The Walking School Bus Routing Problem

Negin Bolkhanian, Matthew Reyers

Department of Mathematics, Simon Fraser University,
Surrey, BC, Canada

Abstract

This paper proposes four heuristic methods for the Walking School Bus Routing Problem (WSBRP). Our work first develops a simulation model based on school catchment regions and census data, creating a proportionally representative sample of students. Each sample is then clustered by four methods and a routing problem is solved on each of these clusters. Evaluation of clustering techniques is done through a scoring function applied to the generated routes and is compared with a modified version of an open source Vehicle Routing Problem (VRP) solver. Results indicate that clustering method based heuristics offer comparable and often improved results in comparison to VRP outputs in a significantly reduced time frame, ideal for the design of applets and real-time services.

1 Introduction

The Walking School Bus Routing Problem (WSBRP) is characterized by a collection of walking leaders that, in total, must service every student node in the network. Nodes are serviced in routes that originate at an individual leader and terminate at the school. The objective of the problem is to minimize cost incurred by the network in which cost is measured by time and distance traveled by all of the nodes in the network. Consideration is also put forth for leader capacity and distance thresholds as these values are not hard constraints in the WSBRP. This is done through a secondary scoring function for further route evaluation.

The WSBRP is a combinatorial optimization problem that designs routes to school from a given set of students, schools, leaders, locations, and roads. This classification as a combinatorial optimization problem may draw comparisons with the School Bus and Vehicle routing problems, though the management of route selection and use of dynamic leader allocation distinguishes the WSBRP [26] [23].

This paper presents a heuristic algorithm that evaluates student routing to school using the Google Maps API. The algorithm is built from 58 schools in the Greater Vancouver Region and utilizes synthetic data from a proportional sampling approach that integrates school catchment regions and census population data. Connection between the synthetic data and the API is achieved by clustering the data first and passing the resulting clusters to the API. Four clustering methods (kNN, k-means, Hierarchical, and Soft) are tested and compared with a publicly available Vehicle Routing Problem (VRP) solver on the same set of data [10]. Findings indicate that the four clustering methods, in combination with the Google Maps API, outperform the VRP on average. Further, the run time for each clustering heuristic is significantly shorter than the VRP, allowing the improved heuristics to be better implemented in potential web applications. Suggestions as to how to create this web application and a starting body of code are provided.

2 Background

Over the past 30 years, there has been a significant decrease in physical activity and increase in obesity rate amongst children. During these years, the obesity rate among children aged 5–17 in Canada has tripled [11]. One of the reasons for the decrease in physical activity is that children are routinely choosing activities that are less physically engaging, such as getting a ride to school instead of walking or biking [15]. According to Canada's Physical Activity Guidelines, children aged 5–17 need several hours of light activity weekly [6]. Children who walk or bike to school meet this objective at a higher rate than those that do not, indicating that active transportation to school can be an important component of healthy living.

To further encourage active transportation, new types of active transportation are being tested. One of these types is the Walking School Bus (WSB). Parents volunteer for their children to be

picked up by a group of other students, with an adult leader, and walk to school. As Walking School Buses are relatively new, there is not yet much infrastructure for them. Many Walking School Buses are as simple as two families taking turns walking their children to school though some organizations have started to build complex versions. The Walking School Bus organization, established in the USA, aims to help develop these programs across the country [22]. This program encourages parents and school officials to participate and collaborate together in running the program. They offer ways for communities to get started and present the best information available though they do not do any route setting.

In exploring the growing challenge of active transportation and childhood obesity, we look to make alternatives such as the Walking School Bus more accessible. Our approach focuses on the design of a system that would support an applet. This objective implies the need for a routing system that resolves rapidly. As the problem of routing Walking School Buses is comparable to Vehicle Routing Problems, similar heuristics could be applied. These heuristics, however, are time consuming to run as they often require many iterations [10]. To compensate for this, we devise a cluster first algorithm that separates the nodes into smaller clusters. Work has been done with cluster first route second algorithms though these have been in the context of VRPs and require different considerations than the WSB addressed by this paper [17]. The smaller clusters can then be solved accurately and quickly by the Google Maps API, resulting in a heuristic that is efficient and accurate.

2.1 Vancouver Safe Routes

In Vancouver there is not currently an official WSB program. As for school originated programs, no school in the Greater Vancouver Region currently holds their own WSB. Some Recreation Centres are adding in a WSB component to their before and after school programs, though those involve registration and commitment to the after school program. For example, Champlain Heights Community Centre in the city of Vancouver has a walking school bus program as an after school activity in which children are picked up from Captain Hook Elementary School as part of their paid after school care. This Recreation Centre is one of the few programs that currently run in the Greater Vancouver Region. Many YMCA's and Recreation Centres of Greater Vancouver have yet to adopt the WSB.

Active transportation to school has received some attention in recent years. The City of Vancouver has started the "School Active Travel Planning" program (SATP) [8] to encourage pupils to walk/bike to school. The program began in 2012 and many schools in the Vancouver area have participated. Their goal is to "improve the safety and comfort of walking and cycling to school, and to encourage more students and families to use active transportation modes to get to and from school" [8]. Route safety is one of the main factors that impacts parents' decisions to allow their kids to walk to school, an aspect of walking that The City of Vancouver is actively trying to improve [18]. The City of Vancouver studies each participating school separately and finds walkable areas requiring improvements and, as a result, creates safer ways to actively transport to school for students. Each

participating school has a map available online in PDF format that shows which routes are safe to walk or bike.

Although the work of SATP has cleaned up walking routes to encourage active transportation, the program does not facilitate any group organization. This means that programs such as the Walking School Bus are not significantly improved by SATP.

The work of the SATP, however, could be enhanced by a systematic approach to routing. This would be done through the introduction of an applet using the results of this paper. The routing that is selected by the applet can be required to either traverse or exclude certain roadways based on varying conditions. Pairing the safe routes with the WSB applet would allow for improved safe walking.

2.2 Implementation of School Travel Planning in Ontario

Green Communities Canada has been promoting active travel since 1996 [13]. They have conducted a feasibility study to implement School Travel Planning (STP), a community-based model of active transportation, in district school boards in Toronto and Wellington-Dufferin Guelph (WDG). The results of one year of coordination and implementation show positive travel behavior changes with safety, environmental, health and financial benefits [13]. Across thirteen schools studied, there was a 4.3% decrease in car use, a 3.2% increase in public transit and an overall 1% increase in cycling and walking [12]. Because of weather conditions, the increase in walking and biking is less than that of public transit. The study reported that during data collection a severe storm occurred in Toronto, preventing families from using active travel to go to school.

