UNIVERSITY OF CALIFORNIA
Santa Barbara

Addressing Facility Workload Balancing in Coverage Problems

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Geography

by

Jing Xu

Committee in Charge:

Professor Alan T. Murray, Chair

Professor Richard L. Church

Professor Ran Wei

December 2021

The dissertation of
Jing Xu is approved:

Professor Richard L. Church

Professor Ran Wei

Professor Alan T. Murray, Committee Chair

December 2021

Addressing Facility Workload Balancing in Coverage Problems

by

Jing Xu

To my parents Hua and Xuezhong

Acknowledgements

Throughout my PhD study, I have received a great deal of support and inspiration. I would like to express my deepest gratitude to the followings:

- My advisor, Prof. Alan T. Murray, for his detailed guidance, valuable feedback and continuous patience.
- My committee members, Prof. Richard L. Church and Prof. Ran Wei, for their knowledge sharing and wise advice on my research and writing.
- The Department of Geography at UCSB and the National Science Foundation project *ICER–1664173* for their generous financial support.
- My friends for celebrating each achievement together and supporting me through difficult times.
- My parents, Hua Wang and Xuezhong Xu, for their unconditional love and encouragement despite we are physically the half a world away.

Without you all, this work would have not been possible.

Curriculum Vitæ

Jing Xu

Education

2021	Doctor of Philosophy in Geography, University of California, Santa Barbara.
2019	Master of Arts in Statistics, University of California, Santa Barbara.
2018	Master of Arts in Geography, University of California, Santa Barbara.
2016	Bachelor of Engineering in Remote Sensing and Information Engineering, Wuhan University.

Professional Experience

2016 - 2021	Teaching Assistant, Department of Geography, University of California, Santa Barbara.
2021 Summer	Data Scientist Intern, Facebook, Inc.
2020 Summer	Research Data Scientist Intern, Facebook, Inc.
2018 - 2020	Graduate Student Researcher, Department of Geography, University of California, Santa Barbara.
2019 Summer	Data Scientist Intern, Express Scripts, Inc.

Publications

- Murray, A.T., Church, R.L., **Xu, J.**, Carvalho, L., Jones, C., & Roberts, D. (2021). Fire and flood vulnerability, and implications for evacuation. In *Geospatial Technology and Smart Cities*, edited by Poonam Sharma, chapter 17 (Springer).
- Murray, A. T.,**Xu, J.**, Baik, J., Burtner, S., Cho, S., Noi, E., Pludow, A., & Zhou, E. (2020). Overview of Contributions in Geographical Analysis: Waldo Tobler. *Geographical Analysis*, 52(4), 480-493.
- Murray, A. T., Carvalho, L., Church, R. L., Jones, C., Roberts, D., **Xu, J.**, Zigner, K., & Nash, D. (2020). Coastal vulnerability under extreme weather. *Applied Spatial Analysis and Policy*, 1-27.
- **Xu, J.**, Murray, A., Wang, Z., & Church, R. (2020). Challenges in applying capacitated covering models. *Transactions in GIS*, 24(2), 268-290.
- Murray, A. T., **Xu, J.**, Wang, Z., & Church, R. L. (2019). Commercial GIS location analytics: capabilities and performance. *International Journal of Geographical Information Science*, 33(5), 1106-1130.
- **Xu, J.**, & Murray, A. T. (2019). Spatial variability in retail gasoline markets. *Asia-Pacific Journal of Regional Science*, 3(2), 581-603.

Publications in progress

- **Xu, J.**, Murray, A. T., Church, R. L., & Wei, R. Service allocation equity in location coverage analytics. Submitted to *European Journal of Operations Research* (09/2021).

- **Xu, J.**, Murray, A. T., Church, R. L., & Wei, R. A heuristic algorithm for maximal covering considering workload balancing. Revised for *Computers, Environment and Urban Systems* (10/2021).

Awards & Honors

2020	Dissertation Fellowship, Graduation Division, University of California, Santa Barbara, USA.
2017-2019	The Jack & Laura Dangermond Student Travel Scholarship, University of California, Santa Barbara, USA (Spring 2017, Winter 2018, Spring 2018, Fall 2018, Spring 2019).
2016	Outstanding Undergraduate, Wuhan University, China.
2015	The Lei Jun Scholarship (proportion: 0.16%), Wuhan University, China.
2014	Annual National Scholarship (proportion: 0.2%), Ministry of Education of China.

Abstract

Addressing Facility Workload Balancing in Coverage Problems

by

Jing Xu

Coverage problems have been important location models and have been widely applied in practice. A major limitation of simple coverage problems is that they do not control allocation, which might lead to unreasonable facility workloads and workload imbalance. Previous studies have been dealt with facility workload related issues in coverage problems, with one of the most popular approach is to impose capacities and/or thresholds. However, capacities and thresholds cannot guarantee facility workload balance and have associated issues in application. This dissertation seeks to evaluate existing approaches that consider workload balance in coverage problems and study alternative approaches to better address facility workload balance. The primary contribution of this research includes: better understanding and systematic evaluation of existing capacitated coverage approaches including their solution characteristics and commercial GIS performance, new modeling approaches explicitly considering facility workload balance in coverage problems that might be applied to other types of location problems, and efficient solution techniques for proposed multi-objective spatial optimization models.

Contents

Acknowledgements **v**

Curriculum Vitae **vi**

Abstract **ix**

List of Figures **xii**

List of Tables **xv**

1 Introduction **1**
1.1 Motivation 1
1.2 Related Research and Context 6
1.2.1 Key Problems 8
1.2.2 Solution Method 13
1.3 Research Objectives 15
1.4 Significance 15
1.5 Structure of Research 16

2 Challenges in Applying Capacitated Covering Problems **18**
2.1 Introduction 18
2.2 Background 22
2.3 Methods 27
2.4 Application Results 34
2.5 Discussion 42
2.6 Conclusion 57

3 Service Allocation Equity in Maximal Covering **59**
3.1 Introduction 59
3.2 Background 63

3.3 Methods . . . 67
3.3.1 MCLP . . . 68
3.3.2 Workload Variation Measure . . . 70
3.3.3 MCLP Extensions to Balance Workloads . . . 75
3.3.4 Forcing Assignment Constraints . . . 85
3.4 Assessment . . . 86
3.5 Application Results . . . 89
3.5.1 San Jose Study . . . 90
3.5.2 Santa Barbara Study . . . 100
3.6 Discussion . . . 103
3.6.1 Withholding Service . . . 103
3.6.2 Capacitated Comparison . . . 105
3.7 Conclusion . . . 108

4 A Heuristic Algorithm for Balancing Workloads in Coverage Modeling 111
4.1 Introduction . . . 111
4.2 Background . . . 113
4.3 Addressing Workload Balancing . . . 118
4.4 Heuristic Algorithm . . . 124
4.4.1 Initial Hybrid Siting . . . 127
4.4.2 Simulated Annealing Based Demand Allocation . . . 129
4.5 Application Results . . . 134
4.5.1 San Jose Postal Service . . . 135
4.5.2 Santa Barbara Fire Response . . . 140
4.5.3 Boston Fire Response . . . 142
4.5.4 Santa Barbara Nutrition . . . 144
4.6 Discussion . . . 146
4.6.1 Hybrid vs Random Siting . . . 148
4.6.2 Strength of Simulated Annealing Based Allocation . . . 150
4.6.3 Sensitivity to Coverage Standard S . . . 155
4.7 Conclusion . . . 156

5 Conclusions 159
5.1 Summary . . . 159
5.2 Theoretical Contributions . . . 162
5.2.1 Better Understanding of Capacitated Methods . . . 163
5.2.2 Explicit Modeling Approaches for Addressing Workload Balancing 166
5.2.3 Efficient Heuristic Solution Method . . . 168
5.3 Future Work . . . 169

References 173

List of Figures

Figure 1.1 Workload balancing research 7

Figure 2.1 Summary of CMCLP solution details for the San Jose case study 37

Figure 2.2 CMCLP-derived facility configuration ($p = 7$) and withholding service in the San Jose case study 39

Figure 2.3 Summary of CMCLP solution details for the Santa Barbara case study 41

Figure 2.4 CMCLP-derived facility configuration ($p = 3$) and facility workloads in the Santa Barbara case study 43

Figure 2.5 MCLP-derived facility configuration ($p = 3$) and facility workloads in the Santa Barbara case study 44

Figure 2.6 CMCLP-derived facility configuration ($p = 2$) and withholding service in the Santa Barbara case study 45

Figure 3.1 Workload imbalance for an MCLP solution 61

Figure 3.2 Pareto optimal solutions for WBMCLP-TotPairDiff (San Jose, $p = 4$) 92

Figure 3.3 Two Pareto optimal solution configurations for WBMCLP-TotPairDiff (San Jose, $p = 4$) 93

Figure 3.4 Pareto optimal solutions of non-benchmark workload balancing models (San Jose, $p = 4$) 94

Figure 3.5 Pareto optimal solutions of approximate workload balancing models mapped to the objective space of WBMCLP-TotPairDiff (San Jose, $p = 4$) 95

Figure 3.6 Pareto optimal solutions for WBMCLP-TotPairDiff (Santa Barbara, $p = 4$) 103

Figure 3.7 Pareto optimal solutions for WBMCLP-TotPairDiff with no forcing assignment constraints (Santa Barbara, $p = 4$) 105

Figure 3.8 A Pareto optimal solution configuration for WBMCLP-TotPairDiff (no forcing assignment constraints) 106

Figure 4.1 Heuristic solution algorithm 125

Figure 4.2 Initial hybrid siting configuration construction 130

Figure 4.3 Simulated annealing demand allocation process 133

Figure 4.4 Two potential types of neighbor allocation 134

Figure 4.5 Pareto optimal front comparison (San Jose postal service delivery) 139

Figure 4.6 Workloads associated with an unbalanced location and allocation decision (San Jose postal service delivery, $p = 3$) 139

Figure 4.7 Workloads associated with a balanced location and allocation decision (San Jose postal service delivery, $p = 3$) 140

Figure 4.8 Pareto optimal front comparison (Santa Barbara fire response) 142

Figure 4.9 Non-dominated solutions identified by heuristic algorithm (Boston fire response, $p = 5$) 145

Figure 4.10 Workloads associated with a balanced location and allocation decision (Santa Barbara nutrition, $p = 5$) 147

Figure 4.11 Non-dominated solutions using hybrid facility initialization vs random facility initialization (San Jose postal service delivery, $p = 7$) 149

List of Tables

Table 2.1 Applications of CMCLP relying on commercial GIS software for direct solution 26

Table 2.2 CMCLP results (total demand served) for San Jose case study 36

Table 2.3 Facility workload range for San Jose case study 38

Table 2.4 Non-closet allocation for San Jose case study 40

Table 2.5 CMCLP results (total demand served) for Santa Barbara case study 42

Table 2.6 Capacity variation impacts on CMCLP results (total demand served) for San Jose case study 46

Table 2.7 Capacity variation impacts on unserved demand within the service standard for San Jose case study 51

Table 2.8 Capacity variation impacts on CMCLP results (total demand served) for Santa Barbara case study 53

Table 2.9 Capacity variation impacts on unserved demand within the service standard for San Barbara case study 56

