Published in final edited form as:

Geogr Anal. 2015 July ; 46(3): 297–320. doi:10.1111/gean.12040.

Assessing Activity Pattern Similarity with Multidimensional Sequence Alignment based on a Multiobjective Optimization Evolutionary Algorithm

Mei-Po Kwan¹, Ningchuan Xiao², and Guoxiang Ding³

¹Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA

²Department of Geography, The Ohio State University, Columbus, OH, USA

³Department of Analytics and Research, Travelers Insurance, Hartford, CT, USA

Abstract

Due to the complexity and multidimensional characteristics of human activities, assessing the similarity of human activity patterns and classifying individuals with similar patterns remains highly challenging. This paper presents a new and unique methodology for evaluating the similarity among individual activity patterns. It conceptualizes multidimensional sequence alignment (MDSA) as a multiobjective optimization problem, and solves this problem with an evolutionary algorithm. The study utilizes sequence alignment to code multiple facets of human activities into multidimensional sequences, and to treat similarity assessment as a multiobjective optimization problem that aims to minimize the alignment cost for all dimensions simultaneously. A multiobjective optimization evolutionary algorithm (MOEA) is used to generate a diverse set of optimal or near-optimal alignment solutions. Evolutionary operators are specifically designed for this problem, and a local search method also is incorporated to improve the search ability of the algorithm. We demonstrate the effectiveness of our method by comparing it with a popular existing method called ClustalG using a set of 50 sequences. The results indicate that our method outperforms the existing method for most of our selected cases. The multiobjective evolutionary algorithm presented in this paper provides an effective approach for assessing activity pattern similarity, and a foundation for identifying distinctive groups of individuals with similar activity patterns.

Introduction

The use of location-aware devices to collect detailed space-time data about individuals has increased dramatically in geographic, health, and social science research in the past decade or so (e.g., Wiehe et al. 2008; Shoval et al. 2011; Wesolowski et al. 2012; Richardson et al. 2013; Shen, Kwan, and Chai forthcoming). These data offer many opportunities for individual-based research to enhance our understanding of complex human spatial behavior and social interactions (Kwan 2004, 2013; Smyth 2001; Raubal et al. 2004; Griffith et al. 2013; Palmer et al. 2013). As part of this endeavor, researchers have sought to derive representative human activity patterns with these high-resolution space-time data using various clustering methods, such as *k*-means and hierarchical clustering (Becken, Simmons,

and Frampton 2003; Schlich and Axhausen 2004; Gao et al. 2010; Chen et al. 2011; Sadahiro, Lay, and Kobayashi 2013). Often these methods are based on similarity analysis that, in turn, relies on various distance measures for evaluating the closeness or similarity between individual space-time behavior (Koppelman and Pas 1985; Joh et al. 2001a; Schlich and Axhausen 2003; Sinha and Mark 2005; Long and Nelson 2013).

During the past three decades, geographers and transport researchers have conducted many studies to develop distance or similarity measures for activity pattern classification (Burnett and Hanson 1982; Hanson and Huff 1986 ; Pas 1983 ; Koppelman and Pas 1985 ; Wilson 1998a, 2001; Joh et al. 2001b,c; Joh et al. 2002). In geographic information science (GIScience), developing new methods for analyzing the space-time trajectories of moving humans or objects has attracted much attention in recent years, perhaps inspired by the time-geographic studies of human activity patterns in the late 1990s and early 2000s (e.g., Hornsby and Cole 2007 ; Laube et al. 2007; Dodge, Laube, and Weibel 2012; Orellana and Wachowicz 2012; Persson and Ellegård 2012). Although these past efforts have made considerable progress in advancing the field, some major challenges still remain. Among these is the difficulty of deriving representative activity patterns while taking into account the multiple dimensions of human activity-travel behavior, which include activity location, purpose, type, sequence, timing, and duration. Frequently existing distance or similarity measures are based on only limited dimensions of human or object movement (i.e., spatial, temporal, or attribute proximity), and little attention has been paid to comparing human activity patterns in terms of many dimensions at the same time (Demšar and Virrantaus 2010). Furthermore, most studies to date tend to focus more on the geometric and statistical properties of movement trajectories (e.g., direction, acceleration, or geometric shape) than on the substantive characteristics of human activities and trips (e.g., travel purpose, travel mode, or the number of accompanying persons) (e.g., Gonzalez, Hidalgo, and Barabási 2008; Dodge, Laube, and Weibel 2012).

This paper addresses these challenges by presenting a unique methodology for assessing the similarity between multidimensional human activity patterns. It conceptualizes sequence alignment as a multiobjective optimization problem, and summarizes the solution to this problem that is given by an evolutionary algorithm. The solution utilizes sequence alignment to code multiple facets of human activities into multidimensional sequences, and to treat similarity assessment as a multiobjective optimization problem that aims to minimize the alignment cost for all dimensions simultaneously. This paper demonstrates the effectiveness of the method by comparing it with a popular existing method called ClustalG using a set of 50 sequences. The results indicate that our method outperforms the existing method for most of the cases. The multiobjective evolutionary algorithm described here provides an effective approach for assessing activity pattern similarity, and a foundation for identifying distinctive groups of individuals with similar activity patterns.

Analysis of human activity patterns

Many issues of interest to geographers, transportation scientists, health researchers, and urban planners will benefit greatly from a better understanding of human activity patterns in space-time (Kwan 1999, 2013; Richardson et al. 2013). The analysis of moving objects and

humans is a relatively young field with many research issues to be explored (Dodge, Laube, and Weibel 2012; Orellana and Wachowicz 2012; Long and Nelson 2013). For instance, how to identify objects with abnormal movement patterns, how to derive distinctive and meaningful patterns from multidimensional trajectory data, and how to analyze multidimensional semantics of moving objects. These are challenging problems because moving objects exhibit multidimensional spatiotemporal characteristics that need to be handled synthetically (Han and Gao 2009; Zhu et al. 2009; Miller and Han 2009). Classifying moving objects into distinctive groups provides one way to discover underlying patterns and interactions.

Human activities and their contexts can be described by inter-related variables that reflect their multiple facets, including the location, sequence, timing and duration of activities, activity type, travel purpose, travel mode, speed, direction, the number of accompanying persons during an activity or trip, and other socio-demographic characteristics (Kwan 2000, 2004, 2012). These various facets are the multiple dimensions of human activity-travel behaviour. To study human activity patterns, appropriate analytical methods need to be applied in order to derive distinctive behavioral patterns. An important basis for all of these methods is a distance or similarity measure that allows researchers to determine how close or similar an individual's activity pattern is relative to that of another individual.

In light of the many different facets of human activity-travel behavior, an effective distance measure should allow us to take into account many characteristics of human activities in addition to the spatial and temporal dimensions. Three of these characteristics deserve particular attention. First, individuals conduct different activities at certain times. Differences in the attributes of these activities (e.g., type and purpose), or *compositional differences*, should be captured when comparing individual activity patterns. Second, the *interdependency* among these dimensions needs to be maintained in a distance or similarity measure (e.g., certain activities can take place only at certain places and/or at certain times). Third, human activities unfold in a sequential order over time. When comparing activity patterns, the distance measure should be able to compare *structural differences* in human activities and their contextual variables (e.g., certain activities have to be performed before specific other activities; Pas 1983). While many past studies have used distance measures to derive human activity patterns, they all have limitations in achieving these three important goals.