Similar to the Vancouver Safe Routes 2.1, finding efficient routes for a WSB based on the numbers of participants and their locations was not the objective of this study. The main focus of the pilot study was to evaluate the implementation of school travel planning. The results of the study indicate the potential for active transportation in Ontario. Therefore, we are expecting that the implementation of the WSB in Vancouver will demonstrate positive results as well.

2.3 Other Jurisdictions and Focus Areas

The accomplishment of the WSB has extended to a number of other influences and concentration areas, for example:

- The Canadian Cancer Society has implemented a WSB, called Trottibus Walking School Bus, in Quebec to encourage young people to integrate walking into their lifestyle [5]. The Trottibus WSB has since won The Play Exchange's grand prize and has been adopted by 100 schools, 30 of which are in Montreal [4].

- VicHealth in the state of Victoria (Australia) conducted a WSB program as a health promotion initiative for four resident committee areas in 2001. The success of the program generated more interest and by 2015 included 61 councils in the program [29].

- In Seattle (US), the Harborview Medical Center has used a WSB program to promote child safety in residential neighbourhoods. Also, a study done by BMC Public Health [20] evaluated the Seattle program as an effective strategy for reducing child obesity among urban, low-income elementary school students.

- A pilot study run in Houston, Texas [21] analyzed the influence of a WSB program on student's physical activity. Findings show that rates of active transportation to school increased, demonstrating a desired outcome.

3 Model

3.1 Relation to Standard Routing Problems

One of the possible comparisons to the WSBRP is that of the School Bus Routing Problem in which bus routes are designed for a fixed number of buses to pick up all of the students in a network. The problem of building walking routes for all students in a catchment is structurally similar to that of the standard School Bus Routing Problem (SBRP). Each student must be picked up by a leader (bus) and dropped off at the school while considering capacity and distance constraints. In previous research, the goals of the SBRP have been to optimize routing subject to a pre-determined number of buses and a fixed capacity on each vehicle [19]. Our work differs from this problem by using an undetermined number of leaders and treats leader capacity as a soft constraint. The different handling of leaders is required for this problem as leaders would be community volunteers. Participation by this group of volunteers would be subject to variation as the program continued. Further, the treatment of leader capacity as a soft constraint allows some leaders to exceed the suggested eight student limit though a solution doing so is penalized.

The context in which we solve the problem also differs from standard SBRPs as we have to dynamically solve the system when a new node is added. Addition of a new node represents someone signing up on the applet to enroll their child or to volunteer as a leader. This challenge forces us to use different heuristics as many current heuristics rely on a large number of iterations and sizable running times. We approach this problem by incorporating clustering into our heuristics. Doing so breaks the combinatorial problem on N nodes down to j combinatorial problems on n_j nodes each.

Another model that bares comparison to the WSBRP is that of the Vehicle Routing Problem in which a fleet of vehicles are sent from a central depot to service every node in the network while accruing the maximum profit. Modifications exist of this combinatorial problem that include time windows and other constraints [17]. The problem aims to categorize customers by their delivery

vehicle while minimizing the total route cost subject to capacity and time constraints [10]. Since determining the optimal solution to VRPs has been proven to be NP-hard [9] [23], the problem is often evaluated by using heuristic and/or meta-heuristic algorithms [10].

The School Bus and Vehicle Routing Problems are closely related and thus both draw comparisons to the Walking School Bus Routing Problem. In a standard VRP, all edges have an associated distance cost and are utilized as directed arcs [17]. Calculations of these costs generally come from point to point distances rather than Euclidean calculations. Further, all non-depot nodes are treated as equal customers and can be traversed in any order desired. Other modifications of the VRP allow for time windows that condense the order of traversal, though these do not constrain the problem to access a leader node first specifically [23].

The WSB design differs from the VRP, and the SBRP by consequence, as some customers have specific constraints. A leader must be the start of a path and each path must contain exactly one leader. This represents the idea that the children walking must always have a supervisor. Walking routes for leaders, however, can overlap and represent the opportunity to fuse WSBs or to leave a leader node unused. As the VRP is designed to generate cycles on the graph, the WSB can treat one edge as redundant due to it causing a cycle instead of a path. This is achieved by making arcs asymmetrically weighted and adjusting the weight on depot-to-leader arcs to be zero. To prevent non-leader nodes from being the first node of the path, all weights on edges inbound to a leader node are set to a large value M except for the depot-to-leader edges. The differences in these two systems are shown in Figure 1 and Figure 2. Yellow nodes represent the leaders and green nodes represent the students.

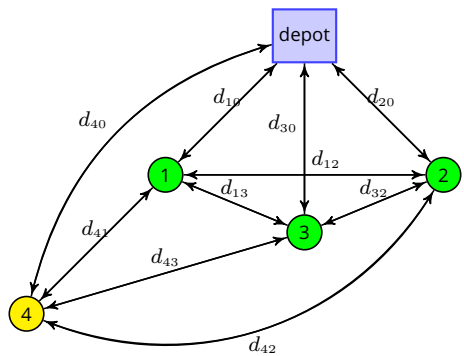


Figure 1: Original Vehicle Routing Problem Example

All edges connecting two non-leader nodes remain unchanged. An example result of running a VRP heuristic on the graph would generate the path in Figure 3. The depot-to-leader edge $e_{0,4}$ is removed to create a path from the VRP-generated cycle.