Table 3.1 Different measures to account for facility workload variation 71

Table 3.2 Solution quality and computational time comparison among workload balancing models (San Jose) 97

Table 3.3 Solution quality and computational time comparison among workload balancing models (Santa Barbara) 102

Table 4.1 Computational results for proposed algorithm (San Jose postal service delivery, $S = 3$) 138

Table 4.2 Computational results for proposed algorithm (Santa Barbara fire response, $S = 1.5$) 142

Table 4.3 Computational results for proposed algorithm (Boston fire response, $S = 1.5$) 144

Table 4.4 Computational results for proposed algorithm (Santa Barbara nutrition, $S = 5$) 147

Table 4.5 Computational results comparing initial hybrid siting vs random siting (San Jose, $S = 3$) 150

Table 4.6 Computational results of solving the allocation problem: simulated annealing vs greedy algorithm vs Gurobi (San Jose postal service, $S = 3$) 152

Table 4.7 Computational results of solving the allocation problem: simulated annealing vs greedy algorithm vs Gurobi (Santa Barbara nutrition) 153

Table 4.8 Computational results for proposed algorithm (San Jose, $S = 5$**)** 156

Table 4.9 Computational results for proposed algorithm (San Jose, $S = 7$**)** 157

Chapter 1

Introduction

1.1 Motivation

Facility location siting has been a frequent practice for humans and human activities, with location related decisions constantly being made by individuals, households, private companies, governments, and others. To support and facilitate locational decision-making, there has been extensive research on location analysis and modeling. When facility siting decisions are made, there are generally two types of implications: service provider (i.e., facility) and service recipient (i.e., customer/demand). For example, the service provider may be faced with how many customers a facility has to serve, whereas the service recipient will be impacted by the travel time to access a facility. One topic that has been of considerable interest is equity in facility siting, especially when addressing public sector contexts. Facility workload, defined as the amount of demand a facility serves, is probably one of the most important elements of equity as significant variation in workloads is undesirable. If a facility is overutilized, it may not have the capacity to

suitably serve all allocated demand. Alternatively, if a facility is underutilized, then it may not be economically viable to be operated. Imbalance in workloads can negatively impact system efficiency, reliability and service quality. Therefore, workload balancing should be carefully addressed as an integral component of the siting process.

While workload balancing is generally important in facility location, it is particularly critical in coverage contexts. For many facility location problems, service provided to a demand is usually related to the travel distance/time between the demand and the facility to which the demand is assigned. Often, a service is not regarded as accessible if the travel distance/time exceeds a threshold. For example, a fire incident cannot be appropriately responded if it is more than eight-minute driving distance away from its closest fire station, as the major goal is to save life and property. This maximum service distance/time for a facility to suitably respond to a demand for service is called coverage (Church and Murray, 2018). With the coverage concept, various problems types have been studied, with supporting models developed. Fundamental coverage problems include the location covering set problem (LSCP) and maximal covering location problem (MCLP). With a given service coverage threshold(s), the LSCP seeks complete coverage of all demand with a minimum amount of facilities (Toregas et al., 1971) while the MCLP (Church and ReVelle, 1974) identifies the maximum demand coverage with a fixed number of facilities. These two models have been applied to wide variety of application contexts and extended in many ways (Daskin and Stern, 1981; Chung, 1986; Current and Storbeck, 1988; Murawski and Church, 2009; Sorensen and Church, 2010; Church

and Murray, 2018). However, these coverage problems only consider facility location, with demand allocation a byproduct of service configuration. This makes it difficult to control facility workloads. Consequently, it is not uncommon to observe facilities that have significantly varied workloads when basic coverage models are relied upon. This is problematic, making workload balancing an important factor to be addressed in coverage modeling.

How to balance facility workloads in location modeling? First, an equity measure is needed to characterize workload variation. Marsh and Schilling (1994) reviewed about 20 different ways that equity measures were used to quantify the variation of "effects" of siting on facilities. These measures can be applied to workloads. There are two ways to balance facility workloads: (1) imposing constraints to restrict workloads, and (2) adding additional objective(s) to directly minimize workload variation.

Among approaches that restrict workloads, a prominent way to control facility workloads is to impose facility capacities and/or thresholds. For example, the capacitated maximal covering location problem (CMCLP) was proposed to prevent the facility workload from exceeding a predefined capacity in MCLP (Chung et al., 1983; Church and Somogyi, 1985; Current and Storbeck, 1988; Elkady and Abdelsalam, 2016; Ferrari et al., 2018). Doing so, however, effectively requires a reformulation of the MCLP in order to track facility allocations. Nevertheless, capacities help balance facility workloads to some extent because it avoids significantly overutilized facility. In addition, imposing capacities has been very popular in location modeling, partly due to the easy access to such model

extensions in commercial GIS (Murray et al., 2019; Xu et al., 2020). However, capacities are not designed for workload balancing, so significant imbalance in workloads can still exist. Also, establishing a trade-off between coverage and workload balancing objectives is not particularly accessible at this time. Therefore, a systematic understanding and evaluation of performance capacitated model characteristics, especially from a workload balancing point of view, is really greatly needed.

The second approach seeking to address workload balance explicitly has been considered using multi-objectives in location modeling. For example, the maximum workload is adopted and minimized as a second objective to reach equitable allocations in center and median problems (Berman et al., 2009; Kim and Kim, 2010; Davoodi, 2019). The minimization of workload range of sited facilities was introduced in the context of p-median problems (Weaver and Church, 1981; Daskin and Tucker, 2018). Other equity measures, like the workload deviation from a system-wise average workload (Garfinkel and Nemhauser, 1970; Zhu and McKnew, 1993), deviation from a target amount, and total pairwise workload difference (Church and Murray, 1993), have been structured through the addition of objectives in location problems along with supporting constraints. These approaches aim to balance workload directly and provide a way to explore the trade-off among competing objectives. However, the relative capabilities of such approaches to address equity remains largely unexplored for coverage problems. In addition, a common consequence of the use of multi-objectives is greatly increased computational difficulty along with alternative optima interpretation challenges. Therefore, further research is

important that explores alternative approaches, seeking to better understand modeling implications and solution potential for effectively and efficiently balancing workloads in coverage problems.

In sum, facility workload balance, an aspect of facility equity, is an important research topic in location modeling, especially in the context of coverage problems that intentionally ignore allocation. Much research interest has been devoted to facility workload related issues in coverage problems, with a popular approach being the use of capacities and/or thresholds. However, the effectiveness of capacity and threshold limits on balancing facility workload is not well understood. Further, there are several important limitations and issues associated with the use of capacities and thresholds that need to be further investigated. While there have been studies that explicitly balance facility workloads through the use of multi-objectives in location modeling, such as minimizing the maximum workload and workload range, their effectiveness and computational efficiency to address coverage concerns remains unknown. Therefore, this dissertation seeks to evaluate existing approaches that consider workload balance in coverage problems and study alternative approaches for addressing facility equity issues. This research will contribute to theories and methods of location science in the sense that it investigates alternative modeling approaches to classic covering models as well as develops spatial optimization methods to solve them.

1.2 Related Research and Context

This section summarizes relevant location models where facility workload balancing remains a challenge, and positions this dissertation (Figure 1.1). The focus of this dissertation is to address workload balancing in coverage problems. The MCLP, as one of the most basic coverage problem, is studied. Related is the p-median problem (PMP) (Hakimi, 1964; ReVelle and Swain, 1970), because MCLP could be formulated as an equivalent PMP (Church and ReVelle, 1976). The existing prominent approach to govern facility workloads in location problems is to add constraints imposing upper and/or lower bounds on workloads. This extended the MCLP to the CMCLP, maximum coverage with threshold (McTHRESH) (Balakrishnan and Storbeck, 1991), maximum coverage with balanced assignment (McBAS) (Gerrard, 1995). Similarly, PMP was extended to the capacitated p-median problem (CPMP) (Mulvey and Beck, 1984; Pirkul, 1987). Current and Storbeck (1988) proved that the CMCLP could be formulated as a CPMP. The dissertation will firstly evaluate the existing capacitated method in solution characteristics and from workload balancing perspective. Alternatively, additional objectives can be added to directly minimize workload variation. One example is the p-median problem with workload range minimization (PMP-Range) (Daskin and Tucker, 2018). But the capability of such multi-objective approach in balance workloads in coverage problems remains unexplored. This dissertation will explore new modeling approaches in this area. Finally, all these key problems were proved NP-hard (Garey and Johnson, 1979), which

means it is difficult to be solved optimally and a polynomial time algorithm is not possible unless $P = NP$. Therefore, this dissertation will also study efficient solution methods for the proposed model.

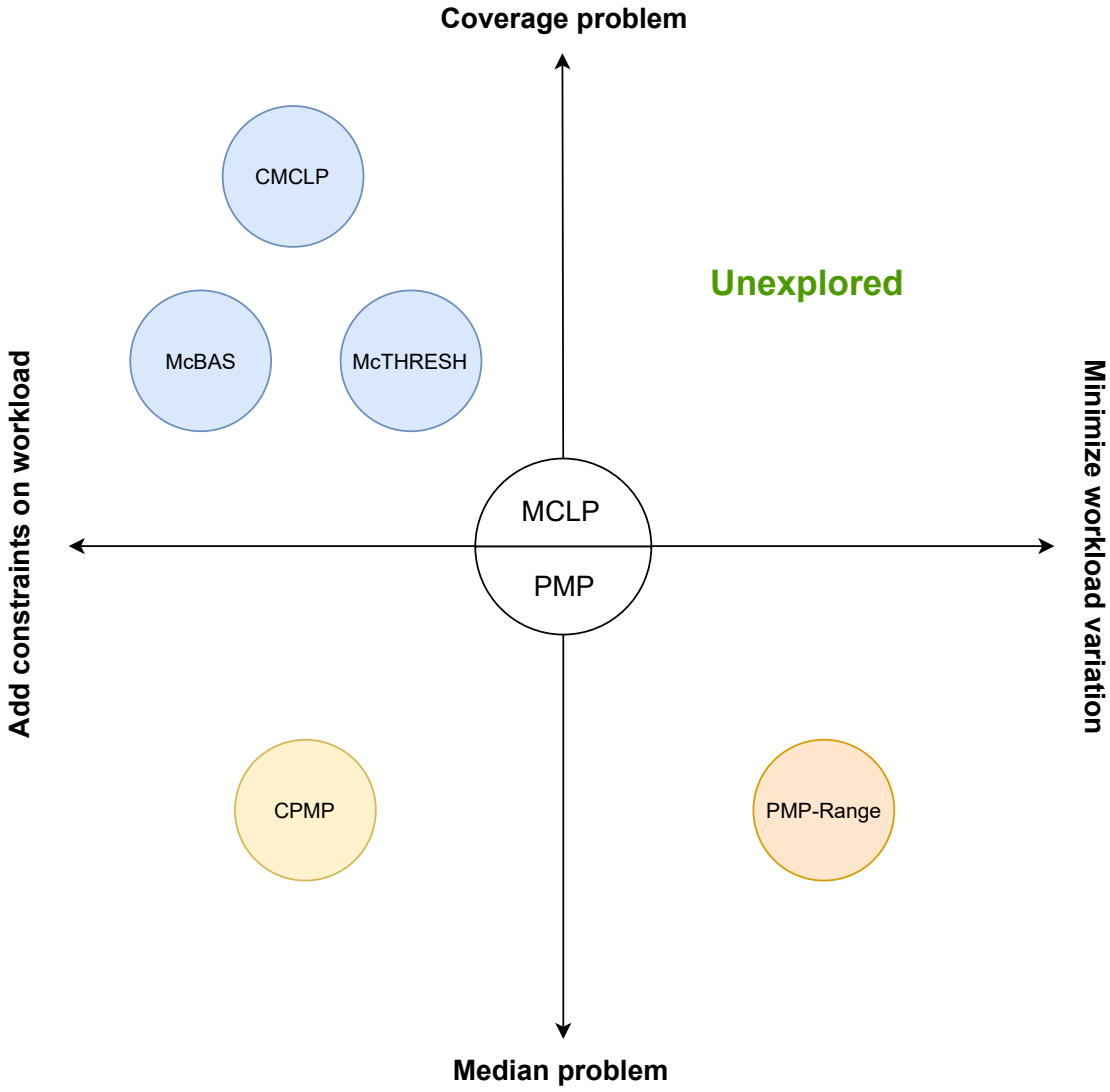

Figure 1.1 Workload balancing research

1.2.1 Key Problems

Consider the following notation:

i = index of demand areas (I entire set)

j = index of potential facilities (J entire set)

d_{ij}= travel distance/cost/time between demand i and facility j

S = service coverage standard

$N_i = \{j|d_{ij} \leq S\}$, the set of facilities that can suitably cover demand i suitably

a_i = amount of demand in area i

p = number of facilities to site

Decision variables:

$$X_j = \begin{cases} 1 & \text{if facility } j \text{ is sited} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{ij} = \begin{cases} 1 & \text{if demand } i \text{ is allocated to facility } j \\ 0 & \text{otherwise} \end{cases}$$

The MCLP is one of the most imporant coverage modeling approaches, seeking to maximize total demand served within the service coverage standard of a fixed number of sited facilities (see Church and ReVelle 1974; Church and Murray 2018). With above notation, the MCLP is formulated as an allocation model as follows:

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{1.1}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{1.2}$$

$$Y_{ij} \leq X_j \quad \forall i \in I, j \in N_i \tag{1.3}$$

$$\Sigma_{j \in J} X_j = p \tag{1.4}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{1.5}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{1.6}$$

The objective (1.1) maximizes the total demand served. Constraints (1.2) require that each demand is allocated to no more than one facility within the service coverage standard. Constraints (1.3) stipulate that a demand cannot be allocated unless a facility is sited. These are often referred to as Balinski constraints, recognized as having integer-friendly solution properties (Church and Roberts, 1983; ReVelle, 1993; Gerrard, 1995; Church and Murray, 2009). Constraint (1.4) requires exactly p facilities to be sited. Constraints (1.5) and (1.6) impose binary restrictions on location and allocation decision variables, respectively. Note that allocation variables are used here and a simpler formulation with no allocation variables can be found in Church and ReVelle (1974). The formulation makes it clear that there is no control on facility workloads. Empirical experience suggests that workloads can vary significantly.

Another fundamental location model is the PMP that seeks to site p facilities so that average (or total weighted) travel time/distance is minimized (Hakimi, 1964; ReVelle and Swain, 1970). The PMP is formulated as follows:

$$\text{Minimize} \quad \Sigma_{i\in I}\Sigma_{j\in J} a_i d_{ij} Y_{ij} \tag{1.7}$$

$$\text{Subject to} \quad \Sigma_{j\in J} Y_{ij} = 1 \quad \forall i \in I \tag{1.8}$$

$$Y_{ij} \leq X_j \quad \forall i \in I, j \in J \tag{1.9}$$

$$\Sigma_{j\in J} X_j = p \tag{1.10}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{1.11}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in J \tag{1.12}$$

The objective (1.7) minimizes the total weighted travel distance. Constraints (1.8) ensure that each demand to be assigned to one facility. Constraints (1.9) are Balinski constraints, limiting the assignment of demand to sited facilities only. Exactly p facilities are sited in Constraint (1.10). Finally, constraints (1.11) and (1.12) are binary integer constraints. Similar to the MCLP, no control on facility workload is imposed here and workloads can vary significantly. In addition, Church and ReVelle (1976) showed that the MCLP can be transformed into an equivalent PMP by manipulating the facility-to-demand distance matrix. Thus, in theory, the MCLP can be solved by any solution techniques developed for the PMP.

One way to balance facility workloads is to impose facility workload capacity and/or thresholds, which are essentially upper and/or lower bounds on allocated demand for service. Denote c_j the capacity of facility j where $c_j > 0$. Capacity constraints are formulated as follows:

$$\Sigma_{i \in I} a_i Y_{ij} \leq c_j X_j \quad \forall j \in J \tag{1.13}$$

The capacity constraints (1.13) require the workload of facility j not to exceed the capacity c_j if the facility is sited. The CMCLP is formulated by adding (1.13) to the MCLP, (1.1)-(1.6) (Chung et al., 1983; Church and Somogyi, 1985; Current and Storbeck, 1988). Similarly, the CPMP is formulated by adding (1.13) to the PMP, (1.7)-(1.12) (Mulvey and Beck, 1984; Pirkul, 1987). Current and Storbeck (1988) showed that the CMCLP can be formulated as a CPMP. This theoretical linkage between the CMCLP and the CPMP enables solution methods developed for the CPMP to be used to solve the CMCLP, though it remains NP-hard so is considered extremely difficult to solve.

Denote t_j the threshold of facility j where $t_j > 0$. Threshold constraints are formulated as follows:

$$\Sigma_{i \in I} a_i Y_{ij} \geq t_j X_j \quad \forall j \in J \tag{1.14}$$

The threshold constraints (1.14) restrict the workload of facility j to be no less than the threshold t_j if it is sited. A model called McTHRESH, was structured and solved in Balakrishnan and Storbeck (1991), involving constrains (1.14) added to the MCLP, (1.1)-(1.6). In addition, there have been studies using both capacities and thresholds

in coverage and median problems. For example, the McBAS model, was formulated by integrating the MCLP, (1.1)-(1.6), and constraints (1.13) and (1.14) by Gerrard (1995) to balance facility workloads in a maximal covering context. Again, the use of capacity and threshold constraints can restrict workloads to be within a specified range $[t_j, c_j]$, but not necessarily balanced unless $[t_j, c_j]$ is a tight range and it is same for all j.

An alternative way to balance facility workloads in location problems is to minimize workload variation through the use of multi-objectives. A recent example is the PMP with a secondary objective that minimizes the workload range (PMP-Range), proposed by Daskin and Tucker (2018). Denote U and L the maximum and minimum workload of sited facilities, respectively, and M a very large positive number. Additional objective and constraints are introduced as follows:

$$\text{Minimize} \quad U - L \tag{1.15}$$

$$\Sigma_i a_i Y_{ij} \leq U \quad \forall j \in J \tag{1.16}$$

$$\Sigma_i a_i Y_{ij} + M(1 - X_j) \geq L \quad \forall j \in J \tag{1.17}$$

$$L \geq 0 \quad \forall j \in J \tag{1.18}$$

(1.15)-(1.18) may be combined with the PMP, (1.7)-(1.12), to introduce the PMP-Range model (Daskin and Tucker, 2018). The objective (1.15) minimizes the difference between the maximum and minimum facility workloads of sited facilities, or rather the workload range. Constraints (1.16) track the maximum workload. Constraints (1.17) track the minimum workload of sited facilities. If facility j is sited, constraint (1.17)

becomes $\Sigma_i a_i Y_{ij} \geq L$. Alternatively, if facility j is not sited, constraint (1.17) becomes $\Sigma_i a_i Y_{ij} + M \geq L$, which does not impose any restrictions since M is a very large positive number. L is non-negative in constraint (1.18). Note that Daskin and Tucker (2018) also introduced other constraints to tighten the model formulation.

While there is other research focused on minimizing other workload variation measures in the context of other location problems, they are not in this introduction. One example is minimizing the total pairwise workload difference in a median like problem (Church and Murray, 1993). Another example is the k-balanced center location problem (k-BCLP) that minimizes the maximum workload in a k-center problem (Davoodi, 2019). Other multi-objective workload balancing problems can be found in Garfinkel and Nemhauser (1970); Weaver and Church (1981); Zhu and McKnew (1993); Kim and Kim (2010); Berman et al. (2009). But little research has focused on how to explicitly balance facility workloads in coverage problems. The research gap here is twofold: lack of systematic evaluation of existing implicit methods from the workload balancing point of view and little exploration of explicit approaches to balance workloads in coverage problems. This is what this dissertation addresses.

1.2.2 Solution Method

All these reviewed problems are NP-hard (Megiddo et al., 1983; Current and Storbeck, 1988; Daskin and Tucker, 2018), suggesting the computational challenges of solving them optimally. As a result, there are many studies focused on developing efficient exact and

heuristic solution techniques. The linear programming with branch-and-bound technique is the most prominent exact solution method that has been applied to solve location problems (e.g., Church and ReVelle 1974; Murray and Tong 2009; Xu et al. 2020). However, more efficient solution methods are needed when the problem instance becomes large and difficult. For example, greedy adding and greedy adding with substitution were developed for solving the MCLP (Church and ReVelle, 1974). Lagrangean relaxation based method is a common one that were used to solve MCLP (Weaver and Church, 1983; Galvão and ReVelle, 1996), CMCLP (Pirkul and Schilling, 1989, 1991; Haghani, 1996), PMP (Narula et al., 1977; Beltran et al., 2006), CPMP (Mulvey and Beck, 1984; Pirkul, 1987) and other problems. Metaheuristics were also used. Tabu search was applied to solving MCLP (Adenso-Diaz and Rodriguez, 1997). Simulated annealing based heuristics were developed for solving MCLP and PMP (Murray and Church, 1996). Genetic algorithm were used to solve the PMP (Bozkaya et al., 2002; Alp et al., 2003), CPMP (Shariff et al., 2013) and PMP-Range (Daskin and Tucker, 2018). Since the MCLP is NP-hard plus the use of mutil-objectives usually increases the computational complexity, it is expected the model that directly minimizes workload variation in coverage problems is computational challenging. So a third focus of this dissertation is to study effective and efficient solution methods for solving the explicit workload balancing model.

1.3 Research Objectives

Three primary research objectives are among a number of goals associated with this dissertation:

1. Examine and investigate important challenges and issues associated with existing approaches that can be used to balance facility workloads in coverage problems, focusing on evaluating solution characteristics of methods accessible in commercial GIS.

2. Formulate and structure alternative modeling approaches that can be used to explicitly balance facility workload in coverage problems.

