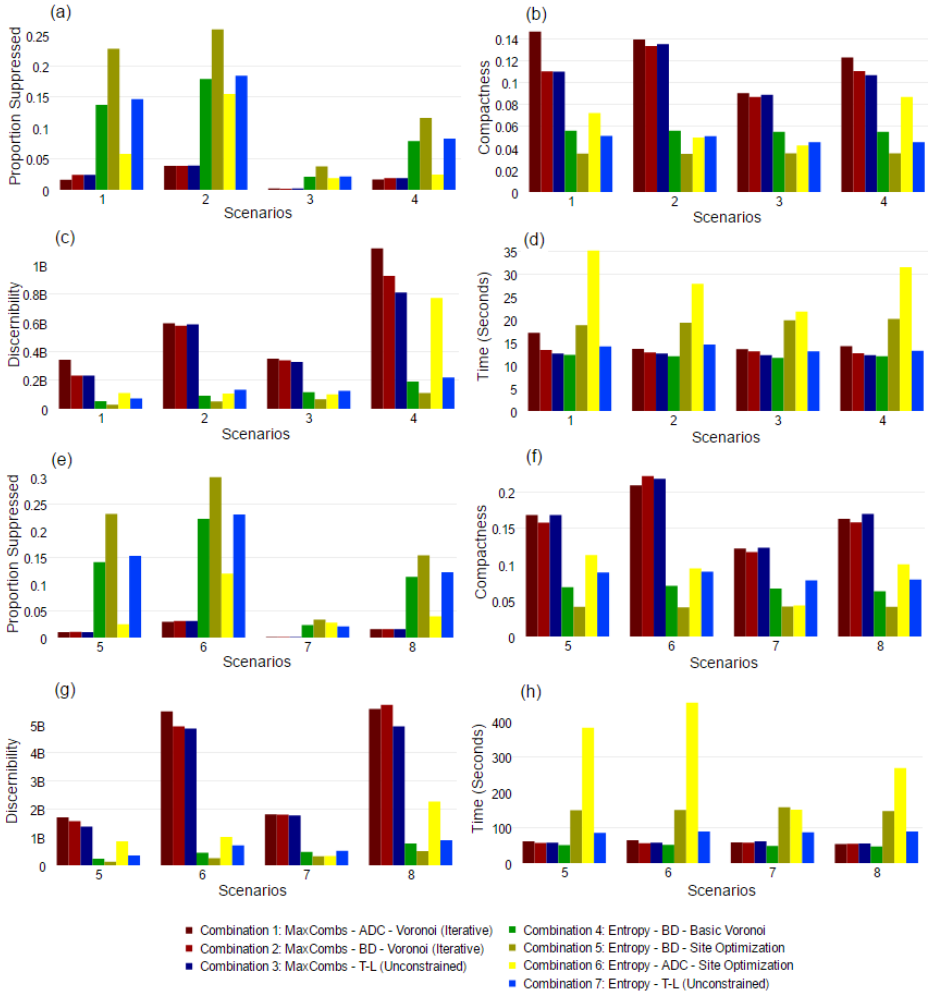
Table 4 Recommended combinations and their motivation.

ID	Site Number	Site Location	Aggregation	Reason
1	Max Combinations	ADC	Iterative Voronoi	Favours suppression
2	Max Combinations	Balanced Density	Iterative Voronoi	Minor trade-off of suppression for minor improvements in other measures (compared to #1)
3	Max Combinations	Transportation-Location (Unconstrained)		Minor trade-off of suppression for minor improvements in other measures (compared to #1)
4	Entropy	Balanced Density	Basic Voronoi	Favours running time and balances other measures
5	Entropy	Balanced Density	Site Optimization	Favours compactness and information loss
6	Entropy	ADC	Site Optimization	Significant improvement in suppression at the cost of significant deterioration in all other measures (compared to #2)
7	Entropy	Transportation-Location (Unconstrained)		Moderate improvement in suppression at the cost of slight to moderate deterioration in all other measures (compared to #2)

2 more often had a slight advantage in compactness and Combination 3 more often had a slight advantage in discernibility. Our recommendation is for a user to select one of these two. Combination 5 consistently had the best performance in compactness and in information loss (Figure 3 b, c, f and g). However, it produces high levels of suppression and has long running times (Figure 3 a, d, e and h). These aspects are very noticeable in contrast to the alternatives that use max combinations. However, for a user who prioritizes compactness or information loss, it is preferable to use entropy for the site number approximation. When comparing Combination 5 to the other combinations which employ entropy site number approximation, the differences in suppression are not as severe and the gains in compactness and information loss are certainly enough to warrant the selection of this combination.