Burnett and Hanson (1982) , for instance, developed an early distance measure to compare the difference between individual activity patterns based on a number of attributes, such as activity type, activity location, travel distance, and travel mode. A distance score was obtained by summing the differences across all attributes. Based on Burnett and Hanson's distance score, Pas (1983) developed a general expression of similarity between two activity patterns by introducing the concept of primary-secondary attributes, and assigning weights to aggregate them as the distance measure between each pair of activity patterns. This measure was further improved by Koppelman and Pas (1985) by using a linear assignment programming method to capture the differences in activity composition. Ma and Goulias (1997) used the standardized z-score of variables to measure the distance between patterns. Cha et al. (1995) employed factor analysis to obtain the distance score between Japanese

overseas travelers. These researchers used pattern classification to facilitate the analysis of activity patterns. However, the distance measures they implemented differentiate activity patterns based only on activity composition, while the sequential and structural aspects of activity patterns were not addressed.

Some recent efforts have started to incorporate both compositional and sequential characteristics of activity patterns into distance measures (e.g., Joh, Arentze, and Timmermans 2001a; Shoval and Isaacson 2007). One example is the feature extraction method based on the Walsh-Hadamard transformation (Recker et al. 1985). In this approach, a set of measurements that define activity patterns is represented by column vectors, and transformation techniques are applied to the column vectors to develop a taxonomy for the pattern space. This method has the advantage of including the sequential order entailed in activity patterns, although it still cannot integrate the multiple dimensions of activity patterns simultaneously into the distance measure. With respect to the limitations of past research about human activity patterns, sequence alignment seems a promising and powerful alternative.

Sequence alignment

Sequence alignment is a method widely used in molecular biology for identifying regions of similarity in DNA, RNA or protein sequences (Kruskal 1983). Compared with all other methods previously discussed, an outstanding strength of sequence alignment is its capability to take into account the multiple attributes as well as the compositional and sequential characteristics of human activity patterns. Researchers in various fields - including sociology (Abbott 1995; Stovel and Bolan 2004), transportation (Joh, Arentze, and Timmermans 2001a), tourism (Bargeman, Joh, and Timmermans 2002, Shoval and Isaacson 2007), and retailing (Joh, Timmermans, and Popkowski-Leszczyc 2003) - have realized the value of sequence alignment. Wilson (1998a, b) was the first to introduce sequence alignment into travel behavior research, where, similar to molecular sequences, each person's daily activities are coded as a sequence of characters to represent different activity characteristics.

Sequence alignment seeks to align two sequences by transforming one into the other using the minimal number of character edit operations (i.e., insertions, deletions, and substitutions; Needleman and Wunsch 1970). The minimal number of edit operations used to align two sequences is defined as the distance (or alignment cost) between them. Consider two example sequences, *LHTH* and *LTWTH*, where each alphabet represents a type of activity. We write one sequence above the other, and an alignment can be obtained by arranging the same or similar characters in the same column. In our example, a possible alignment can be

$$L-H T H$$

written as *L T W T H*, where a dash ('-') represents a gap of length one inserted between the first two elements of the sequence on the top. An insertion in one sequence can be seen as a deletion in the other one, and a substitution can be treated as the combination of a deletion and an insertion operation. In our example, inserting a gap between *L* and *H* in the top sequence is equivalent to deleting the second letter (*T*) from the bottom sequence. Similarly, the second letter in the top sequence (*H*) is substituted by *W*, or equivalently, the third letter in the bottom sequence is substituted by *H*. Therefore, the example alignment can

be written using an operation set $[d_2, d_3, i_3]$, indicating the deletion of the second and third elements from the bottom sequence, and the insertion of an element at the third position to the bottom sequence. After the alignment, these two sequences have the same character at each position and the *alignment score or cost* (edit operations required) is 3. For this example, all other alignment alternatives will require at least 3 operations. Therefore, we can say the distance between the two sequences is 3 (assuming equal weight of insertion and deletion). This is a simple example that illustrates one method of aligning two sequences. Different methods exist for aligning two sequences, each requiring a different set of operations.

Multidimensional sequence alignment (MDSA)

To align multidimensional sequences, conventional sequence alignment methods need to be extended in order to handle the multiple dimensions of these sequences. But multidimensional sequence alignment (MDSA) is significantly more difficult than aligning unidimensional sequences. One strategy is to align the sequences for each dimension separately, and then sum the alignment cost in each dimension to obtain a combined score for the multidimensional sequences. For example, Joh et al. (2001a) propose a hybrid algorithm that first searches for the optimal alignment for each dimension separately using a genetic algorithm, and then combines the unidimensional alignments across the attributes to achieve optimal overall scores. Elements sharing the same operation at the same position for all dimensions are treated as if they are in one element. To illustrate how this strategy works, consider two multidimensional activity sequences, $X=L1H2T3H2$ and $Y=L1T4W3T4H2$, where each activity is composed of two characters. The first character (a letter) indicates the type of activity, and the second (a number) represents the location of the activity. Two possible alignments based on activity and location dimensions are $[d_2, d_3, i_3]$ and $[d_2, i_2, d_4]$ respectively (see Figure 1a), both applied to sequence Y . The alignment costs for each of the two separated dimensions equal 3, whereas the combined alignment (Figure 1b) cost when both dimensions are considered is 5 because a common deletion occurs (i.e., d_2) for both dimensions at position 2.

Although aligning each dimension separately and then calculating the combined costs for all dimensions is straightforward, the main drawback of this strategy is the disconnection in the interdependent relationships among activities and their attributes because the alignment for each dimension tends to be different. As illustrated in Figure 1b, two gap characters are inserted into sequence X , which disconnect inter-related activities and attributes. For example, ‘-2’ in X is aligned without activity type, while ‘T-’ is aligned with missing location information. Moreover, the second element of X ($H2$) is changed to $H3$ after the alignment because a ‘-’ needs to be inserted into the letters in X to achieve an optimal alignment of the letters between X and Y (left panel of Figure 1a), while a gap must be inserted into a different position in the numbers of X to optimize the alignment of the numbers. These two operations force the combination of H and 3 after the two dimensions are combined for X (Figure 1b). This insertion changes the location of the activity and can cause misinterpretation of activity patterns after alignment.

Wilson (1998a) and Wilson et al. (1999) suggest a method for MDSA that involves combining the elements from all dimensions into one element, and then performing a conventional unidimensional sequence alignment (using dynamic programming, for example). The ClustalG software implements this strategy by extending existing unidimensional sequence alignment methods. For example, the operation set for aligning X and Y is $[d_2, d_3, d_4, i_3, i_4]$. Therefore the distance between X and Y is 5 (Figure 1c). In this method, activities that are different in any of their attributes (e.g., T_3 in X and T_4 in Y) are treated as entirely different (which may cause overestimation of the alignment cost). Besides, though ClustalG employs conventional sequence alignment to achieve optimal alignment at the aggregate level, the optimality of the alignment for separate dimensions is not guaranteed, as we subsequently discuss.

To overcome some of these limitations and further advance multidimensional sequence alignment methods, this paper summarizes a formulation of MDSA as a multiobjective optimization problem where alignment is performed in all dimensions in the analysed sequences while the dimensions are evaluated separately. In this formulation, interdependency between different dimensions is maintained while partial differences between activities also are taken into account. However, finding solutions to multiobjective optimization problems is difficult and often computationally intensive (Wang and Jiang 1994; Notredame and Higgins 1996; Zhang and Wong 1997; Joh et al. 2001a). Therefore we developed an evolutionary algorithm-based approach to MDSA. Although evolutionary algorithms (EAs) have been proven to be effective in alignments in various real-world applications (Cai et al. 2000; Joh et al. 2001a; Zhang and Huang 2004), our approach is based on a different alignment representation. In the next two sections, we discuss the design of the special evolutionary algorithm operators (e.g., recombination, mutation) as well as the encoding strategy of our method.