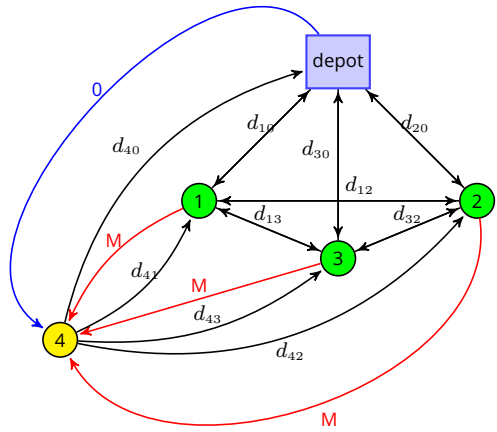


Figure 2: Modified Version of Vehicle Routing Problem Example

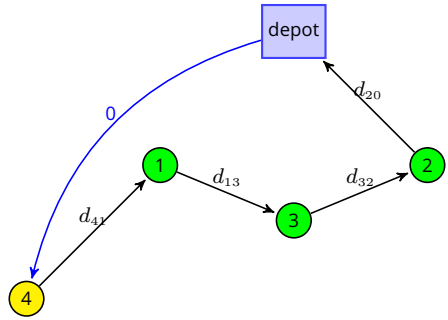


Figure 3: Resulting Cycle from the Vehicle Routing Problem Example

3.2 Assumptions

The general model we use is a modification of School Bus Routing Problem (SBRP) [19]. Relevant assumptions are as follows.

1. All students are picked up from their home
2. The number of leaders is less than the number of students
3. All volunteers are chosen as leaders
4. Leaders can start their routes at independent times
5. Students are a homogeneous group
6. Service time at a node is zero

Assumption 1.) is commonly used in rural SBRP as students are sparse [23]. Urban areas often make use of central bus stops. Though Greater Vancouver represents a predominantly urban area, the concept of a bus is replaced with that of a walking leader and the central bus stops are made unnecessary. Other implementations of the WSB have followed this same design principle [1].

Assumption 2.) is a restriction set on the generation of synthetic data. The reason for this restriction is that some of the clustering methods fail when the number of clusters to create is larger than the number of data points available to cluster. Allowing leader merging prior to routing would allow this assumption to be relaxed.

Assumption 3.) is a simplification of the data generation process. Data could be generated according to a two phase design in which the first phase generates N student nodes and the second phase reclassifies a random number of student nodes into leader nodes. The second phase of the generation would represent the volunteering of leaders. Another approach would be to generate a number of leaders on $[2, N/2]$ and then combines leaders based on a proximity threshold. The remainder of the nodes would then be generated as students. This assumption is in line with the first data generation process.

Assumption 4.) simplifies the incorporation of the hard constraint for school start time. By assuming leaders can start their routes at independent times, each leader is implicitly assumed to start their route at or before the school start time minus their route time.

Assumption 5.) gives each student the same properties. This means that each student has the same walking capacity, time capacity, and walking speed. Future work could explore heterogeneous students based on age differences.

Assumption 6.) reduces the travel time of a route to the time that is spent walking. This results

in the routes generated representing the lower bound on the route's travel time. This is compensated for by building the objective function with respect to distance traveled instead of time.

3.3 Notation

The Walking School Bus Routing Problem (WSBRP) can be mathematically modeled as follows. Each leader $l \in L$ is the leader of a student cluster V_l. The number of students traveling with a leader is then defined to be $|V_l|$ as the leader is assumed to have a student. The route traversed for each cluster starts at the home location of the leader, p_l. The leader then visits each destination p_i, where $i \in V_l$. This results in a path of $p_l - p_{i_1} - ... - p_{i_{|V_l|}} - p_{N+1}$ on each cluster. The travel time for such a path is then the sum of each t_{ij} cost associated with the traversed edges. A similar calculation is done to calculate the distance traveled in the path using d_{ij}.

Let x_{ijk} represent the decision variable for whether edge (p_i, p_j) is traversed by leader k. Then the total time and distance taken for the leader l to collect all of the students and arrive at school can be calculated as

$$\sum_{i,j \in V_l} x_{ijl} * t_{ij}$$

and

$$\sum_{i,j \in V_l} x_{ijl} * d_{ij}$$

respectively. As the objective of the network is to minimize total distance traveled, the effective distance cost of traversing some edge (i, j) is $d_{ij} x_{ijk} c_{ik}$, where c_{ik} represents the number of students with leader k when departing from node i.

- L = Number of leaders available
- D = Maximum walk-able travel distance
- T = Maximum walk-able travel time
- C = Leader threshold capacity
- N = Total number of students to be served
- V_k = Nodes belonging to cluster k
- c_{ik} = Number of students with leader k at node i
- p_i = Pick-up points (where p_{N+1} is school)
- r_i = Total distance traveled by student i
- s_i = Total time traveled by student i

- t_{ij} = Travel time from p_i to p_j
- d_{ij} = Travel distance from p_i to p_j
- $x_{ijk} = \begin{cases} 1, & \text{if leader } k \text{ travels directly from } p_i \text{ to } p_j \text{ where } i \neq j \\ 0, & \text{otherwise} \end{cases}$

3.4 Model Specification and Structure

The objective in the WSBRP is to minimize the cost incurred walking to school. The measure we use to understand cost is the total distance traveled in the network. Distance is measured using Google Maps point to point distances. Although time is also returned from the Google API call, the walking pace used is faster than a child's walking pace. For this reason, distance represents a more accurate objective to consider for the WSBRP.

Calculation of the total distance traveled in the network is done according to the following sum.

$$min \sum_{k=1}^{L} \Big[\sum_{i=1}^{N} \Big(\sum_{j=1}^{N+1} d_{ij} x_{ijk} c_{ik} \Big) \Big] \tag{1}$$

The cost used in the sum is the effective distance cost based on the number of students with the leader at a given time. The soft constraints associated with distance and time management thresholds are used in a scoring function applied after the original route generation. This is due to the limited input available to the Google Maps API as routes are optimized according to an internal objective function.

3.5 Constraints

The constraints associated with the WSB differ from those of the standard SBRP and VRP. Elements unique to walking in a group are not captured by these other formulations. For example, the capacity of a leader is now part of the scoring evaluation rather than a hard constraint as it is possible to have more than the recommended leader to student ratio. Further, the distance a student walks is now used in a scoring function for route evaluation as well as in the objective function. This section will discuss these changes and provide justifications.

Apart from neighbourhood and sidewalk safety, distance is the main aspect either encouraging or discouraging students to walk to school [18]. A growing number of studies suggest that the distance threshold that is practical to walk is approximately one mile [12] [31] [18]. Further, findings suggest that willingness to walk to school is negatively correlated with the distance to be traveled [31]. For these reasons, distance walked per student is measured against a threshold distance of 1 mile (1.6 km) in the scoring function. As time and distance traveled are closely related in an urban environment, the scoring on distance will serve as a scoring on time implicitly.