For a user who prioritizes compactness and information loss but is willing to trade some quality in these measurements for better levels of suppression, Combination 6 offers a significant improvement in suppression (Figure 3 a and e). However, in addition to degradation in compactness and information loss, this combination has longer running times. For a less severe trade-off, Combinations 4 and 7 offer a good compromise (Figure 3 a-h). Of these two, Combination 4 is typically slightly superior in levels of suppression and in running time by a slight margin as well. In fact, Combination 4 is consistently the fastest out of all of the combinations. In terms of compactness, Combination 4 performed better in all scenarios in Western Canada while Combination 7 performed better in all scenarios in Eastern Canada exposing a potential dependence upon the input data in one or both of these combinations. The two combinations are roughly tied in terms of levels of information loss. Thus, as Combination 4 is superior in suppression and running time, it would be our recommendation here.

Fig. 3 Recommended Combinations Comparison Graphs.



5 Conclusions

In this paper, we have focused on the problem of geographic-based anonymization with special emphasis on geographic precision. VBAS, the system we introduced to address this problem, is designed in a modular fashion to allow for different approaches to be supplied for its components. This provides the ability to easily test the effectiveness of various approaches and combinations of such approaches. We have developed a working implementation of the system which we use for testing and that includes implementations of different approaches that can be used in the system components. We have also adapted and included existing linear programming heuristics which can be used to create aggregated regions with roughly equal population levels.

We have run a series of tests using synthetic data sets which we have generated. Through these tests, we are able to clearly demonstrate the trade-offs between different objectives that are inherent in the process of anonymization. Furthermore, we have shown that while the linear programming heuristics can be applied to this problem, they are, for the most part, only useful in an unconstrained setting, indicating that the constraints used in this way are not very effective. Based on the results of the comparisons of all approaches, we have made recommendations to users about which approaches are appropriate based on the user's requirements. Additionally, we recommended combinations of approaches to choose from based on user requirements. With these recommendations, a user with minimal background knowledge of how to anonymize data can use VBAS in an effective manner.

The use of VBAS along with these recommendations provides an easy-to-use process for generating anonymized data with high precision geographic information, as is often required in various data analytics research. With this work completed, it is of interest to expand our comparison of VBAS with other systems of geographic-based anonymization. In particular, we now plan to compare VBAS to systems using perturbation-based anonymization. As these systems use a fundamentally different approach to protect privacy, the same data utility metrics cannot be applied. We will therefore employ a comparison of data utility with regards to spatial analysis.

Acknowledgements The authors gratefully acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grants No. RGPIN-2015-05390 and No. RGPIN-332-2011.

References

Armstrong M, Rushton G, Zimmerman D (1999) Geographically Masking Health Data To Preserve Confidentiality. Stat Med 18:497–525

Arzberger P, Schroeder P, Beaulieu A, et al (2004) Promoting Access to Public Research Data for Scientific, Economic, and Social Development. Data Sci J 3:135–152

Aurenhammer F, Klein R (2000) In: Sack, J-R and Urrutia, J (ed) Handbook of Computational Geometry, Elsevier Science Publishers B.V. North-Holland, chap Voronoi Diagrams, pp 201–290

Bayardo RJ, Agrawa R (2005) Data Privacy Through Optimal k-Anonymization. In: Proc. 21st ICDE '05, pp 217–228

Benitez K, Malin B (2010) Evaluating Re-identication Risks with Respect to the HIPAA Privacy Rule. J Am Med Inform Assoc 17:169–177

Boulos M, Cai Q, Padget JA, et al (2006) Using Software Agents to Preserve Individual Health Data Confidentiality in Micro-Scale Geographical Analyses. J Biomed Inform 39:160–170

Bridwell SA (2007) In: Miller HJ (ed) Societies and Cities in the Age of Instant Access, Springer Netherlands, chap The Dimensions of Locational Privacy, pp 209–225

CGAL (1995) The Computational Geometry Algorithms Library. URL http://www.cgal.org/, [Accessed May, 2014]

Charnes A, Cooper WW (1954) The Stepping Stone Method of Explaining Linear Programming Calculations in Transportation Problems. Management Science 1:49–69