3. Develop efficient solution methods for proposed models that consider facility workload balance in coverage problems.

1.4 Significance

Coverage problems have been important location models and have been widely applied in practice. A major limitation of simple coverage approaches is that they do not control allocation, which might lead to unreasonable facility workloads, workload imbalance, inequity, service performance degradation, etc. Previous studies have dealt with facility workload related issues in coverage problems, with one of the most popular approaches

being to impose capacities and/or thresholds. However, capacities and thresholds cannot guarantee facility workload balance and have associated issues in application. This dissertation seeks to evaluate existing approaches that consider workload balance in coverage problems and propose alternative approaches to better address facility workload balance. The primary contribution of this research includes: better understanding and systematic evaluation of existing capacitated coverage approaches including their solution characteristics and commercial GIS performance; new modeling approaches explicitly considering facility workload balance in coverage problems that might be applied to other types of location problems; and efficient solution techniques for proposed multi-objective spatial optimization models.

1.5 Structure of Research

This remainder of the dissertation is structured as follows.

Chapter 2 investigates existing modeling approaches for balancing facility workloads in maximal covering, focusing on the most popular method – CMCLP. The CMCLP is assessed in terms of facility workload balance, theoretically and empirically. Rigorous mathematical proof and empirical studies are given to support this. In addition, since the CMCLP is available through commercial GIS software, this chapter compares solutions produced by commercial GIS software with optimal solutions in terms of both quality and efficiency.

Chapter 3 studies five different workload balancing measures and corresponding modeling approaches that can be used to explicitly balance facility workloads in the context of maximal covering. Approaches are evaluated comparatively, with completeness, inferiority and maximum gap measures introduced to support this. Two empirical studies are conducted to evaluate and compare the proposed five workload balancing models, examining their effectiveness and computational efficiency.

Chapter 4 proposes a heuristic algorithm for the proposed workload balanced maximal covering model, due to the observation of great computational difficulties. The proposed algorithm incorporates interchange along with simulated annealing, taking advantage of problem-specific knowledge to derive high-quality solutions in an efficient manner. Four empirical studies are conducted to demonstrate the strength of the algorithm.

Chapter 5 summarizes major research findings and theoretical contribution, as well as directions for future work.

Chapter 2

Challenges in Applying Capacitated Covering Problems

2.1 Introduction

Location covering models have been important spatial analytic approaches, used to support strategic planning, management and decision making in public and private sector contexts. There is a long history in GIS of coverage application and development efforts, including the work of Gerrard et al. (1997), Murray et al. (2008), Straitiff and Cromley (2010), Downs et al. (2014) and Xiao and Murray (2019), as well as Murray and Tong (2007), Tong and Church (2012), Church and Li (2016), Wei and Murray (2016) and Murray et al. (2019). The coverage concept relates to service provision, acknowledging that response criteria like access and accessibility are fundamental. Church and Murray (2018) characterized coverage as a maximum distance or travel time standard for personnel at a facility to respond to a demand for service. Similarly, this may be viewed

This chapter represents a revised version of a paper published in *Transactions in GIS*, co-authored with Dr. Alan T. Murray, Zifan Wang and Dr. Richard L. Church.

from a central place theory perspective where coverage relates to the range of a good or service, reflecting the maximum distance/time a customer is willing travel to consume the good/service (see Christaller 1966). Many examples of coverage standards are detailed in Church and Murray (2018). A classic analysis situation involves fire service, where coverage is associated with a desired response by firefighting personnel within some stipulated time standard, e.g., eight minutes. Given a response standard, a prominent covering model is the maximal covering location problem (MCLP) introduced in Church and ReVelle (1974). It seeks to site a given number of facilities in such a manner that suitable coverage within the standard is provided to the highest total demand possible.

Interestingly, an important underlying assumption in the MCLP is that facilities being sited have unlimited capacities. That is, service by any one facility can be provided to as much demand as possible as long as recipients are within the coverage standard. This assumption, however, is problematic in many ways. Most facilities have limits on service capabilities due to physical, political, structural, regional and other reasons. For example, firefighting personnel at a particular station are limited in the number of calls they can reasonably respond to as there are only so many hours in the day. In addition, the assumption of unlimited service capabilities can lead to significant variation in facility workloads, which is highly problematic and inequitable. For example, Murray and Gerrard (1997) showed how workloads can vary between located facilities when no utilization limits are imposed. In particular, one facility serves 47.8% of total demand while another facility only serves 3.1%, a factor of more than 15 times that of the less utilized facility.

Such a significant deviation in workloads can cause employee dissatisfaction, low morale, poor productivity and other negative effects. Beyond this, variability in workloads may degrade service quality and lead to marginal economic returns. For these reasons and others, unlimited capacity assumptions may be untenable (Current and Storbeck, 1988; Church and Murray, 2018).

The prominent approach for dealing with the assumption of unlimited service capabilities is to constrain total service provided by any one facility. That is, add capacity constraints to the model. Indeed, this has been true for coverage models. The capacitated maximal covering location problem (CMCLP), as an example, was introduced along these lines, employing constraints to track and prevent the workload (total service) for each facility from exceeding an established limit (Chung et al., 1983; Church and Somogyi, 1985; Current and Storbeck, 1988; Liao and Guo, 2008; Elkady and Abdelsalam, 2016; Ferrari et al., 2018). Service availability and facility busyness can be managed through the use of a capacity. Additionally, capacity limits can be employed and structured to help balance facility workloads in a system, shifting demand from heavily burdened facilities to those with available resources. Capacities too are useful for avoiding situations where over-utilized facilities degrade service quality as well as circumventing inefficiency associated with under-utilized facilities.

While using capacity limits on facilities is appealing in many ways, there are challenges and issues in their application. Little attention has been devoted to such considerations. There are at least five challenges associated with the introduction of capacities. The

first is that most approaches for specifying appropriate facility capacities are subjective. Marianov and Serra (1998) suggested two methods for establishing and/or deriving a facility capacity: use total demand multiplied by a scaling factor or use average historical workloads. Both assume expert knowledge and/or the existence of historical information. A second challenge is that a facility's capacity is not necessarily a strict limit, so there may be some degree of flexibility. Accordingly, uncertainty is introduced through the use of strict capacities and this may in turn severely impact service system effectiveness and efficiency. Third, the use of capacities can result in an undesired allocation response where demand is denied service or dispatched to a further away service facility. A fourth challenge is that capacities might fail to reflect actual workloads. If a service provider lacks the authority or control to restrict people from accessing services, facility capacity may be exceeded during operation. A final challenge is that the addition of capacities often significantly increases computational processing in model solution (Current and Storbeck, 1988; Pirkul and Schilling, 1988, 1989; Church and Murray, 2018), sometimes beyond computing capabilities. With these challenges in applying capacitated models, it is clear that further research is necessary.

Interestingly, access to capacitated modeling approaches has greatly expanded. One can devise exact and heuristic approaches for capacitated models through the use of open source packages and libraries in R and Python as well as others. Capacitated models can also be structured and solved in general optimization software such as Xpress and Gurobi. An increasingly popular option is commercial GIS software with user-friendly

access through point-and-click interfaces, making it possible for any user to model and solve various location problems, including capacitated coverage models (Murray et al., 2019). A direct result of this can be observed in Table 2.1 as there has been a significant increase in applications specifically relying on GIS packages to solve capacitated models in various substantive contexts. While access to these analytical methods is encouraging in many ways, a number of additional issues do arise, such as what is the actual model being implemented and solved, how good are obtained solutions, etc. Thus, there is a critical need for a systematic assessment of such capabilities.

This chapter examines important challenges and issues in applying capacitated covering models, focusing on evaluating allocation response and exploring solution characteristics of capacitated coverage models available through commercial software. Background literature associated with the research is provided in the next section. Then, mathematical details of coverage modeling is detailed in the methods section. Two empirical studies, one focusing on postal service in San Jose, CA and another locating nutrition programs in Santa Barbara, CA, are detailed to highlight performance issues and characteristics. Finally, this chapter ends with discussion and conclusions.

2.2 Background

Location models have played an important role in supporting facility siting decision making in numerous areas, such as emergency medical service, fire response, goods delivery,

school districting, species preservation, and manufacturing, just to name a few. Location problems have broad application. Further, they have been categorized according to the following primary types: median, coverage, center, dispersion, hub and competitive (Church and Murray, 2009; Daskin, 2011). One of the most prominent location models is the location set covering problem (LSCP) that deems system service to be adequate if all demand is within a given travel distance/time of a facility. Toregas et al. (1971) proposed the LSCP, seeking complete coverage using a minimal number of facilities. Substantial interest in covering models has followed that seminal work. Given budget and resource realities, the MCLP (maximal covering location problem) was subsequently formalized to support siting a certain number of facilities in order to achieve the most coverage of demand possible (Church and ReVelle, 1974) and applied broadly (Chung, 1986; Gerrard et al., 1997; Downs et al., 2014; Murray, 2016). Many related studies have sought to extend the LSCP and MCLP in various ways, including addressing backup issues (Daskin and Stern, 1981; Hogan and ReVelle, 1985; Bianchi and Church, 1988), probabilistic facility availability (Chapman and White, 1974; Daskin, 1982, 1983; Sorensen and Church, 2010), and continuous space service (Murray and O'Kelly, 2002; Murray et al., 2008; Straitiff and Cromley, 2010; Tong and Church, 2012; Wei and Murray, 2016), to name a few.

One of the more important extensions has been the introduction of facility capacities to deal with the assumption of unlimited service to demand. Associated with coverage modeling, the CMCLP adds capacity limits to the MCLP. The CMCLP was first for-

mulated by Chung et al. (1983) who also developed a substitution based heuristic to solve it. Church and Somogyi (1985) proposed an alternative CMCLP formulation that allows partial assignments and multiple facilities at a site. A max-flow based heuristic called BASC (Balancing Access and Service Coverage) was developed to solve this model. Hogan and ReVelle (1985) used separable programming to solve the CMCLP. Current and Storbeck (1988) formulated the capacitated LSCP and CMCLP, with discussion of the potential for solving a CMCLP using a technique developed for a capacitated p-median and generalized assignment problems. Continued interest and related work includes minimizing total travel distance in addition to maximizing the total served demand (Church, 1974; Pirkul and Schilling, 1991; Haghani, 1996), considering backup service (Pirkul and Schilling, 1989; Narasimhan et al., 1992), imposing multiple facility capacity levels (Yin and Mu, 2012) and developing efficient solution methods (Straitiff and Cromley, 2010). Interestingly, the CMCLP is accessible in contemporary GIS software for general usage. In recent years, broad application of CMCLP can be observed in academic journals, conference proceedings, thesis/dissertations and research reports (see Table 2.1).

The studies summarized in Table 2.1 all utilized commercial GIS software to structure and solve the CMCLP, though their application foci differ significantly. Beyond its broad utility and relevance, a primary reason for the observed uptake in reported applications of the CMCLP is that GIS provides easy access to this approach, serving as an integrated environment for data acquisition, management, manipulation, analysis and display for users. While increased utilization of this location analytic approach is indeed encourag-

ing for improving fundamentally important service systems, the performance behavior of GIS software for solving capacitated models is effectively unknown. At present there are no established solution quality characteristics nor an understanding of computational efficiency in solving the CMCLP by embedded functions in commercial GIS. Critical evaluation is essential because the primary solution approaches for coverage models provided in GIS are heuristics (see Esri 2019; Caliper 2019; Murray et al. 2019). Heuristics are techniques that identify a solution to a model, and are often computationally efficient, but cannot prove or verify anything about the quality of the obtained solution (Church and Murray, 2008). Accordingly, heuristics cannot guarantee an optimal solution will be found. In addition, GIS software, like ArcGIS, uses one heuristic to solve a variety of different problem types through the technique of Hillsman editing (Esri, 2019). Murray et al. (2019) summarized that GRASP (and Teitz and Bart) strategies with a path relinking metaheuristic, originally designed for the p-median problem, are used for solving covering problems in ArcGIS. They found that the provided heuristic did not identify optimal solutions in the majority of evaluated problem instances (LSCP and MCLP). Given that the CMCLP is known to be much harder to solve compared to the MCLP and LSCP due to capacity limits, concern for solution quality is warranted. As a result, solution quality and performance in applying the CMCLP remains unclear in many ways when a commercial GIS is utilized.