Finally, consideration must be made for how many students can walk with a single leader. The capacity of a leader is flexible though safety concerns become apparent above certain thresholds. In a pilot study conducted in Ottawa, the ratio of leader to children was set to be 1:10. This ratio was deemed to pose a safety risk so it was later revised to 1:8 [1]. The Canadian Cancer Society's Trottibus, the Walking School Bus in Ontario and New Brunswick, also suggests each leader should have no more than eight students [5]. The capacity threshold is therefore decided to be 1:8.

4 Data

4.1 Vancouver School Catchment

Data regarding the actual locations of children interested in participating in such a program is unavailable to us. Because of this absence of data, the model we designed is based on simulated data. Using the polygonal school boundaries available in online resources [7], school catchments are defined for the schools of Greater Vancouver. Each school can then have students generated within its catchment boundaries for consideration in this model. We recognize that some students attend schools outside of their appropriate catchment region due to pre-existing circumstances and do not include these students in our model. The justification of such is that students beyond the polygonal school boundary are likely beyond a reasonable walking distance to school.

Of the students that are generated, another sampling process selects which nodes are to be leader nodes. Selected nodes indicate a parent willing to volunteer as a bus leader. Descriptions as to how these data points are generated is covered in the simulation discussion.

The construction of our model is based on 58 schools from the Greater Vancouver Region. The original data collected from the online resource contained a total of 73 school boundaries. The removed schools were excluded because of duplicate locations in Google Maps. These duplicates occurred because schools were identified by name automatically in Google Maps, generating slight inconsistencies that do not occur with Longitude and Latitude. As we did not have the Longitude and Latitude for the schools, names were used in place.

As the data for school boundaries is encoded according to European Petroleum Survey Group (EPSG) standards, transformations are required to create Longitudinal and Latitudinal coordinates. All Greater Vancouver schools exist in the EPSG: 26910 region which covers North America from 126 degrees West to 120 Degrees West. Conversion of EPSG: 26910 data to Longitudinal and Latitudinal data requires a set of transformations incorporated in the “spTransform" function of the “sp" package in R [24] [2].

4.2 Population Density

The simulation method is based on accurate population data from the Canadian Census. Accessed through SimplyAnalytics [25], the Census data can be filtered at the geographic block level

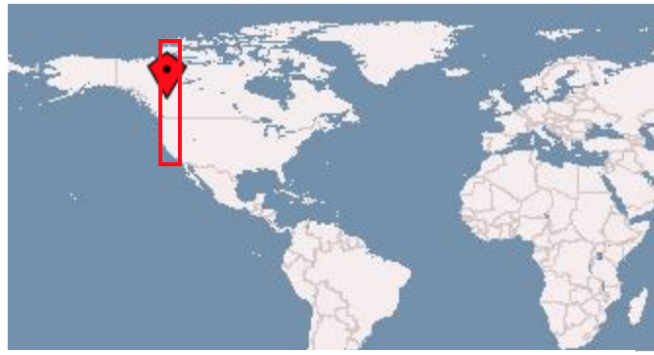


Figure 4: EPSG: 26910 region [3]

and on the specific ages of 5-14. Doing so creates a series of polygons that represent city blocks and contain the number of children aged 5-14 in that block. The incorporation of these population groupings allows for proportional sampling to be used in place of simple random sampling. This creates samples that are more closely representative of the underlying population and allows for more realistic simulations.

4.3 Generation of Synthetic Data

The simulation of this scenario focuses on geographic point generation. Uniform random sampling of points within a school's catchment zone is a simple exercise in R using the polygons we have defined, but is not necessarily reflective of the actual region. For example, uniform random sampling could generate multiple student nodes in a park or other location that is of low population density. Because this is not faithful to our problem, we enhanced our sampling by informing it with Census data. Using Simply Analytics, we collected data at a block by block level for the number of children aged 5 to 14 living in that location [25]. The resulting data was then overlaid with the school catchment region to redefine the density with which to sample from a given location. This gives each block the following probability of being sampled.

$$P(\text{sample a point from this block}) = \frac{\text{Number of children on this block}}{\text{Number of children in this catchment region}}$$

The result of this formula is that more dense locations are more likely to be sampled in each draw. As sampling is done with replacement, the more densely populated areas are, on average, sampled more frequently.

The entire process of generation is that of a Two Stage Design [27]. A region is defined as the union of many subregions and a sample of these subregions is drawn with replacement. For each time a subregion is drawn, a point is generated within that subregion. The resulting collection of points is roughly representative of the underlying population of children in that region.

Leaders are selected from the generated data by a uniform random process. The leaders represent the start of a school bus in the data. Remaining nodes are deemed to be students and are to be assigned to one of the leaders. Each path to the school must contain exactly one leader. These restrictions result in the following formulation of constraints for clustering.

$$\sum_{k=1}^{L} \Big(\sum_{i \neq j}^{N} x_{ijk} \Big) = 1 \text{ for all nodes } j \neq N+1 \tag{2}$$

$$|V_l \cap L| = 1 \text{ for all clusters } l \in L \tag{3}$$

5 Clustering Methods

With the aforementioned constraints in mind, the data is then clustered according to four methods: k-Nearest Neighbours, k-means, Hierarchical, and Soft. An illustration of the four methods are shown in figure 5.

5.1 K-Nearest Neighbours Clustering

The implementation of k-Nearest Neighbours (kNN) involves describing neighbourhoods on a training set. This training set is the set of leaders. By mapping the leaders first and describing the neighbourhood size, kNN creates neighbourhoods on the map such that a new point added to the graph will be classified as being in the neighbourhood of one of the leader nodes. Each leader has a neighbourhood of size corresponding to a user chosen parameter. The neighbourhood size for this problem has been set to one. Reasoning for this is that each entering node needs to be assigned specifically to one leader. If the parameter were set to any higher value, each point would have to have its group membership randomly selected from the overlapping neighbourhoods as the training set only contains one member of each neighbourhood (leaders). Distance measures of closeness are set to Euclidean. This provides a quick clustering approach but does not account for actual geographic distance. Because of the connectivity of Greater Vancouver, however, this does not seem to be a cause of any problems though specific street distances would likely improve results modestly.

5.2 K-means Clustering

k-means clustering is a popular Machine Learning problem that involves classifying points based on the nearest cluster mean. The algorithm is an iterative process in which the first step is to assign initial centres. Because of the NP-Hard nature of k-means, even in two dimensions, heuristics are often used [28] [30]. To prevent the problems of local optima associated with heuristics, random centres are chosen and the process of finding cluster membership is done multiple times.