MDSA as a Multiobjective Optimization Problem

The goal of solving a multiobjective optimization problem is to minimize (without loss of generality) a set of objective functions simultaneously (Cohon 1978). Figure 2 portrays an example of minimizing two objective functions, where the objective function values of each solution are used as the coordinates to plot the solution. Solution A is said to dominate solution B because A has smaller values than B for both objective functions. Solutions A , C , and D , however, do not dominate each other and are not dominated by any other solutions in the figure. These solutions are called non-dominated (Van Veldhuizen and Lamont 2000; Tan et al. 2001). The non-dominated solutions are the optimal solutions to a problem that form a front that is generally referred to as the Pareto front.

For two multidimensional sequences, we define an objective function for each dimension. For example, the alignment in Figure 1c can be evaluated by calculating a distance score for two unidimensional alignments: L - HTH and $LTWTH$, and 1-232 and 14342. Formally, we calculate the sum-of-pairs alignment score (Gonnet et al. 2000) of each alignment. Let S_1 and S_2 be two unidimensional sequences with length m and n , respectively. Each sequence element is a character from a finite alphabet set Σ that does not include the reserved gap character ('-'). Let J be the length of a sequence after alignment. J satisfies the condition of

$\text{MAX}(m,n) = J = m+n$. The alignment of the two sequences can be written as a $2 \times J$ matrix $A=(a_{ij})$ that satisfies the following three conditions: (1) the matrix is padded with gap characters so that $a_{ij} \in \Sigma' = \Sigma \cup \{-\}$, (2) removing the gap characters from each row reproduces the corresponding sequence, and (3) no column exists that contains only gap characters. The left or right panel in Figure 1a is an example of A . The sum-of-pairs alignment score of alignment A can be computed as the sum of character substitution costs at each position:

$$\sum_{j=1}^J d(a_{1j}, a_{2j}), \quad (1)$$

where d defines a symmetric matrix that contains substitution costs between each pair of characters in Σ . In this study, we use the Levenshtein distance (Myka and Güntzer 1996) to define the cost matrix which can be written as

$$d(a, b) = \begin{cases} 2 & a, b \in \Sigma \text{ and } a \neq b \\ 1 & (a = '-' \text{ and } b \in \Sigma) \text{ or } (a \in \Sigma \text{ and } b = '-') \\ 0 & a = b \end{cases} \quad (2)$$

The MDSA can be formulated as a multiobjective optimization problem that minimizes the sum-of-pair distance scores for the dimensions simultaneously:

$$\text{MIN}: SP_i = \sum_{j=1}^J d(a_{1ij}, a_{2ij}), \quad \forall i \quad (3)$$

where SP_i is the sum-of-pairs distance scores of dimension i for an alignment A , and $a_{1ij}, b_{2ij} \in \Sigma'$ are the j -th characters in dimension i from the first and second sequence, respectively. In our case, we have two dimensions, and therefore $1 \leq i \leq 2$.

Design of a Multiobjective Evolutionary Algorithm for MDSA

EAs have been widely used to search for solutions to optimization problems involving large, complex, and poorly understood search spaces (Deb 2001; Tan et al. 2001; Xiao et al. 2002; Xiao et al. 2007). An EA can be used to produce a diverse set of alternative solutions, and to search for the Pareto-front for a multiobjective optimization problem. At the beginning of an EA, a population of random solutions is created and then evaluated by corresponding objective functions. Solutions close to the current Pareto front receive high fitness function values, and have a high chance to be selected in the next generation of the population. Then a breeding process is used to generate a new population. First, a certain percentage of the fittest individuals from the current generation is copied to fill the new generation directly. To create the remaining portion of the new generation, parent solutions are selected from the current population based on their fitness values, and then are recombined to generate new solutions. A small portion of the new solutions is altered (mutated) on a random basis. The EA repeats this process until a pre-defined termination condition, such as the maximum number of generations, is met. For a multiobjective optimization problem, maintaining the diversity among the solutions is also important to avoid converging to a single point on the Pareto front.

Graph representation of sequence alignment

An alignment between any two sequences can be represented as a path in a directed acyclic graph, as illustrated in Figure 3a for an alignment between the two sequences discussed in the second section of this paper. In this graph, activities in X are arranged as rows and Y as columns. A horizontal arrow corresponds to the insertion of a gap in the sequence in the rows, or the deletion of characters in the columns. A vertical arrow corresponds to the insertion of a gap in the sequence on the columns, or the deletion of characters in the rows. Finally, a diagonal arrow corresponds to either substitution if the characters in a row and column are different, or identity if the characters are the same. Each possible alignment is a chain of connected arrows between the *start* and *end* nodes on the graph. Each path is evaluated for the two dimensions using the cost matrix defined by equation (2). In our example, for the 6 arrows along the path in Figure 3a, the costs for the first dimension (letters) are 0, 1, 1, 1, 0, and 0, respectively, with a sum-of-pairs distance of 3, while the costs for the second dimension (numbers) are 0, 1, 1, 1, 2, and 0, respectively, with a sum of 5.

For a single objective alignment problem, finding the optimal alignment is a problem of searching for the least-cost path on the acyclic graph, with the length of the least-cost path being no longer than the sum of both sequences to be aligned. Needleman and Wunsch (1970) proposed a dynamic programming (DP) algorithm to find the least-cost path for unidimensional sequences (for details, see Appendix A). Although the DP is useful in finding optimal alignments, it is not suitable when alignments must be evaluated using multiple objectives. Finding optimal solutions to a multiobjective problem often requires a search algorithm. From the graph, the number of possible alignments between two sequences of length n is approximately $2^{2n} / \sqrt{\pi n}$ (Waterman 1989), which makes exhaustive enumeration of all possible alignments for long sequences infeasible. The literature generally suggests that heuristic methods, such as evolutionary algorithms, are efficient for finding high quality solutions to multiobjective problems (see, for example, Deb 2001; Xiao et al. 2007).

Encoding of alignment solutions

Each individual solution (i.e., an alignment in this paper) in an EA also is called a chromosome, a term borrowed from evolutionary biology, and must be encoded as a series of genes in an effective form in order for the EA to work. In our research, we encoded an alignment between two sequences as a path between the *start* and *end* points in a directed acyclic graph (Figure 3a). Let m and n be the length of two sequences to be aligned. To represent the path, we use a chromosome of variable length composed by J genes, where $\text{MAX}(m, n) \leq J \leq m+n$. Each gene contains two bits representing the direction of moves of an arrow (Figure 3b). The first bit represents the horizontal direction, and the second bit represents the vertical direction. A value of 1 means a move exists along that direction, 0 denotes otherwise. Therefore, a horizontal or vertical arrow can be encoded as {10} or {01} genes in a chromosome, respectively. A diagonal arrow can be treated as the combination of moves along both horizontal and vertical dimensions, and is encoded as {11}. The alignment in Figure 3a is encoded as the chromosome in Figure 3c with 6 genes.

Given the encoding strategy, the initial population can be randomly generated at the beginning of an algorithm run. Each individual in the initial population is created by simulating a path from the *start* to *end* points. The initialization operation first generates a chromosome of $m+n$ genes, and then each gene is randomly assigned one of the three directions {01, 10, and 11} with an equal weight. This procedure can yield invalid solutions. For example, a chromosome of 9 genes containing the {11} direction is invalid for aligning X and Y . A repair process needs to be used to ensure the validity of each initial solution. Let m be the number of columns (i.e., the number of horizontal arrows in a line) and n the number of rows (i.e., the number of vertical arrows) in the graph. In Figure 3a, we have $m = 5$ and $n = 4$. To repair a solution, we trace the path from the start node until it reaches the last row or column, whichever appears first, and count the number of arrows (l'), rows (n'), and columns (m'). If $n' < n$, we assign {01} to $n-n'$ genes, starting with position $l'+1$, and the final length of the chromosome is $l'+n-n'$. Similarly, if $m' < m$, we assign {10} to $m-m'$ genes, starting with position $l'+1$, making the final length of the chromosome $l'+m-m'$. Otherwise, the chromosome is valid and the length is l' .