In addition to uncertainty in GIS-based heuristic performance, other issues associated with capacitated covering models are also a concern. Pirkul and Schilling (1991)

Table 2.1 Applications of CMCLP relying on commercial GIS software for direct solution

Reference	Application	Publication
Erfani et al. (2019)	Waste storage stations	Journal
Lemire et al. (2019)	Biomass depots	Journal
Alho et al. (2018)	Bays for urban freight vehicles	Journal
Erfani et al. (2018)	Waste storage stations	Journal
Helo et al. (2018)	Distribution centers	Research report
Sharma et al. (2018)	Biorefinery, depot, and storage stations	Journal
Teixeira et al. (2018)	Power plants	Journal
Tiggelaar (2016)	Neonatal intensive care units	Thesis
Manliclic (2016)	Hydrogen fueling stations	Dissertation
Anhorn and Khazai (2015)	Emergency shelters	Journal
Shahid and Mas Machuca (2015)	Optical devices	Thesis
Burciu et al. (2015)	Port activities	Conference proceeding
Sánchez et al. (2015)	Pellet plants	Journal
Naharudin (2014)	School	Thesis

suggested that the service standard is a way to ensure performance characteristics and not a mechanism for withholding service. They argued that it would be unreasonable to withhold service to demand beyond the standard. They extended the basic CMCLP to include an additional objective that minimizes the total weighted distance for demand beyond the standard, where allocation of all demand is made regardless of the standard. Haghani (1996) was similarly focused on allocation of demand within and outside the standard while incorporating both lower and upper capacity limits for each facility. Worth noting as well was the work of Church (1980) who developed a multi-objective model that minimized the weighted distance of assigned demand, maximized the cov-

erage, and minimized the number of facilities which could not reach a desired target of assigned demand (a lower threshold) that was used to design solid waste planning regions for the Tennessee Valley Authority. However, little attention has been given to the allocation response of demand within the service standard, including holding back service to some demand and dispatching some demand to further away facilities when capacities are imposed.

2.3 Methods

Since capacitated models are available in commercial GIS software such as ArcGIS and open-source software including Python and R libraries, easily set up and solved in general optimization packages such as Gurobi, Xpress, LINGO and others, and the CMCLP has been applied in a number of different settings, it is critical to assess current capabilities in solving the CMCLP, especially given that it has increasingly been utilized through GIS as part of integrated planning environments.

Consider the following notation:

i = index of demand areas (I entire set)

j = index of potential facilities (J entire set)

d_{ij}= travel distance/cost/time between demand i and facility j

S = service coverage standard

$N_i = \{j | d_{ij} \leq S\}$

a_i = amount of demand in area i

c_j = capacity of facility j

p = number of facilities to site

Decision variables:

$$X_j = \begin{cases} 1 & \text{if facility } j \text{ is sited} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{ij} = \begin{cases} 1 & \text{if demand } i \text{ is allocated to facility } j \\ 0 & \text{otherwise} \end{cases}$$

One detail to point out regarding N_i is that it is the set of potential facility sites that can suitably cover demand i. This is predicated on the nature of service combined with the associated standard S. Normally, it is assumed that $a_i > 0$ and $c_j > 0$. With the above notation, the CMCLP can be formulated as follows:

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{2.1}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{2.2}$$

$$\Sigma_{j \in J} X_j = p \tag{2.3}$$

$$\Sigma_i a_i Y_{ij} \leq c_j X_j \quad \forall j \in J \tag{2.4}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{2.5}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{2.6}$$

The objective, (2.1), seeks to maximize the total demand served within the coverage standard. This is equivalent to the objective of the MCLP, but involves the use of allocation variables to indicate which demand is served by a specific facility. Constraints (2.2) indicate that each demand is to be allocated to at most one facility. Exactly p facilities are to be sited in Constraint (2.3). Capacity limits are imposed in Constraints (2.4). The left hand side of Constraints (4) sums the total demand allocated to facility j. The right hand side of Constraints (2.4) establishes the upper limit on demand that can be allocated to facility j. There are two roles served by Constraints (2.4): first, the facility workload of a facility cannot be greater than its upper limit if it is sited; second, a demand can never be allocated to a facility that is not sited. Constraints (2.5) define binary conditions for the location variable associated with each potential facility site. Constraints (2.6) define binary conditions on each allocation variable.

Formulation of the CMCLP along the lines structured using (2.1)-(2.6) requires $\Sigma_{i \in I}|N_i| + |J|$ decision variables, where $|\cdot|$ is the number of elements in the associated set. The number of constraints for this approach is $\Sigma_{i \in I}|N_i| + |I| + 2|J| + 1$.

The MCLP can be viewed as a special case of a CMCLP with an unlimited capacity for each facility. Specifically, the MCLP is equivalent to the CMCLP when $c_j = \infty$ for all j. Constraints (2.4) would then become $\Sigma_i a_i Y_{ij} \leq \infty$ when $X_j = 1$. Thus, no limits are imposed on facility workloads. Alternatively, Constraints (2.4) take the following form $\Sigma_i a_i Y_{ij} \leq 0$ when $X_j = 0$. This then forbids demand from being allocated to facilities that have not been sited.

When allocation variables are utilized, Balinski constraints are sometimes added to enhance structural properties, giving a model so called integer-friendliness when solved by exact methods (Church and Roberts, 1983; ReVelle, 1993; Gerrard, 1995; Church and Murray, 2009). These constraints take the following form:

$$Y_{ij} \leq X_j \tag{2.7}$$

The result is that $\Sigma_{i \in I}|N_i|$ constraints are added, bringing the total number of constraints to $2\Sigma_{i \in I}|N_i| + |I| + 2|J| + 1$. Again, the benefit is that Balinski constraints often enhance solution efficiency, if linear programming with branch-and-bound is used.

Worth noting is that a formulation nuance of the MCLP, the situation where a CMCLP has $c_j = \infty$ for all j, is to adopt binary coverage decision variable that indicates whether or not a demand i is suitably covered (Church and ReVelle, 1974). This avoids the use of allocation decision variables Y_{ij}, requiring significantly fewer variables and constraints. However, it is impossible to derive facility workloads without tracking allocation explicitly in the formulation.

Since the primary emphasis of the CMCLP is capacity limits, further discussion of the role of capacities is necessary. First, more balanced facility workloads are an expected byproduct due to the fact that each facility bounds the demand that it can be allocated. Second, significant computational complexity arises when capacity limits are introduced. Third, some demand within the coverage standard might not be allocated or could be dispatched to further away facilities in order to satisfy capacity constraints.

First, the upper bound of the workload range of a CMCLP solution is never greater than that of a corresponding optimal MCLP solution. Let J^* be the set of sited facilities, The range of facility workload for any solution is as follows:

$$\max_{j \in J^*}\{\Sigma_i a_i Y_{ij}\} - \min_{j \in J^*}\{\Sigma_i a_i Y_{ij}\} \tag{2.8}$$

This suggests the following proposition.

Proposition. The workload range of an optimal CMCLP solution, UB^{CMCLP}, is bounded by the corresponding workload range of an optimal MCLP, UB^{MCLP}, implying that $UB^{CMCLP} \leq UB^{MCLP}$.

Proof. Let R_j be the set of demand that can be suitably covered by facility j. Given capacity constraints (2.4), the workload of a facility j cannot exceed its capacity bound as well as the total demand it is able to cover. That is $\Sigma_i a_i Y_{ij} \leq \min\{c_j, \Sigma_{i \in R_j} a_i\}$ for all j. Thus, the maximal facility workload of a sited facilities in a CMCLP solution is $\max_j\{\min\{c_j, \Sigma_{i \in R_j} a_i\}\}$. The minimal workload of a sited facility j is $\min_{i \in R_j} a_i$. Therefore, the minimal facility workload in the system is $\min_j\{\min_{i \in R_j} a_i\}$. Accordingly, an upper bound of the CMCLP solution $UB^{CMCLP} = \max_j\{\min\{c_j, \Sigma_{i \in R_j} a_i\}\} - \min_j\{\min_{i \in R_j} a_i\}$.

When $c_j = \infty$ in constraints (2.4), the model is the MCLP. In this case, the maximal workload of sited facilities would be $max_j \Sigma_{i \in R_j} a_i$. The minimal facility workload of sited facilities for the MCLP is still $\min_j\{\min_{i \in R_j} a_i\}$. Thus, an upper bound of the MCLP

solution $UB^{MCLP} = max_j\{\Sigma_{i\in R_j} a_i\} - \min_j\{\min_{i\in R_j} a_i\}$. Since $\max_j\{\min\{c_j, \Sigma_{i\in R_j} a_i\}\} \leq max_j\{\Sigma_{i\in R_j} a_i\}$, $UB^{CMCLP} \leq UB^{MCLP}$.

It is easy to see that with tight capability limits c_j, the workload range in capacitated problems would be more restricted with a smaller UB^{CMCLP}. In addition, workloads of sited facilities will tend to approach/reach imposed limits c_j. This is because the objective (2.1) seeks maximum total demand served. ■

Second, capacity limits introduce computational difficulties. There are two conditions for a facility in a CMCLP solution: not sited or sited. If $X_j = 0$, a facility j is not sited, the allocation decision associated with j in the CMCLP are trivial. Because of capacity constraint (2.4), for facility j the condition becomes $\Sigma_i a_i Y_{ij} \leq 0$. Then, Y_{ij} for all i. Alternatively, if $X_j = 1$, the CMCLP can be viewed as a 0-1 knapsack problem for each sited facility j:

$$\text{Maximize} \quad \Sigma_{i\in R_j} a_i Y_{ij} \tag{2.9}$$

$$\text{Subject to} \quad \Sigma_{i\in R_j} a_i Y_{ij} \leq c_j \tag{2.10}$$

$$Y_{ij} = \{0, 1\} \tag{2.11}$$

Note that R_j is a set that indicates those demand i that can be suitably served within the coverage standard of facility j. The 0-1 knapsack problem is an integer linear programming problem and is NP-hard (Gary and Johnson, 1979). Branch-and-bound procedures coupled with linear programming are a general purpose integer linear pro-

gramming technique (Church and ReVelle, 1974; ReVelle, 1993). The idea is to solve a linear programming problem that ignores binary restrictions on a node of a tree, obtaining an upper bound for a maximization problem. If some variables violate binary constraints, a branch of the tree is created by fixing a fractional variable to 0 and 1. Each problem instance is then re-solved. This continues until all variables are integer for a node and the node's objective value is greater than or equal to that of any other terminal node (Efroymson and Ray, 1966). The introduction of capacity c_j would tend to make allocation variables Y_{ij} fractional when solving the relaxed linear problem. Here, $\Sigma_{i \in R_j} a_i > c_j$ is assumed without loss of generality. Define r such that $\Sigma_{i=1}^{r-1} a_i \leq c_j$ and $\Sigma_{i=1}^{r} a_i > c_j$ and $|R_j|$ the number of demand areas that can be suitably covered by facility j. An optimal solution of the relaxed problem is (Wolsey and Nemhauser, 1999): $Y_{ij} = 1$ for $i = 1, 2, \ldots, r-1$, $Y_{ij} = 0$ for $i = r+1, r+2, \ldots, |R_j|$ and $Y_{rj} = \frac{c_j - \Sigma_{r=1}^{r-1} a_i}{a_r}$, which is likely to be fractional in order to satisfy capacity constraint (2.10). Thus, it can be difficult for a branch-and-bound procedure to find an integer solution for the 0-1 knapsack problem. Therefore, it is even more computationally intensive to solve a CMCLP that is comprised of many 0-1 knapsack problems.