The heuristically optimal assignment is then the groupings that minimize the total sum of squared distances to the mean of each cluster for nodes belonging to said cluster. If there is a tie for the minimum sum of squares, the assignment is chosen randomly from those of equal sum of squares.

As one of the key components of our routing involves fixing one leader per cluster, the k-means algorithm is run without the leaders involved. Instead, the number of clusters to be generated is set equal to the number of leaders that exist within a given school's generated points. After the decision on clusters by the k-means algorithm laid out by Hartigan and Wong [14], each cluster is assigned to a leader according to the Hungarian Algorithm. This is done by creating an n x n matrix in which entry (i, j) corresponds to the geodesic distance between the mean of cluster i and leader j. As our scoring utilizes distance walked as the cost associated with a route, mapping a leader to a cluster through the Hungarian Algorithm represents a minimization of cost associated with leader cluster matching. Justification of such an approach can be obtained by recognizing this as an Assignment Problem or a Bipartite Matching problem in which a perfect matching is to be obtained [16]. Both of these instances have been shown to be applications of the Hungarian Algorithm and are equivalent to the process of leader cluster assignment described.

5.3 Hierarchical Clustering

Hierarchical clustering differs from previous clustering methods as no structure or number of clusters is predefined. Instead points all start as being classified as their own cluster. This is known as Agglomerative Hierarchical Clustering. Iteratively, the two closest clusters are merged into one cluster. Closeness, in the context of our hierarchical implementation, is geodesic distance. The process completes when all clusters have been merged into one. This is represented by a tree in which clusters are bound to each other at a given height in the tree, representing the proximity of the joined clusters.

The use of hierarchical clustering in our work involved the clustering of non-leader nodes only. As the result of hierarchical clustering is one ultimate cluster, an earlier version of the cluster tree is needed such that the number of clusters created is equal to the number of leaders. This is established by running hierarchical clustering to completion and then cutting the created tree at the lowest height in which the appropriate number of groups are generated. For each of these leaderless groups, an average point is then calculated with coordinates as follows:

$$(x_i, y_i) = \Big(\sum_{j \in \text{clust i}} \frac{x_j}{n_{clust}}, \sum_{j \in \text{clust i}} \frac{y_j}{n_{clust}} \Big)$$

Using the created middle points for each cluster, a distance matrix from each cluster centre to each leader is created according to geodesic distance. The Hungarian Algorithm is then similarly applied to this distance matrix, pairing leaders with clusters.

5.4 Soft Clustering

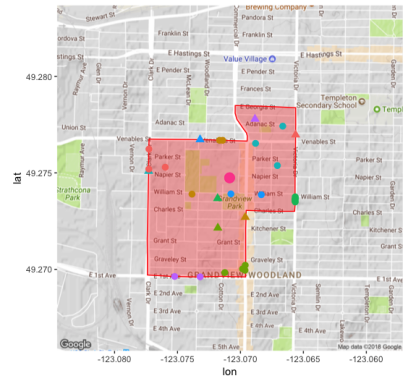
Soft Clustering is an approach based on the idea that a data point can belong to multiple clusters simultaneously. This is represented by having each data point assigned a degree of membership coefficient which expresses the degree of certainty with which that data point is believed to belong to a given cluster. A value is assigned to each data point for each cluster.

Our approach requires students to belong to exactly one cluster and one leader, indicating that the base implementation of Soft Clustering is insufficient for our work. Instead, we make a modification in which a data point is grouped in the cluster with which it has the highest degree of membership coefficient, with ties being broken randomly. This forces each node into exactly one cluster. As leaders are incorporated in the clustering and represent the original centers there is no need to map leaders to nodes through the Hungarian Algorithm. Instead clusters already contain both nodes and a leader.

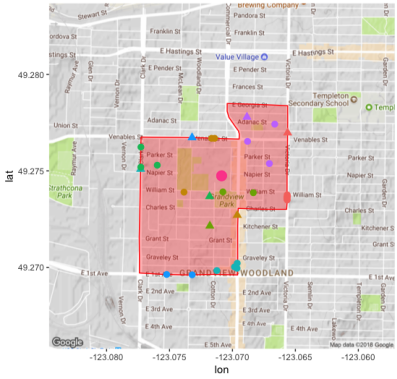
5.5 VRP Solver

To compare the results of the cluster first route second heuristics, we used an open source VRP solver [10]. The solver uses Bing maps to calculate the distances between each pair of nodes. As mentioned earlier, the VRP comparison will be solved on the modified graph in section 3.1. This will allow the VRP solver to evaluate the same problem as the WSBRP. This specific implementation of the solver uses the Adaptive Large Neighbourhood Search heuristic to evaluate solutions [10]. The solver is given 60 seconds of run time and calculates approximately 800 solutions on our 16 GB RAM computer with an Intel Core i7 processor. Modifications to the solver, such as vehicle capacity constraints and daily work limits, cause the algorithm to only generate 60-100 solutions in the given time on the same computer. We provided an increased run time allocation for these constrained cases to generate more solutions. In testing, we provided the heuristic 60 seconds and 600 seconds of run time on the same test region and obtained the same routing. As we did not observe any gain, the remainder of testing was carried out using 60 seconds of run time.

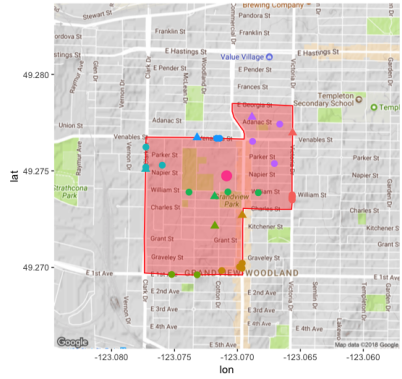
(a) kNN Clustering



(b) $k$means Clustering



(c) Hierarchical Clustering



(d) Soft Clustering

Figure 5: Illustration of the four clustering methods for Britannia Community Elementary located in Greater Vancouver. The red shaded area represents the school catchment, the big pink dot denotes the school, the circles are student nodes and the triangles are leader nodes.