Fitness assignment and selection

Each individual solution in the current population of an EA can be evaluated using the two objective functions defined by Equation (3). An iterative Pareto ranking procedure (Fonseca and Fleming 1993) can be used so that the non-dominated solutions in the current population are ranked with a rank level initially set to 0. We then increase the rank level by 1, and use it to rank the non-dominated solutions for the unranked individuals in the remaining population. This process repeats until all individuals in the population are ranked (see the ranks in Figure 2). After ranking, the following fitness value can be assigned to each individual:

$$f_i = 0.5 + \frac{N_{\text{pop}} - R_i}{N_{\text{pop}}} \quad (4)$$

where i denotes the i -th individual with a rank R_i , N_{pop} is the size of or number of individuals in the population, and the constant 0.5 is used to ensure that the minimum fitness value is 0.5 and that each individual has a chance to be selected.

To encourage diversity in the solutions, the fitness value of each individual is adjusted using a technique called fitness sharing (Goldberg and Richardson 1987), such that

$$f'_i = \frac{f_i}{\sum_j N_{\text{pop}} \text{sh}(d_{ij})} \quad (5)$$

where function $\text{sh}(d)$ is defined as

$$\text{sh}(d) = \begin{cases} 1 - \left(\frac{d}{\sigma_{\text{share}}}\right)^\alpha & \text{if } d < \sigma_{\text{share}} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where α is a constant coefficient and σ_{share} is a distance threshold. Parameter σ_{share} defines a neighbourhood such that fitness values of solutions in that neighbourhood decrease. A commonly used α value is 2, while σ_{share} often is problem specific. After sharing, fitness values for solutions that are close to each other in the solution space (e.g., *A* and *D* in Figure 2) decreased, while solutions around less crowded areas (e.g., *C* in Figure 2) receive higher fitness values so that they are more likely to be selected in the next generation, making the search process more explorative.

After each solution is assigned a fitness value, a selection operator is used to determine individuals chosen to produce the offspring. We used a method called stochastic universal sampling (Baker 1987). In this method, the individuals in a population are mapped onto continuous line segments, where the length of each segment equals the fitness value of the corresponding individual. To select M individuals from the population, we first define a step interval as $P = \sum_{i=1}^{N_{\text{pop}}} f'_i / M$, and then randomly generate a number between 0 and P . Using that random number as the starting position, the process advances M steps and the individual line segments containing these steps subsequently are selected. Individuals with high fitness values may be selected multiple times. The selected individuals are shuffled before they are used to create new solutions.

Recombination

A recombination operation is used to generate a new alignment from two selected parent alignments. However, conventional recombination operators cannot be directly employed in our problem because our special encoding method allows for variable lengths of chromosomes (different paths may have different lengths) and the validity of a chromosome (i.e., {00} is not a valid direction code for a gene). We developed a new recombination operator for our research (Figure 4).

The chromosome representations for the two parent alignments in Figures 4a and 4b are {11,10,11,11,11} and {11,01,11,10,11,10}, respectively (note the different lengths). A recombination point RP is randomly located on the path of parent 1, and the segment from *start* to RP in parent 1, {11,10}, is directly copied to the offspring. Then segments of parent 2 are copied backward from the *end* point to a point either directly underneath or right at the RP point, forming a temporary offspring of {11,10,*,*, 10,11,10}, where the '*' symbol indicates the directions that are not found in the two parents (Figure 4c). Finally, a sequence of vertical or horizontal arrows is used to replace the '*' symbols to complete the offspring as {11, 10, 01, 01, 10, 11, 10} (Figure 4d). This recombination operator ensures the validity of the offspring, and attempts to keep as much of the genetic information from both parents as possible. In our design, each pair of parents is applied twice, so that two children are generated. In the EA, a parameter called recombination probability (P_{rec}) is used to control the execution of the recombination operation. If a random number drawn from a uniform distribution between 0 and 1 is smaller than P_{rec} , the recombination operation is executed twice to yield two new individuals; otherwise, the two parents are directly copied to the new population.

Mutation

A mutation operation is used to slightly alter a solution to introduce new genetic information and to increase the diversity of a population. In a mutation operation, an individual is randomly picked from the newly generated offspring with a mutation probability (P_{mut}). Figure 5 illustrates the mutation process developed in this study. In this example, the selected individual is encoded as {11,11,10,01,11,10}. A short segment between two randomly selected points along the path of the alignment is replaced by another random segment between these two points. In our example, the original segment between points P_1 and P_2 , {11,10} are replaced by {10,11}, and we have a new chromosome {11,10,11,01,11,10}.

Local search

One critical feature of an EA design is balance between exploiting existing information using operations such as a recombination, and exploring new information using mutation operations (Eiben and Schippen 1998). Although the mutation operation generally can find new information that has not been explored before, researchers have demonstrated that incorporating other local search mechanisms in an EA can effectively increase search performance (see, for example, Xiao 2006). In our research, if a new solution undergoes a mutation process, we also apply a DP to a subset of a solution to improve its Pareto ranking. First this process randomly chooses a short segment along the path on the graph (e.g., from P_1 to P_2 in Figure 5b), and then one of the dimensions is randomly selected to construct two unidimensional subsequences. Next, the DP algorithm developed by Needleman and Wunsch (1970) is used to obtain an optimal alignment of the two subsequences, which is used to replace the original alignment. Appendix A formulates a brief description of the DP algorithm.

Summary

An outline of the MOEA for MDSA is provided in Figure 6. The MOEA continues until reaching the maximum number of generations ($MaxGen$) allowed. The algorithm selects a total of $(1-r_{elite}) \times N_{pop}$ individuals from a population (Line 8), and uses the selected individuals to create the same number of offspring (Line 9). Here, r_{elite} is a ratio between 0 and 1. A portion of the offspring undergo mutation and the DP algorithm (Line 10). Finally the offspring are inserted into a population (Line 11) to replace $100 \times (1-r_{elite})$ percent of its individuals. During each generation, the highest ranked $r_{elite} \times N_{pop}$ individuals in the current population are kept. This technique, called elitism (Deb 2000; Coello Coello et al. 2007), is used to maintain the best solutions found through EA generations. Together with the fitness sharing method [Equation (5)], the EA is encouraged to search for a diverse set of non-dominated solutions to the problem. Line 12 is used to ensure that no duplicated solutions are present in the population, which may occur after the recombination, mutation, and local search operations. In our MOEA, duplicated individuals are removed, and new individuals are generated randomly and inserted into the population.

Computational Experiments

We implemented the MOEA for MDSA in Matlab™ R2006b, using the Genetic and Evolutionary Algorithm Toolbox for Matlab (GEATbx, see Pohlheim 2006) library. All experiments were conducted using a computer with an AMD Athlon™ 64X2 Dual Core CPU of 2.81 GHz and a RAM of 3 GB. The data used to test the performance of the MOEA were individual activity-travel diaries collected in the Portland Metropolitan Region, Oregon through a 1994–95 Household Activity and Travel Behaviour Survey (Kwan 2000; Weber and Kwan 2002, 2003). During the survey, each individual was asked to provide detailed information about household and personal characteristics, and details of all out-of-home and in-home activities lasting at least 30 minutes in a two-day period assigned to them. Attributes collected include activity locations, travel mode, the time activities started and ended, and the purposes of travel activities.

To evaluate output from the MOEA, 50 car-drivers' activity-travel diaries from 30 households were selected and coded into 50 two-dimensional sequences. The first dimension describes the type of activities using a capital letter, and 13 letters are used to represent 13 types of activities (Table 1). The second dimension uses a three-digit number to represent the census block group of an undertaken activity - there are 465 census block groups. The block group scale was chosen to represent the location of activities, rather than using certain nominal variables (e.g., "home" and "work") to capture the changing patterns of activity locations. The events in these sequences include daily activities at a 10-minute interval, along with different activities occurring within that time interval. All sequences have at least 144 activities in the 24-hour time span. Figure 7 shows two example sequences of persons' daily space-time activity patterns.