Third, the provided allocation response might not be desirable. Let J^* be the set of selected facilities and a demand i is within the covering standard of selected facilities, that is $J^* \cap N_i \neq \emptyset$. The demand i would not be served if for all $j \in J^* \cap N_i$, $Y_{ij} = 0$ is part of optimal solutions of associated 0-1 knapsack problems. A demand would be

dispatched to a further away facility if allocation to a closer sited facility violates the capacity limit and there is sufficient capacity to accommodate it elsewhere.

The CMCLP is an important capacitated model, but users face several challenges caused by capacity limits. The introduction of capacity limits imposes a more restrictive upper bound of the facility workload range, so it leads to a more balanced system to some extent. However, capacity constraints mean that linear programming based solution approaches will likely encounter highly fractional solutions, making it far more difficult for a branch-and-bound technique to identify an optimal integer solution. Therefore, due to capacities, the CMCLP can be computationally intensive to solve optimally. In addition, some demand within the covering standard might not be served or may be dispatched to a further away facility in order to satisfy capacity restrictions. The associated properties of the CMCLP mean that greater understanding and further empirical evaluation is essential.

2.4 Application Results

Two application case studies are utilized in the empirical assessment that follows. These studies help to highlight practical issues in applying the CMCLP, but also serve to provide insight regarding heuristic performance of commercial GIS software. The CMCLP is structured and solved by Location-Allocation functionality available in Network Analysis toolbox of ArcGIS (version 10.5). To establish a comparative basis, the CMCLP is also

structured and solved optimally using an exact solver, FICO Xpress (version 8.4) that uses the simplex algorithm combined with branch-and-bound. Corresponding MCLP instances are also structured and solved for comparison. Origin-destination matrices are exported from GIS, characterizing travel distance/cost between potential facilities and demand. Facility capacity, c_j, is established through evaluation of the maximal demand that can be suitably covered without capacity limits (i.e. the MCLP optimal objective value) divided by the number of facilities then multiplied by a coefficient α. Specifically, $c_j = \alpha \times$(MCLP optimal objective value)$/p$. The factor, α, adjusts facility capacity and $\alpha = 1$ initially. All empirical results are obtained using a desktop personal computer (Intel Xeon E5 CPU 2.30 GHz with 64 GB RAM and 4 cores).

The first case study involves analysis and planning to support postal service in the City of San Jose. The accuracy and efficiency of postal service is essential for business. Demand for service is considered using 32 ZIP Code Tabulation Areas (U.S. Census Bureau, 2018b) with a total population (2010) of 1,023,791 people. The population of each area is regarded as service demand. The ZIP Code Tabulation Area centroid is used to represent the demand point, as well as considered as a potential postal service facility location. Travel information between demand and potential facilities is computed based on a network extracted from The City of San Jose (2019). The covering standard to access a facility is a maximal service distance of 5 miles. The resulting problem has 32 demand points and 32 facility sites. CMCLP (and MCLP) instances with 1 to 10 facilities (i.e., varying p from 1 to 10) are explored.

Table 2.2 CMCLP results (total demand served) for San Jose case study

p	**ArcGIS**	**Xpress**	**Optimality deviation (%)**
1	393,050	393,050	0.000
2	636,850	636,850	0.000
3	771,132	781,776	1.362
4	805,660	844,022	4.545
5	827,304	844,022	1.981
6	844,022	844,102	0.009
7	844,022	864,200	2.335
8	844,022	889,601	5.124
9	872,883	889,681	1.888
10	857,749	889,681	3.589

The summary of findings for each CMCLP can be found in Table 2.2 and Figure 2.1(a). ArcGIS is unable to find optimal solutions in 8 of 10 problem instances. This is because the heuristic in ArcGIS cannot guarantee that optimal solutions are found while the exact solver using Xpress can verify solution optimality. Only when p is equal to 1 or 2, the heuristic in ArcGIS successfully identifies optimal solutions. When $p = 8$, the optimality deviation is 5.124%, which amounts to 45,579 people. This means that the optimal configuration of postal service facilities can serve 45,579 more people within the service covering standard than what was provided by the heuristic solution generated by ArcGIS. This is significant because the additional costs and effort associated with adopting a non-optimal solution are substantial, if a comparable level of service coverage is to be achieved. The median optimality deviation is 1.934%. Solution time is summarized in Figure 2.1(b). The computational time for ArcGIS is stable with an average of 0.553 seconds across the 10 application instances. The computational time for Xpress increases with p. Solution

time is significantly longer, requiring 44.635 seconds for $p = 7$. Note that solving the CMCLP and MCLP in Xpress requires the origin-destination matrix to be exported from ArcGIS, which takes 2.230 seconds.

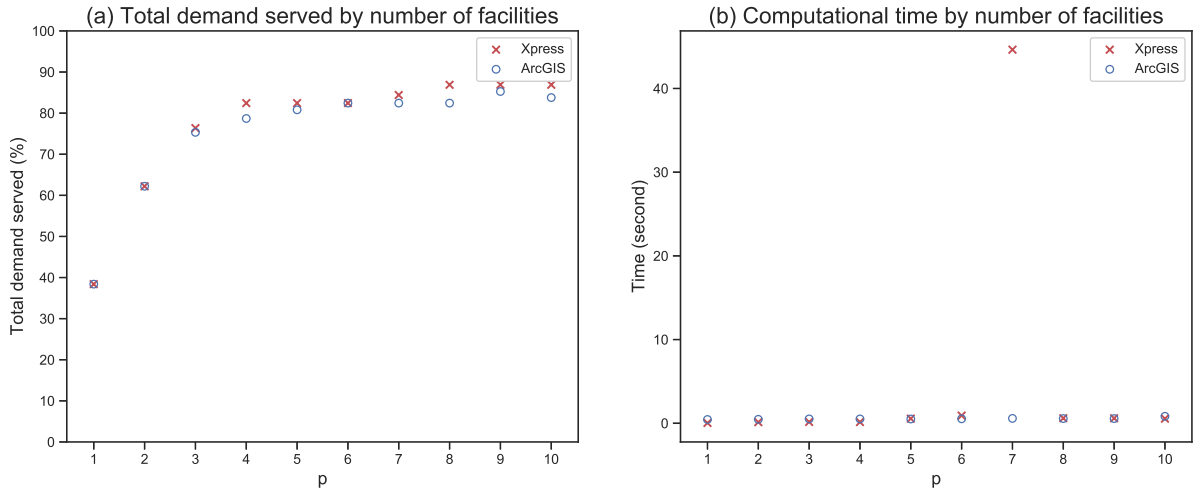

Figure 2.1 Summary of CMCLP solution details for the San Jose case study

Allocation offered by optimal solutions is also investigated. First, with imposed facility capacities, facility workload ranges, $\max_{j \in J^*}\{\Sigma_i a_i Y_{ij}\} - \min_{j \in J^*}\{\Sigma_i a_i Y_{ij}\}$, are significantly less than workload ranges in the case of the MCLP (Table 2.3), except when $p = 2$. The average workload range of optimal CMCLP solutions is 44,455 people compared to an average of 214,669 people for corresponding MCLP solutions. Second, there exists demand unserved even though it is within the service covering standard when $p = 7$ (Figure 2.2). Demand 31, representing 25,401 people, does not have access to service provided by the seven site facilities for the optimal allocation using the CMCLP. However, the distances between Demand 31 and two selected facilities (Facilities 17 and 23)

are around 3 miles, less than the maximal service distance of 5 miles. This means that 25,401 people, which account for 2.855% of total demand within the covering standard, are not provided service. In addition to withholding service, there are 381,945 people from 10 demand areas that are allocated to non-closest facilities when $p = 7$, leading to additional 719,795.961 miles in total travel (Figure 2.2). Most notably, Facilities 17 and 23 are sited but Demand 17 and 23 (in nearby areas) would be allocated to other facilities. More non-closest allocation can be found in all application instances except the case where one facility is sited (Table 2.4). A significant amount of demand is allocated to further away facilities, resulting in 671,531.689 more miles in total travel distance on average.

Table 2.3 Facility workload range for San Jose case study

p	**MCLP**	**CMCLP**
1	0	0
2	2,174	13,904
3	121,358	17,273
4	288,112	13,214
5	293,392	32,769
6	287,010	76,464
7	270,833	91,833
8	317,003	77,721
9	258,795	65,525
10	308,015	55,849

Analysis and planning to support the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) in the Santa Barbara area (Goleta, Santa Barbara and Carpinteria) is also investigated. This federally funded program aims to provide