6 Analysis

The result of each clustering method for the 58 regions is evaluated first for the total distance traversed in its solution and per route. The route comparisons are based on the threshold values for distance, D, and time, T. A route that has no student exceed the threshold time or distance is a successful route. A clustering method that has no routes for a given region surpass the threshold distance or time is a successful method on that region. Because of geographic region differences, some regions generate multiple successful methods while others generate none. In the case of either multiple or zero successful methods, a secondary evaluation method is applied in the form of the scoring function. The scoring function assigns a score that expresses the degree of violation in each of the constraints as well as the balance in the network. The formulation of this is as follows.

$$DistScore = \sum_{i=1}^{N} max(0, r_i - D)$$

$$TimeScore = \sum_{i=1}^{N} max(0, s_i - T)$$

$$Score = DistScore + TimeScore$$

Following this scoring process, the methods are first ranked based on their corresponding success in generating regions without threshold violations. The result of this ranking shows that among the four clustering methods, soft clustering handles the threshold requirements better than the other methods . Despite regional differences, soft clustering satisfies the threshold requirements in more than one third of the regions. k-NN clustering is the next best with respect to threshold requirements while k-means clustering and hierarchical clustering satisfy only 7% and 10% of the regions respectively.

Cluster	$Satisfied$	$Violated$	$\%\ Satisfied$
k-NN	16	42	27.5%
k-means	4	54	6.89%
Hierarchical	6	52	10.3%
Soft	21	37	36.2%

Table 1: Clustering results: time and distance threshold compliance for the four clustering methods over the 58 regions

The effectiveness of each clustering method can be further understood by including other metrics. For regions that have all methods failing to meet threshold values, some methods still perform closer to the desired threshold values and represent a result worth capturing. Let e_{ik} be the extra distance associated with node i when clustered by clustering method k in a given region. Let

a_k then be used to denote the average distance violation in the nodes that violate the distance threshold as follows:

$$e_{ik} = max(0, r_i - D) \tag{4}$$

$$a_k = \frac{\sum_{i=1}^{N} e_{ik}}{|e_{ik} > 0|} \tag{5}$$

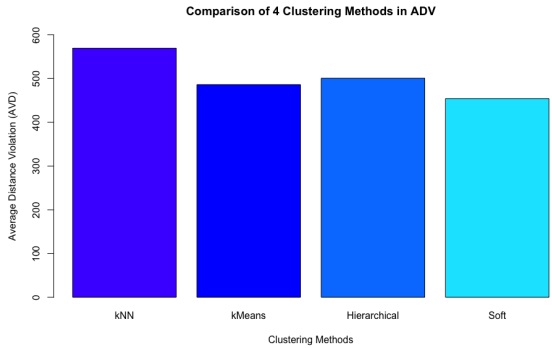


Figure 6: Comparison of Distance Violation of each Clustering Method on Average

The results of averaging the distance violation across clustering methods and regions (figure 6) indicates that soft clustering slightly outperforms on average for violated cases. By comparison k-NN clustering, a method that performed nearly as well as soft clustering in terms of satisfaction, generates the worst average distance violation. This chart only compares average distance violations calculated against the number of times distance violations occur. When compared with figure 7, a more complete understanding of method behaviour over the 58 regions can be established.

Figure 7 indicates that apart from two regions in which all of the clustering methods passed the threshold values on all routes, k-means clustering routinely displayed the largest average distance violation on a region. Soft clustering demonstrated the fewest maximum average distance violations.

Another important threshold to validate on is the leader capacity. We described each leader homogeneously, giving each a capacity of $C = 8$ for students. If all routes generated by a clustering method for a given region satisfy the capacity threshold, regardless of distance and time thresholds, it is counted as a success with respect to capacity. This information is captured in the contin-

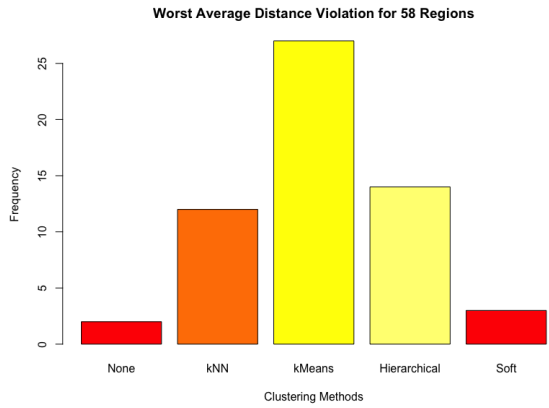


Figure 7: Comparison of Average Distance Violation across the 58 Regions

gency tables presented in Table 2.

(a) k-NN Contingency table

Capacity / Distance	Satisfied	Violated	$Total$
Satisfied	14	2	16
Violated	23	19	42
Total	37	21	58

(b) k-means Contingency table

Capacity / Distance	Satisfied	Violated	$Total$
Satisfied	3	1	4
Violated	44	10	54
Total	47	11	58

(c) Hierarchical Contingency table

Capacity / Distance	Satisfied	Violated	$Total$
Satisfied	5	1	6
Violated	40	12	52
Total	45	13	58

(d) Soft Contingency table

Capacity / Distance	Satisfied	Violated	$Total$
Satisfied	18	3	21
Violated	31	6	37
Total	49	9	58

Table 2: Comparison of clustering distance violations against capacity violations

The validation of the cluster first heuristics is then provided by comparison with the VRP solver. Use of the solver is through a spreadsheet interface and represents a manual process. Due to external time constraints, we could not validate the results on all 58 regions. Instead we used a random sample to select 20 regions and applied the VRP solver. Each of the four clustering methods and VRP solver used the same leaders and students for these tests. Because the VRP solver uses the Bing API to generate route distances and times, we generated the routes with the

VRP solver but collected the distance and time measures from the Google API. This maintains the consistency within the comparison. The results of these comparisons are then captured in Table 3.