Configuring MOEA parameters

The MOEA requires a set of parameters, including a population size or N_{pop} , a maximum number of generations ($MaxGen$), a P_{rec} , a P_{mut} , a ratio of elites (r_{elite}), and a fitness sharing coefficient (σ_{share}). The literature does not furnish clear guidelines about how to choose the r_{elite} value. We choose a value of 0.1 so that 90 percent of the population becomes new individuals during each generation. The value of σ_{share} also is dependent on the problem under study, and we chose a value of 20 after many trial-and-error tests.

For the other parameters, we developed a set of experiments to explore their impacts on EA performance. The goal was to undertake a fair and quick exploration of the effectiveness of different parameter combination in order to guide parameter choices in future work. We used the two sequences in Figure 7 to run our tests. After completing the MOEA, we computed the equally weighted sum of the two objective function values for each non-dominated solution using the following equation:

$$SP = \frac{1}{2} \sum_{i=1}^2 SP_i \quad (7)$$

where SP_i is the objective function of the i -th dimension defined in equation (3). The minimum SP value in the non-dominated solutions is used to indicate the EA performance.

We pre-set the values to be tested for the parameters (Table 2). Each combination of these parameter values was tested for 10 runs of the MOEA, with the resulting SP values averaged.

Figure 8 presents impacts of N_{pop} , P_{rec} , and P_{mut} , with *MaxGen* set to 100. The x-axis represents the mutation probability, the y-axis represents the average SP value or running time, and the panel variable is the number of individuals used in each generation. Figure 8a shows that small SP scores can be achieved with high mutation probabilities, especially for relative small population sizes (e.g., 50 and 100). SP scores tend to decrease with an increase in population size. However, increasing population size and mutation probability may significantly increase computing time, especially for sizes greater than 200 (Figure 8b). When choosing the mutation probability and the population size, a trade-off needs to be made between the distance score and time cost to achieve reasonable and efficient results. In this case, we chose an N_{pop} of 200 and a P_{mut} of 0.1.

Figure 9 shows the impacts of *MaxGen*, along with P_{rec} and P_{mut} , on the MOEA performance. N_{pop} is set to 200 for these experiments. Here the average distance score does not significantly decrease when *MaxGen* is greater than or equal to 100 (Figure 9a), although computing time dramatically increases when *MaxGen* is greater than 200 (Figure 9b). In other experiments, we set the value of *MaxGen* to 100 for efficiency considerations. Finally, Figures 8 and 9 do not show significant impacts of different recombination probabilities on either average distance scores or running time. Therefore, we set P_{rec} to 0.8.

Evaluation of MOEA performance

We first demonstrate the effectiveness of the MOEA using a pair of sequences from our data set. Figure 10 shows the objective space using the non-dominated solutions found in different generations. This figure clearly shows that the MOEA maintains a population of solutions that moves toward the lower-left part of the graph (i.e., from those marked as 1 and 2 to those as plus signs). Some solutions can be found in many generations (e.g., those marked as both 1 and 2, or those marked with both a plus and a square), indicating that elitism has effectively retained non-dominated solutions in the population. Compared with the results from a run without using the fitness sharing method (the grey circles in Figure 10), the sharing mechanism proves useful in promoting diversity of the solution population, because more solutions produce a more complete covering of the objective space.

Next, we compare the MOEA with the existing software package ClustalG, using all 50 sequences for 1,225 pairwise alignments (excluding self alignment). ClustalG is chosen here because it is able to maintain the interdependency relationships among activities when measuring the distances between space-time activity patterns. Because ClustalG does not search for non-dominated solutions in a multiobjective context, we again use the SP value [equation (7)] of each alignment found by ClustalG for assessment. For each MOEA run, we pick the non-dominated solution with the smallest SP value. For example, in the results of the specific pair of sequences shown in Figure 10, the solution that exhibits the smallest sum is marked as an open triangle (the distance scores of 29 and 127), whereas the solution found by ClustalG is marked as the green triangle (with distance scores of 45 and 127). In this

particular example, the MOEA can find an alignment that exhibits a smaller SP value than that found by ClustalG.

For each alignment problem, we run the MOEA 20 times. No significant difference exists between the minimum and mean SP values in the 20 runs of MOEA ($t = -0.2125$ and $p\text{-value} = 0.8318$) for the 1,225 alignments. This finding indicates the robustness of the MOEA because multiple runs yield similar results. Comparing the mean SP values from the 20 MOEA runs with ClustalG results, for the 1,225 alignments, 1,202 SP values from MOEA are lower than those from ClustalG, and only 23 MOEA SP values are higher. Figure 11 shows the QQ-plot between the SP values obtained with the two methods; the dots are mostly above the diagonal line. The SP values yielded by the two methods exhibit a significant difference ($t = -52.0781$, $p\text{-value} < 2.2\text{e-}16$). These numbers confirm our analysis in the second section of the paper: ClustalG searches the optimal alignment for combined elements, but cannot guarantee true optimal solutions for each separate dimension.

Conclusions and Discussion

This paper presents a new and unique methodology for evaluating the similarity among individual activity patterns: conceptualizing sequence alignment as a multiobjective optimization problem, and solving this problem with EAs. As the results indicate, the MOEA developed in this study provides an effective approach for assessing activity pattern similarity, and a foundation for identifying distinctive groups of individuals with similar activity patterns. The study also demonstrates that our method outperforms ClustalG for most of the cases we examined.

Although various methods have been used in past studies by geographers, geographic information scientists, and transportation researchers, most of these methods are not effective for handling the composition, interdependency, and structure of people's everyday activities, and few studies to date have explicitly incorporated geographic location into the analysis (with the notable exception of Shoval and Isaacson 2007). Existing studies tend to focus more on extracting the geometric and statistical properties of movement trajectories than on the substantive characteristics of human activities (e.g., activity purpose). In our approach, time serves as the referential framework for organizing all other activity dimensions, which unfold and change over time with an activity sequence. Activity location and other semantic attributes of human activities are explicitly represented and simultaneously registered to specific time points (or intervals) in each sequence, and are encoded with specific characters or numbers (see Figures 1b and 7). Furthermore, many semantic attributes of human activities besides location and time can be taken into account simultaneously using the method summarized in this paper. The most prominent strength of multidimensional sequence alignment is that it does not reduce the dimensionality of a problem (it assesses activity pattern similarity using many dimensions at the same time). Instead, it entails a trade-off between two or more dimensions (see Figures 2 and 10), which potentially allows researchers to gain more insight into human space-time behaviour. However, it is accomplished at the cost of increased computational intensity.

Specifically, our method conceives multidimensional sequence alignment as a multiobjective optimization problem that aims to minimize the distance scores for all dimensions simultaneously, which not only helps to maintain the interdependency relationships of different activity attributes, but also provides the flexibility of handling various numbers of attribute dimensions. When compared with other algorithms, our MOEA can use different distance measures for different activity dimensions through defining a corresponding substitution matrix in its objective function. For example, a spatial proximity matrix can be defined for the spatial dimension to compare real geographic locations, and various semantic matrices can be defined for comparing activity pattern differences with respect to specific attribute dimensions (e.g., activity type or activity purpose). This flexibility removes the constraints of existing algorithms, and lays a foundation for the clustering of movement trajectories of humans or objects with many interacting dimensions. Using the MOEA, the number of activity dimensions that can be handled for high-dimensional activity data becomes flexible by defining a corresponding number of objectives in multidimensional solution space.