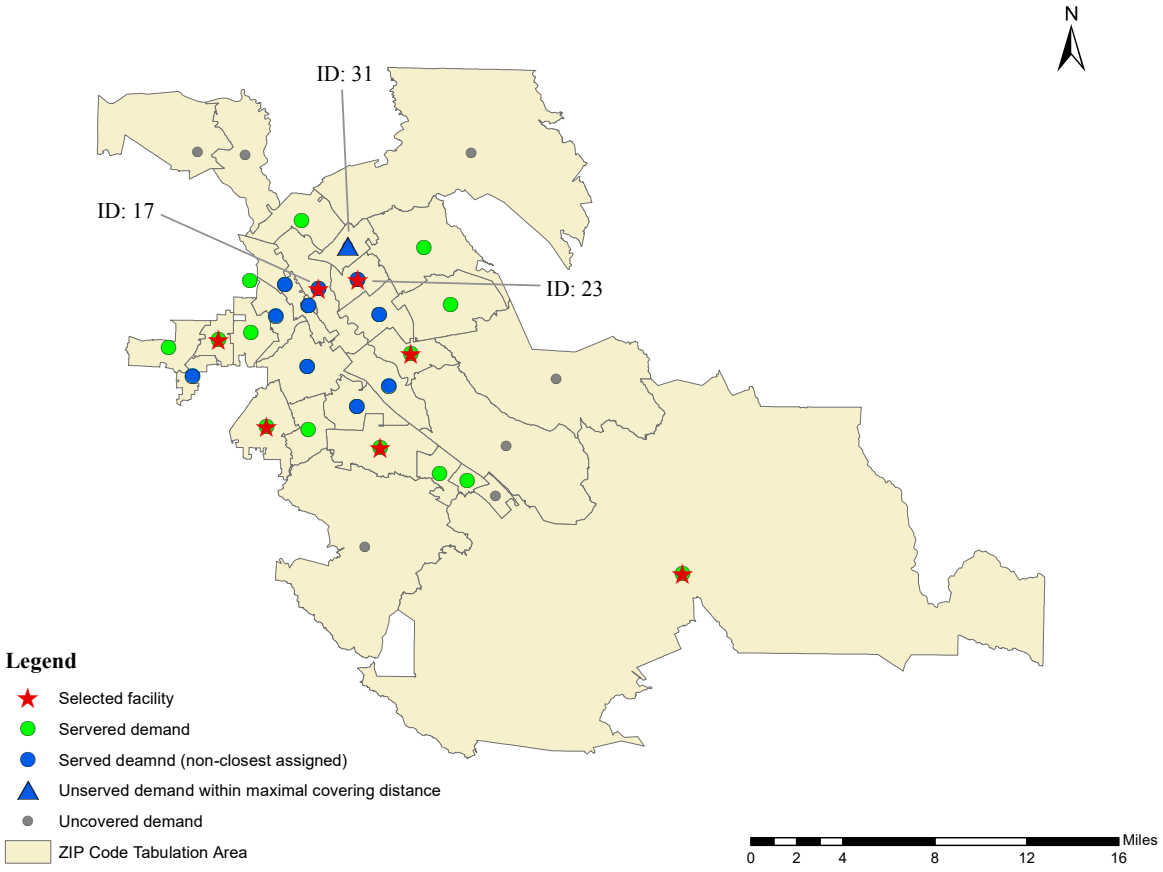

Figure 2.2 CMCLP-derived facility configuration ($p = 7$) and withholding service in the San Jose case study

nutrition, healthy foods, breastfeeding education and health care service across the United States. The locations of WIC facilities and associated service allocation in Santa Barbara are considered. There are 2,070 census blocks in the region with a total population (2010) of 200,450 people. The population of each block is used as a proxy for service demand, represented using the centroid. There are 82 locations identified as potential WIC facilities. The road network is constructed using data downloaded from U.S. Census

Bureau (2018a). The service coverage standard is assumed to be 5 miles. Siting 1 to 8 facilities under capacity restrictions is considered.

Table 2.4 Non-closet allocation for San Jose case study

p	**Demand not allocated to closet facilities**	**Total weighted distance to closet facilities (mile)**	**Total weighted distance to allocated facilities (mile)**	**Distance difference (mile)**
1	0	0	0	0
2	86,976	212,777.599	385,873.873	173,096.274
3	94,907	198,957.989	394,689.289	195,731.300
4	164,565	377,184.379	542,493.900	165,309.522
5	332,247	588,347.791	1,358,672.029	770,324.238
6	276,989	193,914.388	959,206.769	765,292.382
7	381,945	741,504.511	1,461,300.472	719,795.961
8	530,828	747,646.767	1,894,780.189	1,147,133.422
9	570,992	746,412.817	2,247,774.870	1,501,362.053
10	435,246	229,336.481	1,506,608.215	1,277,271.735

The summary findings for applying the CMCLP can be found in Table 2.5 and Figure 2.3(a). Of the 8 problem instances solved using the ArcGIS heuristic, only one is optimal ($p = 1$). Thus, 87.5% of the application instances are not solved optimally by the heuristic. The optimal configuration of selecting 3 facilities is given in Figure 2.4. These 3 selected facilities can serve 182,749 people while facilities found using the ArcGIS heuristic can only serve 170,144 people. This means that the configuration suggested by ArcGIS would not serve 12,605 people within the standard as compared to what can be optimally served. The median optimality deviation is 1.685%. Computational time is presented in Figure 2.3(b). Generally, both ArcGIS and Xpress have an increasing trend with p. ArcGIS requires significantly less computational time with an average

of 30.844 seconds. In contrast, solution using Xpress requires an average of 1,232.300 seconds. Note that exporting of the origin-destination matrix from ArcGIS for Xpress takes 22.420 seconds.

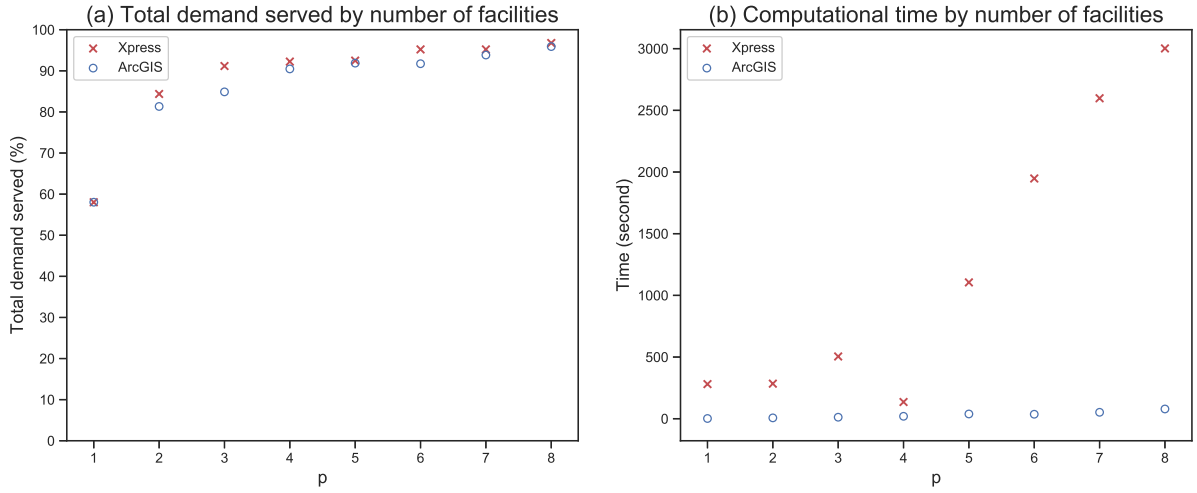

Figure 2.3 Summary of CMCLP solution details for the Santa Barbara case study

Characteristics of CMCLP allocations were also evaluated. First, significantly less facility workload variation is observed in CMCLP solutions compared to the MCLP, which is not surprising. The average workload ranges are 6,046 and 55,292 people for the CMCLP and MCLP, respectively. When $p = 3$, the most utilized facility serves 5,330 more people than the least utilized facility in the CMCLP (Figure 2.4) while the range reaches 84,323 people in corresponding MCLP solutions (Figure 2.5). Second, there are two among eight problem instances giving optimal allocation strategies stipulating a total of 6,216 people from 369 blocks within the standard would not be served. When $p = 2$,

some 2.714% of demand within 5 miles from selected two facilities, 4,719 people, would not be served (Figure 2.6). It can be observed as well that some unserved demand within the maximal service distance is located much closer than some served demand. Service is withheld for 1,497 people within the maximal service distance when selecting eight facilities. Third, in the eight application instances, a total of 452,446 people are not allocated to their closest sited facilities and have to travel a total of 803,713.691 miles further for service.

Table 2.5 CMCLP results (total demand served) for Santa Barbara case study

p	**ArcGIS**	**Xpress**	**Optimality deviation (%)**
1	116,327	116,327	0.000
2	163,028	169,122	3.603
3	170,144	182,749	6.897
4	181,337	184,909	1.932
5	184,158	185,376	0.657
6	183,873	190,853	3.657
7	188,110	190,853	1.437
8	192,231	193,980	0.902

2.5 Discussion

There are a number of potential issues surrounding the application of the CMCLP worth further investigation and discussion. More empirical assessment was conducted on the impacts of different facility capacities, accomplished by varying the factor α in both case

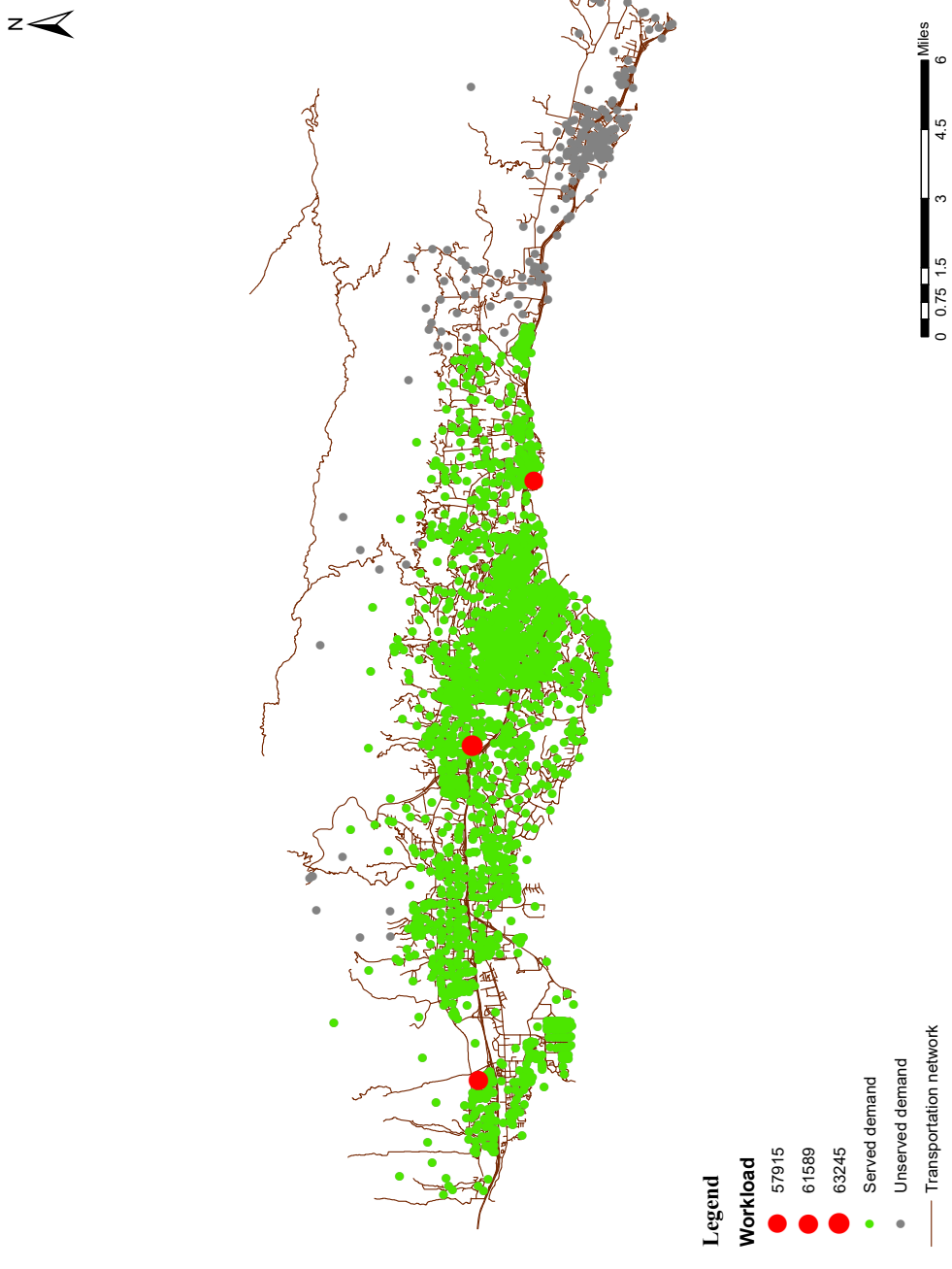

Figure 2.4 CMCLP-derived facility configuration ($p = 3$) and facility workloads in the Santa Barbara case study