Cluster	$Satisfied$	$Violated$	$Total$	$\%\ Satisfied$
K-NN	5	15	20	25%
K-Means	1	19	20	5%
Hierarchal	2	18	20	10%
Soft	7	13	20	35%
VRP	2	18	20	10%

Table 3: Comparison of Clusters and VRP

Capacity / Distance	Satisfied	Violated	$Total$
Satisfied	2	0	2
Violated	17	1	18
Total	19	1	20

Table 4: VRP Contingency table: distance violations against capacity violations

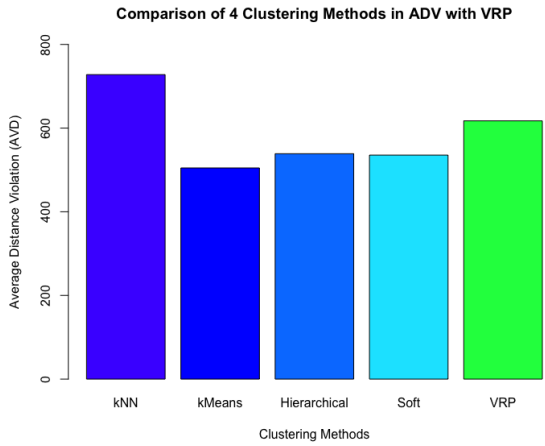


Figure 8: Average Distance Violation of the four clustering methods and VRP for selected regions

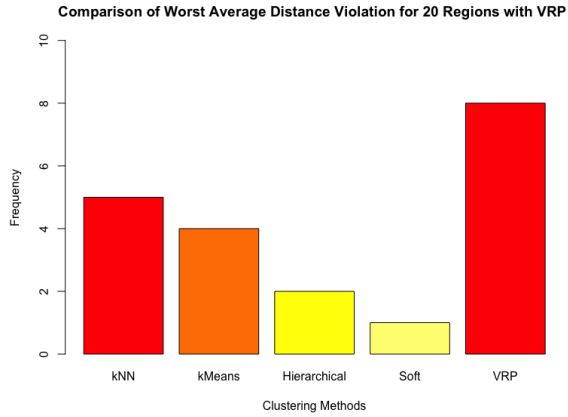


Figure 9: Worst Average Distance Violation comparison with VRP on selected regions

7 Conclusion

In this paper, we proposed four cluster first route second heuristic algorithms for evaluating the WSBRP on 58 Greater Vancouver school catchments. Each method clusters the points in a different manner and the results express this. In the context of our study, the most consistent clustering method is soft clustering. Over the 58 regions, soft clustering satisfied the time and distance thresholds 21 times, representing a 36.2% success rate. This represents a more consistent performance than the other methods which only satisfied in 27.5%, 10.3%, and 6.9% of regions. Further, in the regions that soft clustering violated the distance and time thresholds, the average distance violation was the minimum of the four clustering methods. This is reiterated by figure 7 which demonstrates only 3 of the 58 regions had the longest routings produced by soft clustering.

With respect to table 2, the satisfaction of the capacity threshold can also be examined. Each of the methods performed similarly in the case that distance thresholds were met. In these cases, kNN clustering satisfied 14/16 capacity constraints, $k$means clustering satisfied 3/4 capacity constraints, hierarchical clustering satisfied 5/6 capacity constraints, and soft clustering satisfied 18/21 capacity constraints. These are equivalent to 87.5%, 75%, 83.3%, and 85.7% respectively. Considering now the violated distance threshold regions, the clustering methods score 23/42, 44/54, 40/52, and 31/37. These correspond to 54.7%, 81.5%, 76.9%, and 83.8% respectively. Though $k$means, hierarchical, and soft clustering all score above 75% in terms of capacity satisfaction when the distance threshold is violated, kNN fails to do so and only generates 54.7%. This indicates that despite kNN performing well in terms of distance satisfaction, the regions in which it fails to satisfy generally fail against more thresholds as well. This is not common across the other clustering methods.

To bring these tests in to scope, the results are compared with that of the open source VRP solver [10]. The solver uses the Adaptive Large Neighbourhood Search heuristic in its calculations and is solved on the modified network outlined in figure 2. This solver has been successfully used previously in the health care and tourism industries, as per specific case studies [10].

In the implementation of the VRP solver for this project, early exploration found the solver to be generating routes far beyond the capacity of a leader. To account for this, we used the route capacity constraint in the solver to limit each route to eight students unless there was an insufficient amount of leaders to visit all students while only traveling with at most eight students.

The results of the VRP solver against the clustering methods can be seen in figures 8 and 9 as well as in tables 3 and 4. On the 20 selected regions, VRP satisfied the distance and time thresholds in 2 of the regions. All other clustering methods performed near the same levels as they did for the 58 regions, meaning that the VRP performed as well as hierarchical and $k$means clustering for distance threshold satisfaction. In terms of the average distance violation incurred on regions that did not have the distance threshold satisfied, VRP generated the second highest value behind only that of kNN clustering. Further, VRP most frequently generated the longest routes in the sampled regions, doing so in 8 of the 20.

The performance of the VRP solver on this problem suggests that our cluster first route second algorithms are of use, specifically those involving soft clustering. Reasoning as to how the clustering algorithms outperformed the VRP solver can be found in the use of the Google API. Each call to the API allows a user to specify whether to optimize a path through the given waypoints (students) based on time taken. Our use of this feature returned near-optimal paths for the clusters we generated. This means that the routing is most improved by improving the clustering process. As the VRP solver clusters points according to a different algorithm and API (Bing), the solver may be disadvantaged. Rerunning this solver with different clustering methods and the Google API may result in more closely comparable results.

7.1 Further Studies and Improvements

The purpose of this study was to create a heuristic algorithm that could be used effectively in the design of a WSB applet. At this point, a starting body of code and an efficient heuristic exist though the applet does not. The code can be found at https://github.com/mreyers/MATH-402-Projects. To build the applet, one could use R Shiny for design. R Shiny allows for interactive web pages with underlying R code. This would allow simple integration of the code in Github with the applet. The basic functionality that the applet should support includes node insertion, node deletion, school insertion, and school deletion. Each of the node insertion and deletion functionalities are manageable by minor modifications to the code base as the simulation was built with dynamic leaders and designed to have quick run times. School insertion and deletion would be the ability of a school

to join or leave the program at a time of their choice. This would involve deletion of the school and the related nodes or the creation of a school and the option for an enlisting student to choose that school as their catchment region. Other functionalities, such as the incorporation of pick-up and drop-off time windows, can also be designed though the code in Github does not currently support these.

In the context of the modeling described in this paper, future work should investigate the joining of leaders and potential routes prior to route assignment. Doing so would allow more flexibility with the capacity thresholds and would be more in line with the idea of a Walking School Bus. This could be done by considering leader proximity prior to routing as well as the quantity of leaders in proportion to the quantity of students.