However, significant challenges to using MDSA and MOEA remain. For example, because time is the referential framework for organizing all other activity dimensions, analysis of motion patterns or extracting space-time patterns (e.g., cyclical or repeating movements) is difficult. Further, when comparing patterns with primary and secondary activities, coding human activities appropriately and defining a good substitution cost matrix to be used in the objective function are important. For example, given three persons' daily activities of "home-work-home", "home-work-lunch-work-home" and "home-shopping-home", if the distance measure is defined as the number of character edit operations, the distance score between the first and the third sequences is smaller than that between the first and the second. However, the first two sequences should be more similar because both are commute-related activity patterns. Therefore, sequence alignment depends heavily on how activity sequences are coded, and how the substitution cost matrix is defined. In this case, one option is to code the sequences with the consideration of activity duration, with each character representing a certain amount of time, which eliminates the biases caused by sequence length differences. Because secondary activities often are dependent on primary activities, in the substitution cost matrix, smaller substitution costs should be defined for primary-secondary activities than for any other activities. Effectively defining substitution costs requires a good substantive understanding of human activity-travel behavior and domain-specific expert knowledge.

Another topic for future research is the computational intensity of sequence alignment methods and EAs. One problem relates to the combinatorial explosion as a result of the use of finer spatial and temporal granularity in sequence alignment (Joh, Arentze and Timmermans 2001). In this study, for example, activity attributes are organized along a temporal sequence of 144 time intervals (10-minute intervals over a 24-hour period), and 465 geographic locations are explicitly coded (the study area in Shoval and Isaacson [2007] has only 26 zones). When finer temporal and spatial scales are used, the combinatorial explosion becomes significant, and poses a considerable challenge for applying sequence alignment. Furthermore, when compared to the DP method used in ClustalG, the MOEA is more computationally intensive due to its iterative nature. To increase the search speed and

make the algorithm feasible for tackling large problems, the single-population MOEA needs to be extended to a multiple-population MOEA, and implemented on a massively parallel architecture for high-dimensional applications (Xiao and Armstrong 2003). In addition, although we have proposed the MOEA to compute the distance between multidimensional sequences, the test of the reliability and significance of the distance scores for multidimensional sequences needs to be investigated in future research (Wilson 2006).

Acknowledgments

Mei-Po Kwan's work on this article was supported by the following grants: NSF BCS-1244691, NIH R01DA032371-01, and NSFC 41228001. In addition, part of this research was supported by a grant from National Natural Science Foundation of China (No.71272030).

References

- Abbott A. Sequence Analysis: New Methods for Old Ideas. *Annual Review of Sociology*. 1995; 21:93–113.
- Back, T.; Homeister, F. Extended Selection Mechanisms in Genetic Algorithms. In: Belew, RK.; Booker, LB., editors. *Proceedings of the Fourth International Conference on Genetic Algorithms*; 13–16 July 1991; San Diego, USA. San Francisco, CA: Morgan Kaufmann; 1991. p. 92–99.
- Baker, JE. Reducing Bias and Inefficiency in the Selection Algorithm. In: Grefenstette, JJ., editor. *Proceedings of the Second International Conference on Genetic Algorithms and their Application*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1987. p. 14–21.
- Bargeman B, Joh C-H, Timmermans H. Vacation Behavior Using a Sequence Alignment Method. *Annals of Tourism Research*. 2002; 29:320–37.
- Becken S, Simmons D, Frampton C. Segmenting Tourists by Their Travel Pattern for Insights into Achieving Energy Efficiency. *Journal of Travel Research*. 2003; 42(1):48–56.
- Burnett P, Hanson S. The Analysis of Travel as an Example of Complex Human Behavior in Spatially-constrained Situations: Definition and Measurement Issues. *Transportation Research A*. 1982; 16:87–102.
- Cai, L.; Juedes, D.; Liakhovitch, E. Congress on Evolutionary Computation. Piscataway, NJ: IEEE Service Center; 2000. Evolutionary Computation Techniques for Multiple Sequence Alignment; p. 829–35.
- Cha S, McCleary KW, Uysal M. Travel Motivations of Japanese Overseas Travelers: A Factor-cluster Segmentation Approach. *Journal of Travel Research*. 1995; 34(1):33–39.
- Chen J, Shaw S-L, Yu H, Lu F, Chai Y, Jia Q. Exploratory Data Analysis of Activity Diary Data: A Space-time GIS Approach. *Journal of Transport Geography*. 2011; 19:394–404.
- Coello Coello, CA.; Lamont, GB.; Van Veldhuizen, DA. *Evolutionary Algorithms for Solving Multi-Objective Problems*. 2. Berlin: Springer; 2007.
- Cohon, JL. *Multiobjective Programming and Planning*. New York: Academic Press; 1978.
- Deb, K. *Multi-Objective Optimization Using Evolutionary Algorithms*. New York: Wiley; 2001.
- Demšar U, Virrantaus K. Space-time Density of Trajectories: Exploring Spatio-temporal Patterns in Movement Data. *International Journal of Geographical Information Science*. 2010; 24(10):1527–42.
- Dodge S, Laube P, Weibel R. Movement Similarity Assessment Using Symbolic Representation of Trajectories. *International Journal of Geographical Information Science*. 2012; 26(9):1563–88.
- Eiben AE, Scippers CA. On Evolutionary Exploration and Exploitation. *Fundamenta Informaticae*. 1998; 35(1–4):35–50.
- Fonseca, CM.; Fleming, PJ. Genetic Algorithms in Multiobjective Optimization: Formulation, Discussion and Generalization. In: Forrest, S., editor. *Proceedings of the Fifth International Conference on Genetic Algorithms*. San Mateo, CA: Morgan Kaufmann; 1993. p. 416–23.