Clifton KJ, Gehrke SR (2013) Application of Geographic Perturbation Methods to Residential Locations in the Oregon Household Activity Survey. Transportation Research Record 2354:40–50

Cooper L (1972) The Transportation-Location Problem. Operations Research 20:94–108

Croft W, Shi W, Sack J-R, et al (2016) Location-Based Anonymization: Comparison and Evaluation of the Voronoi-Based Aggregation System. Int J Geogr Inf Sci 30:2253–2275

Emam KE, Brown A, AbdelMalik P (2009) Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk. J Am Med Inform Assoc 16:256–266
Emam KE, Dankar FK, Neisa A, et al (2013) Evaluating the Risk of Patient Re-identification from Adverse Drug Event Reports. BMC Med Inform Decis 13
Ford LR, Fulkerson DR (1956) Solving the transportation problem. Management Science 3:24–32
Gionis A, Tassa T (2008) k-Anonymization with Minimal Loss of Information. IEEE Trans Knowl Data Eng 21:206–219
Goodchild M, Massam B (1969) Some Least-Cost Models of Spatial Administrative Systems in Southern Ontario. Geografiska Annaler 51:86–94
Greenberg B, Voshell L (1990) Relating Risk of Disclosure for Microdata and Geographic Area Size. In: Proc. SRMS, Am. Stat. Assoc., pp 450–455
Hawala S (2001) Enhancing the "100,000 Rule" on the Variation of the Per Cent of Uniques in a Microdata Sample and the Geographic Area Size Identified on the File. In: Proc. Annu. Meeting Am. Stat. Assoc., pp 1–6
Jung, H-W and Emam, K E (2014) A Linear Programming Model for Preserving Privacy when Disclosing Patient Spatial Information for Secondary Purposes. Int J Health Geogr 13
Lowrance W (2006) Access to Collections of Data and Materials for Health Research: A Report to the Medical Research Council and the Wellcome Trust. Medical Research Council and the Wellcome Trust pp 1–39
Lyseen AK, Nohr C, Sorensen EM, et al (2014) A Review and Framework for Categorizing Current Research and Development in Health Related Geographical Information Systems (GIS) Studies. Yearb Med Inform 9:110–124
Mohammed N, Fung BCM, Hung PCK, et al (2009) Anonymizing Healthcare Data: A Case Study on the Blood Transfusion Service. In: Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data, pp 1285–1294
Morreau, J-P (2009) TRANSPOR.CPP. URL http://jean-pierre.moreau.pagesperso-orange.fr/cplus.html, [Accessed Apr, 2016]
Olson KL, Grannis SJ, Mandl KD (2002) Privacy Protection Versus Cluster Detection In Spatial Epidemiology. Am J Public Health 96:2002–2008
Rezaeian M, Dunn G, Leger SS, et al (2007) Geographical Epidemiology, Spatial Analysis and Geographical Information Systems: a Multidisciplinary Glossary. J Epidemiol Commun H 61:98–102
Samarati P (2001) Protecting Respondents Identities in Microdata Release. IEEE Trans Knowl Data Eng 13:1010–1027
Statistics Canada (2014) Individuals File, 2011 National Household Survey (Public Use Microdata Files). URL http://www5.statcan.gc.ca/olc-cel/olc.action?objId=99M0001X2011001&objType=46&lang=en&limit=0, [Accessed Mar, 2015]
Statistics Canada (2015) Dissemination Area (DA). URL http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo021-eng.cfm, [Accessed Mar, 2015]
Sweeney L (2002) k-Anonymity: A Model for Protecting Privacy. Int J Uncertain Fuzz 10:557–570
Thomas Y, Richardson D, Cheung I (2008) In: Thomas Y, Richardson D, Cheung I (eds) Geography and Drug Addiction, Springer Netherlands, chap Integrating Geography and Social Epidemiology in Drug Abuse Research, pp 17–26
Vora A, Burke DS, Cummings DAT (2008) The Impact of a Physical Geographic Barrier on the Dynamic of Measles. Epidemiol Infect 136:713–720
Young C, Martin D, Skinner C (2009) Geographically Intelligent Disclosure Control for Flexible Aggregation of Census Data. Epidemiol Infect 23:457–482
Zhou A, Qu B, Li H, et al (2011) Multiobjective Evolutionary Algorithms: A Survey of the State of the Art. Swarm and Evolutionary Computation 1:32–49