Finally, as this study worked solely with synthetic data, our work could be improved by collecting real world data. Regions such as Ontario and Greater Vancouver have either attempted to implement Walking School Buses or safe walking routes and represent possible areas of collection. Ontario specifically has done data collection on the program and may be interested in larger scale data collection. In gathering real data associated with WSB implementation, the work in this paper can be refined, verified, or disputed.

References

[1] Wallace Beaton. Case Study The Ottawa Walking School Bus Pilot Project. http://www.saferoutestoschool.ca/wp-content/uploads/2017/10/Ottawa-WSB-Pilot-Case-study-May-2015.pdf, 2015. Accessed March 02, 2018.

[2] Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. Applied Spatial Data Analysis with R, Second Edition. Springer, NY, 2013.

[3] Howard Butler, Christopher Schmidt, Dane Springmeyer, and Josh Livni. NAD83 / UTM zone 10N: EPSG Projection - Spatial Reference. http://spatialreference.org/ref/epsg/nad83-utm-zone-10n/, 2007. Accessed March 15, 2018.

[4] Canadian Cancer Society. The Canadian Cancer Society's Trottibus Walking School Bus. http://www.cancer.ca/en/about-us/for-media/media-releases/quebec/2016/semaine-trottibus/?region=qc, 2016. Accessed March 15, 2018.

[5] Canadian Cancer Society. https://www.trottibus.ca/en/, 2018. Accessed March 01, 2018.

[6] Canadian Society for Exercise Physiology. Canadian Physical Activity Guidelines. http://www.csep.ca/guidelines/, 2011. Accessed February 15, 2018.

[7] City of Vancouver. Open Data Catalogue. http://data.vancouver.ca/datacatalogue/publicPlaces.htm, 2015. Accessed January 26, 2018.

[8] City of Vancouver. Walk+Bike+Roll: School Active Travel Planning. http://vancouver.ca/streets-transportation/school-active-travel-planning.aspx, 2018. Accessed February 20, 2018.

[9] Jacques Desrosiers, Yvan Dumas, Marius M. Solomon, and François Soumis. Time Constrained Routing and Scheduling. In Network Routing, volume 8 of Handbooks in Operations Research and Management Science, pages 35 – 139. Elsevier, 1995.

[10] Güneş Erdoğan. An Open Source Spreadsheet Solver for Vehicle Routing Problems. Computers & Operations Research, 84:62–72, August 2017.

[11] Government of Canada. Tackling Obesity in Canada: Childhood Obesity and Excess Weight Rates in Canada. https://www.canada.ca/en/public-health/services/publications/healthy-living/obesity-excess-weight-rates-canadian-children.html, 2018. Accessed March 01, 2018.

[12] Green Communities Canada. School Travel Planning in the City of Toronto and Wellington-Dufferin-Guelph. August 2017.

[13] Green Communities Canada. Active and Safe Routes to School. http://www.saferoutestoschool.ca/, 2018. Accessed March 01, 2018.

[14] John A Hartigan and Manchek A Wong. A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1):100–108, 1979.

[15] Tess Kalinowski and Krinstin Rushowy. Study Looks at Declining Trend of Kids Walking to School. https://www.thestar.com/news/gta/transportation/2016/04/05/fewer-children-walking-to-school-metrolinx-report-says.html, 2018. Accessed March 18, 2018.

[16] Harold W Kuhn. The Hungarian Method for the Assignment Problem. Naval Research Logistics (NRL), 2(1-2):83–97, 1955.

[17] Gilbert Laporte. Fifty Years of Vehicle Routing. Transportation Science, 43(4):408–416, 2009.

[18] Chanam Lee, Xuemei Zhu, Jeongjae Yoon, and James W. Varni. Beyond Distance: Children's School Travel Mode Choice. Annals of Behavioral Medicine, 2013.

[19] LYO Li and Z Fu. The School Bus Routing Problem: a Case Study. Journal of the Operational Research Society, 53(5):552–558, 2002.

[20] Jason A Mendoza, David D Levinger, and Brian D Johnston. Pilot Evaluation of a Walking School Bus Program in a Low-Income, Urban Community. BMC Public Health, 9(1):122, 2009.

[21] Jason A Mendoza, Kathy Watson, Tom Baranowski, Theresa A Nicklas, Doris K Uscanga, and Marcus J Hanfling. The Walking School Bus and Children's Physical Activity: a Pilot Cluster Randomized Controlled Trial. Pediatrics, pages peds–2010, 2011.

[22] National Center for Safe Routes to School . Starting a Walking School Bus. http://www.walkingschoolbus.org/, 2018. Accessed January 12, 2018.

[23] Junhyuk Park and Byung-In Kim. The School Bus Routing Problem: A Review. European Journal of operational research, 202(2):311–319, 2010.

[24] Edzer J. Pebesma and Roger S. Bivand. Classes And Methods For Spatial Data in R. R News, 5(2):9–13, November 2005.

[25] SimplyAnalytics. Census 2017 Total 5 to 9 Population. Total Population by Age. http://app.simplyanalytics.com.proxy.lib.sfu.ca/index.html, 2018. Accessed February 10, 2018.

[26] Sam R Thangiah, Adel Fergany, Bryan Wilson, Anthony Pitluga, and William Mennell. School Bus Routing in Rural School Districts. In Computer-aided Systems in Public Transport, pages 209–232. Springer, 2008.

[27] Steve K. Thompson. Sampling, pages 183–198. John Wiley & Sons, Inc., 3 edition, 2012.

[28] Andrea Vattani. The Hardness of K-means Clustering in the Plane. http://cseweb.ucsd.edu/~avattani/papers/kmeans.pdf, 2009.

[29] VicHealth. Record Number of Children and Schools Get Walking. https://www.vichealth.vic.gov.au/search/record-breaking-participation-in-vichealths-walk-to-school-campaign, 2015. Accessed March 15, 2018.

[30] Haizhou Wang and Mingzhou Song. Optimal k-means Clustering in One Dimension by Dynamic Programming. The R journal, 3(2):29, 2011.

[31] Huaguo Zhou, P E Assistant, Shaoqiang I E Chen, and Peter Hsu. Analyze Factors Affecting Students' Travel Modes Using Multi-Perspectives Diagnosis Approach. In Traffic and Transportation Studies 2010, pages 545–556. American Society of Civil Engineers, 2010.