- Gao Y, Zheng B, Chen G, Li Q. Algorithms for Constrained k-nearest Neighbor Queries over Moving Object Trajectories. *Geoinformatica*. 2010; 14:241–76.
- Goldberg, DE. Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley; 1989.
- Goldberg, DE.; Richardson, J. Genetic Algorithms with Sharing for Multimodal Function Optimization. In: Grefenstette, JJ., editor. Genetic Algorithms and Their Applications: Proceedings of the Second International Conference on Genetic Algorithms. Hillsdale, NJ: Lawrence Erlbaum Associates; 1987. p. 41-49.
- Gonnet GH, Korostensky C, Benner S. Evaluation Measures of Multiple Sequence Alignments. *Journal of Computational Biology*. 2000; 7:261–76. [PubMed: 10890401]
- Gonzalez MC, Hidalgo CA, Barabási A-L. Understanding Individual Human Mobility Patterns. *Nature*. 2008; 453:779–82. [PubMed: 18528393]
- Griffith DA, Chun Y, O’Kelly ME, Berry BJL, Haining RP, Kwan M-P. *Geographical Analysis: Its First Forty Years*. Geographical Analysis. 2013; 45(1):1–27.
- Han, JW.; Gao, J. Research Challenges for Data Mining in Science and Engineering. In: Kargupta, H.; Han, JW.; Yu, P.; Motwani, R.; Kumar, V., editors. Next Generation of Data Mining. New York: CRC Press; 2009. p. 1-25.
- Hanson S, Huff J. Classification Issues in the Analysis of Complex Travel Behavior. *Transportation*. 1986; 13:271–93.
- Hornsby KS, Cole S. Modeling Moving Geospatial Objects from an Event-based Perspective. *Transactions in GIS*. 2007; 11(4):555–73.
- Joh C-H, Arentze TA, Timmermans HJT. Multidimensional Sequence Alignment Methods for Activity-travel Pattern Analysis. *Geographical Analysis*. 2001a; 33(3):247–70.
- Joh C-H, Arentze T, Timmermans H. Pattern Recognition in Complex Activity Travel Patterns: Comparison of Euclidean Distance, Signal-processing Theoretical, and Multidimensional Sequence Alignment Methods. *Transportation Research Record*. 2001b; 1752:16–22.
- Joh C-H, Arentze T, Timmermans H. A Position-sensitive Sequence-alignment Method Illustrated for Space-time Activity-diary Data. *Environment and Planning A*. 2001c; 33:313–38.
- Joh C-H, Arentze T, Hofman F, Timmermans H. Activity Pattern Similarity: A Multidimensional Sequence Alignment Method. *Transportation Research Part B: Methodological*. 2002; 36(5):385–403.
- Joh C-H, Timmermans HJP, Popkowski-Leszczyc PTL. Identifying Purchase-history Sensitive Shopper Segments Using Scanner Panel Data and Sequence Alignment Methods. *Journal of Retailing and Consumer Services*. 2003; 10:135–144.
- Koppelman, FS.; Pas, EI. Travel-activity Behavior in Time and Space: Methods for Representation and Analysis. In: Nijkamp, P.; Leitner, H.; Wrigley, N., editors. Measuring the Unmeasurable. The Hague: Martinus Nijhoff; 1985. p. 587-623.
- Kruskal, JB. An Overview of Sequence Comparison. In: Sankoff, D.; Kruskal, J., editors. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Reading, MA: Addison-Wesley; 1983. p. 1-44.
- Kwan M-P. Gender, the Home-work Link, and Space-time Patterns of Non-Employment Activities. *Economic Geography*. 1999; 75(4):370–94.
- Kwan M-P. Interactive Geovisualization of Activity-travel Patterns Using Three-dimensional Geographical Information Systems: A Methodological Exploration with a Large Data Set. *Transportation Research C*. 2000; 8:185–203.
- Kwan M-P. GIS Methods in Time-geographic Research: Geocomputation and Geovisualization of Human Activity Patterns. *Geografiska Annaler B*. 2004; 86:267–80.
- Kwan M-P. The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers*. 2012; 102(5):958–68.
- Kwan M-P. Beyond Space (As We Knew It): Toward Temporally Integrated Geographies of Segregation, Health, and Accessibility. *Annals of the Association of American Geographers*. 2013; 103(5)
- Laube P, Dennis T, Forer P, Walker M. Movement Beyond the Snapshot: Dynamic Analysis of Geospatial Lifelines. *Computers, Environment and Urban Systems*. 2007; 31:481–501.

- Long JA, Nelson TA. A Review of Quantitative Methods for Movement Data. *International Journal of Geographical Information Science*. 2013; 27(2):292–318.
- Ma J, Goulias KG. A Dynamic Analysis of Person and Household Activity and Travel Patterns Using Data from the First two Waves in the Puget Sound Transportation Panel. *Transportation*. 1997; 24:309–31.
- Miller, HJ.; Han, JW. Geographic Data Mining and Knowledge Discovery: An Overview. In: Miller, HJ.; Han, JW., editors. *Geographic Data Mining and Knowledge Discovery*. 2. New York: CRC Press; 2009. p. 1-26.
- Myka, A.; Güntzer, U. Fuzzy Full-Text Searches in OCR Databases. In: Adams, NR.; Bhargava, BK.; Halem, M.; Yesha, Y., editors. *Digital Libraries: Research and Technology Advances, ADL'95 Forum*. Berlin: Springer; 1996. p. 131-148.
- Needleman SB, Wunsch CD. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*. 1970; 48:443–53. [PubMed: 5420325]
- Notredame C. Recent Progress in Multiple Sequence Alignment: A Survey. *Pharmacogenomics*. 2002; 3:131–44. [PubMed: 11966409]
- Notredame C, Higgins DG. SAGA: Sequence Alignment by Aesthetic Algorithm. *Nucleic Acids Research*. 1996; 24:1515–24. [PubMed: 8628686]
- Orellana D, Wachowicz M. Exploring Patterns of Movement Suspension in Pedestrian Mobility. *Geographical Analysis*. 2012; 43(3):241–60. [PubMed: 22073410]
- Palmer JR, Espenshade TJ, Bartumeus F, Chung CY, Ozgencil NE, Li K. New Approaches to Human Mobility: Using Mobile Phones for Demographic Research. *Demography*. 2013 forthcoming.
- Pas EI. A Flexible and Integrated Methodology for Analytical Classification of Daily Travel-activity Behavior. *Transportation Science*. 1983; 17:405–29.
- Persson O, Ellegård K. Torsten Hägerstrand in the Citation Time Web. *The Professional Geographer*. 2012; 64(2):250–61.
- Pohlheim, H. [last accessed September 9, 2013] GEATbx: Introduction. 2006. <http://www.geatbx.com>
- Ratti C, Williams S, Frenchman D, Pulselli R. Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis. *Environment and Planning B: Planning and Design*. 2006; 33:727.
- Raubal M, Miller HJ, Bridwell S. User-centred Time Geography for Location-based Services. *Geografiska Annaler B*. 2004; 86:245–265.
- Recker WW, McNally MG, Root GS. Travel/activity Analysis: Pattern Recognition, Classification, and Interpretation. *Transportation Research A*. 1985; 19:279–96.
- Richardson DB, Volkow ND, Kwan M-P, Kaplan RM, Goodchild MF, Croyle RT. Spatial Turn in Health Research. *Science*. 2013; 339(6126):1390–92. [PubMed: 23520099]
- Sadahiro Y, Lay R, Kobayashi T. Trajectories of Moving Objects on a Network: Detection of Similarities, Visualization of Relations, and Classification of Trajectories. *Transactions in GIS*. 2013; 17(1):18–40.
- Schlich R, Axhausen KW. Habitual Travel Behaviour: Evidence from a Six-Week Travel Diary. *Transportation*. 2003; 30:13–36.
- Schlich, R.; Axhausen, KW. *Arbeitsbericht Verkehrs- und Raumplanung*. Vol. 296. Zürich: ETH Zürich; 2004. Analysing interpersonal variability for Homogeneous Groups of Travellers. IVT
- Sinha G, Mark DM. Measuring Similarity between Geospatial Lifelines in Studies of Environmental Health. *Journal of Geographical Systems*. 2005; 7(1):115–36.
- Shen Y, Kwan M-P, Chai Y. Investigating Commuting Flexibility with GPS Data and 3D Geovisualizations: A Case Study of Beijing, China. *Journal of Transport Geography*. Forthcoming.
- Shoval N, Isaacson M. Sequence Alignment as a Method for Human Activity Analysis in Space and Time. *Annals of the Association of American Geographers*. 2007; 97(2):282–97.
- Smyth, CS. Mining Mobile Trajectories. In: Miller, HJ.; Han, J., editors. *Geographic Data Mining and Knowledge Discovery*. New York: Taylor & Francis; 2001. p. 337-61.
- Stovel K, Bolan M. Residential Trajectories: Using Optimal Alignment to Reveal the Structure of Residential Mobility. *Sociological Methods & Research*. 2004; 32:559–98.

- Tan KC, Lee TH, Khoo D, Khor EF. A Multiobjective Evolutionary Algorithm Toolbox for Computer-aided Multiobjective Optimization. *IEEE Transactions on Systems, Man and Cybernetics, Part B*. 2001; 31:537–56.
- Van Veldhuizen DA, Lamont GB. Multiobjective Evolutionary Algorithms: Analyzing the State-of-the-art. *Evolutionary Computation*. 2000; 8:125–47. [PubMed: 10843518]
- Wang D, Jiang T. On the Complexity of Multiple Sequence Alignment. *Journal of Computational Biology*. 1994; 1:337–48. [PubMed: 8790475]
- Waterman, MS. *Mathematical Methods for DNA Sequences*. Boca Raton, FL: CRC Press; 1989.
- Weber J, Kwan M-P. Bringing Time Back in: A Study on the Influence of Travel Time Variations and Facility Opening Hours on Individual Accessibility. *The Professional Geographer*. 2002; 54:226–40.
- Weber J, Kwan M-P. Evaluating the Effects of Geographic Contexts on Individual Accessibility: A Multilevel Approach. *Urban Geography*. 2003; 24:647–71.
- Wilson C. Activity Pattern Analysis by Means of Sequence-alignment Methods. *Environment and Planning A*. 1998a; 30:1017–38.
- Wilson C. Analysis of Travel Behavior Using Sequence Alignment Methods. *Transportation Research Record*. 1998b; 1465:52–59.
- Wilson C. Activity Patterns of Canadian Women: An Application of ClustalG Sequence Alignment Software. *Transportation Research Record*. 2001; 1777:55–67.
- Wilson C. Reliability of Sequence-alignment Analysis of Social Processes: Monte Carlo Tests of ClustalG Software. *Environment and Planning A*. 2006; 38:187–204.
- Wilson, C.; Harvey, A.; Thompson, J. ClustalG: Software for analysis of activities and sequential events. Paper presented at the Workshop on Longitudinal Research in Social Science: A Canadian Focus; Windermere Manor, London, Ontario, Canada. October 25–27, 1999; 1999.
- Xiao N. An Evolutionary Algorithm for Site Search Problems. *Geographical Analysis*. 2006; 38(3): 227–47.
- Xiao, N.; Armstrong, MP. A Specialized Island Model and its Application in Multiobjective Optimization. In: Cantú-Paz, E., et al., editors. *Genetic and Evolutionary Computation — GECCO 2003*. Lecture Notes in Computer Science. Vol. 2724. Berlin: Springer; 2003. p. 1530–40.
- Xiao N, Bennett DA, Armstrong MP. Using Evolutionary Algorithms to Generate Alternatives for Multiobjective Site Search Problems. *Environment and Planning A*. 2002; 34(4):639–56.
- Xiao N, Bennett DA, Armstrong MP. Interactive Evolutionary Approaches to Multiobjective Spatial Decision Making: A Synthetic Review. *Computers, Environment and Urban Systems*. 2007; 31:232–252.
- Zhang, G.; Huang, D. Aligning Multiple Protein Sequence by an Improved Genetic Algorithm. *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks*; 2004. p. 1179–83.
- Zhang C, Wong AKC. A Genetic Algorithm for Multiple Molecular Sequence Alignment. *Bioinformatics*. 1997; 13:565–81.
- Zhu, FD.; Yang, XF.; Han, JW.; Yu, P. Mining Frequent Approximate Sequential Patterns.” In: In: Kargupta, H.; Han, JW.; Yu, P.; Motwani, R.; Kumar, V., editors. *Next Generation of Data Mining*. New York: CRC Press; 2009. p. 66–87.

Appendix A: Dynamic Programming for Sequence Alignment

To briefly describe the dynamic programming method developed by Needleman and Wunsch (1970), we use A and B to denote two sequences to be aligned, and A_i ($1 \leq i \leq m$) and B_j ($1 \leq j \leq n$) to denote the i -th and j -th characters in the sequences, respectively. We construct an $m \times n$ matrix and use $S(i, j)$ to denote the value at cell (i, j) . Before filling in the values in the matrix, we set $S(0, 0) = 0$, $S(i, 0) = i \times d$, and $S(0, j) = j \times d$, where d is the cost of inserting a gap into a sequence. Using the costs defined in equation (2), we have $d = 1$. Moreover, we also apply equation (2) to compute $d(A_i, B_j)$ as the distance score between the i -th and j -th

characters in A and B , respectively. We then recursively set $S(i, j)$ to be the minimum of three values: $S(i-1, j-1) + d(A_i, B_j)$, $S(i, j-1) + d$, and $S(i-1, j) + d$, representing three edit operation directions: diagonal (substitution or identity), horizontal (inserting a gap into A after the i -th character), and vertical (inserting a gap into B after the j -th character), respectively. After completing the process, the directions chosen are used to trace back the optimal alignment starting from cell (m, n) , and hence the optimal cost is $S(m, n)$.

X: *L* – *H* *T* *H* 1 2 3 – 2
 Y: *L* *T* *W* *T* *H* 1 4 3 4 2

(a) Alignments for two separated dimensions

X: *L*1 –2 *H*3 *T* – *H*2
 Y: *L*1 *T*4 *W*3 *T*4 *H*2

(b) Combined optimal alignments

X: *L*1 – *H*2 *T*3 *H*2
 Y: *L*1 *T*4 *W*3 *T*4 *H*2

(c) Multidimensional alignment in ClustalG

Figure 1.

Multidimensional alignment for the two sequences $X=L1H2T3H2$ and $Y=L1T4W3T4H2$.

Figure 2. Solutions for a 2-objective optimization problem. Each dot is drawn using the two objective function values of the corresponding solution. The number associated with each dot is the Pareto ranking of that solution. The curve represents a hypothetical Pareto front.

Figure 3. Graph representation of a sequence alignment. (a) Alignment as a path on a directed acyclic graph. (b) Encoding edit operations. (c) The chromosome corresponding to the path in (a).

Figure 4. Recombination operation that uses the two parent solutions (a) and (b) to create a temporal solution (c) that it modifies to form the child solution (d).

Figure 5.
Mutation operator

1 $t \leftarrow 0$
 2 initialize population $G(t)$
 3 evaluate $G(t)$ using equation (3)
 4 obtain Pareto ranks and fitness values of individuals in $G(t)$
 5 **do**
 6 $t \leftarrow t + 1$
 7 $G(t) \leftarrow G(t-1)$
 8 select $(1-r_{\text{elite}}) \times N_{\text{ind}}$ individuals from $G(t)$
 9 create offspring by recombining the selected individuals
 10 perform mutation and the DP local search on the offspring
 11 insert offspring to $G(t)$
 12 evaluate $G(t)$ using equation (3)
 13 obtain Pareto ranks and fitness values of individuals in $G(t)$
 14 **while** ($t < \text{MaxGen}$)
 15 report non-dominated solutions in $G(t)$, $0 \leq t \leq \text{MaxGen}$

Figure 6.

Outline of the multiobjective evolutionary algorithm for MDSA.

(a) Average distance scores

(b) Average running time

Figure 8.

Relationships between weighted sum of distance scores (A) or running time (B) and recombination probabilities, population sizes, and numbers of individuals. The square and diamond symbols represent trends within and across panels, respectively.

(a) Average distance scores

(b) Average running time

Figure 9.

Relationships between weighted sum of distance scores (A) or running times (B) and recombination probabilities, maximum numbers of generations, and numbers of generations. The square and diamond symbols represent trends within and across panels, respectively.

Figure 10.

The solution space of aligning a pair of sequences. Distance scores for the two dimensions are the two axes. Non-dominated solutions found in each generation are used in this figure. The grey circles are from a run of the MOEA without using the fitness sharing method.

Figure 11.

QQ plot of the weighted sum of the objective function values (SP) for results from the MOEA (average of 20 runs) and ClustalG. Each circle represents the results for one of the 1,225 pairs.

Table 1

Coding of Daily Activities

Code	Activity Type
P	Pick-up/drop-off
S	Shopping
V	Services
M	Medical
E	Meals
R	Social/Recreation
H	Leisure, sport, hobbies
W	Work
U	Education
I	In-home activity
O	Other and unknown
T	Travel
L	Rest

Table 2

Selection of algorithm parameters

Parameters	Values
Number of individuals (<i>Nind</i>)	50, 100, 200, 500, 1000
Maximum number of generations (<i>MaxGen</i>)	25, 50, 100, 200, 500
Recombination probability (<i>RecProb</i>)	0.6, 0.8, 1.0
Mutation probability (<i>MutProb</i>)	0.05, 0.1, 0.25, 0.4