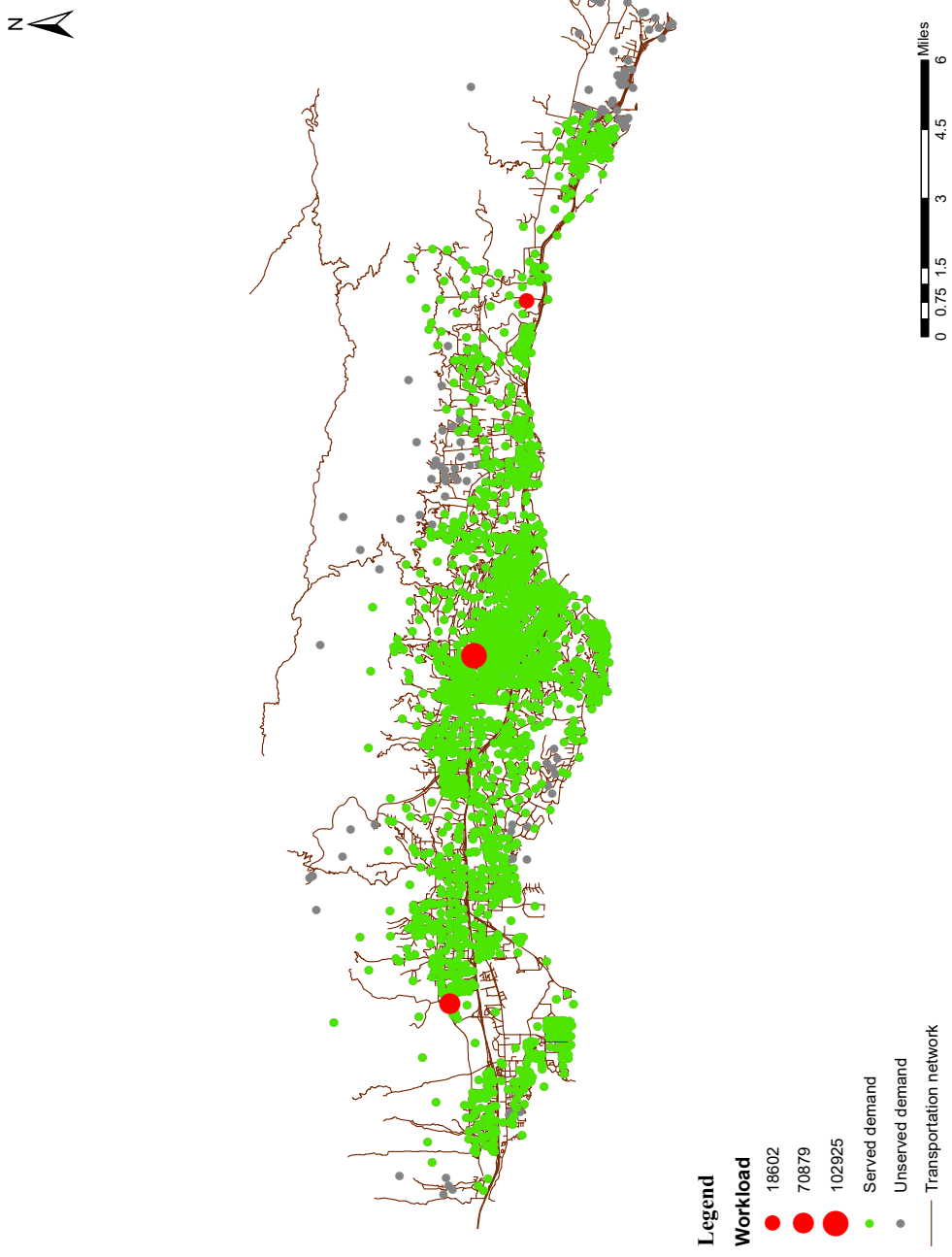

Figure 2.5 MCLP-derived facility configuration ($p = 3$) and facility workloads in the Santa Barbara case study

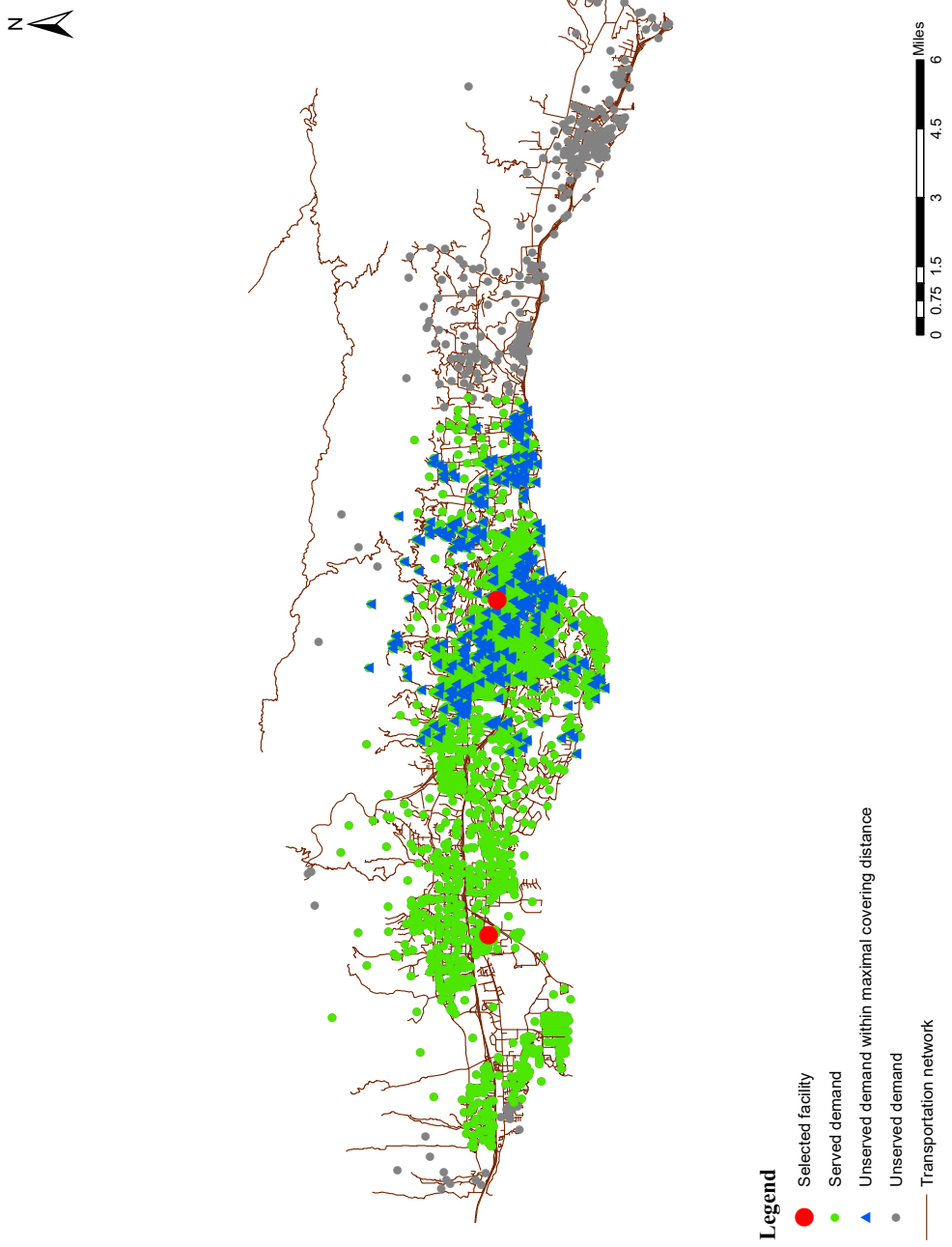

Figure 2.6 CMCLP-derived facility configuration ($p = 2$) and withholding service in the Santa Barbara case study

studies. Specifically, α is varied from 0.2 to 2.0 in intervals of 0.2. For example, an α of 0.2 means a relatively low facility capacity, that is 20% of the average workload of facilities compared to the corresponding MCLP optimal solution. With each α value, evaluation of the CMCLP along the lines reported in Table 2.2 and Figure 2.1 was repeated for applications. As a result, an additional 90 problem instances for the postal service context and 72 more instances for WIC in Santa Barbara were considered. In total, 180 different CMCLP applications were examined.

Table 2.6 Capacity variation impacts on CMCLP results (total demand served) for San Jose case study

α	p	**ArcGIS**	**Xpress**	**Optimality deviation (%)**
0.2	1	78,374	78,374	0.000
	2	131,735	132,114	0.287
	3	159,803	161,677	1.159
	4	167,009	167,009	0.000
	5	159,548	163,624	2.491
	6	164,018	164,018	0.000
	7	134,293	134,293	0.000
	8	108,892	108,892	0.000
	9	108,892	108,892	0.000
	10	108,892	108,892	0.000
0.4	1	157,156	157,202	0.029
	2	264,349	264,737	0.147
	3	322,806	325,825	0.927
	4	340,641	343,876	0.941
	5	342,823	353,705	3.077
	6	367,464	370,797	0.899
	7	366,738	366,738	0.000
	8	367,689	370,770	0.831
	9	364,880	364,880	0.000
	10	339,210	352,882	3.874

Table 2.6 (continued)

α	p	**ArcGIS**	**Xpress**	**Optimality deviation (%)**
0.6	1	235,665	235,790	0.053
	2	395,615	397,050	0.361
	3	486,299	488,758	0.503
	4	504,302	514,176	1.920
	5	534,404	538,939	0.841
	6	538,605	555,742	3.084
	7	566,347	574,434	1.408
	8	524,931	569,022	7.749
	9	535,039	574,071	6.799
	10	580,533	586,804	1.069
0.8	1	313,456	314,330	0.278
	2	524,418	528,883	0.844
	3	645,574	650,480	0.754
	4	673,863	686,062	1.778
	5	715,271	718,441	0.441
	6	705,637	740,221	4.672
	7	742,753	766,188	3.059
	8	717,011	757,925	5.398
	9	756,991	771,725	1.909
	10	746,635	759,854	1.740
1.0	1	393,050	393,050	0.000
	2	636,850	636,850	0.000
	3	771,132	781,776	1.362
	4	805,660	844,022	4.545
	5	827,304	844,022	1.981
	6	844,022	844,102	0.009
	7	844,022	864,200	2.335
	8	844,022	889,601	5.124
	9	872,883	889,681	1.888
	10	857,749	889,681	3.589
1.2	1	393,050	393,050	0.000
	2	662,036	662,036	0.000
	3	790,056	815,457	3.115
	4	830,111	844,022	1.648
	5	844,022	889,601	5.124
	6	844,102	889,601	5.115

Table 2.6 (continued)

α	p	**ArcGIS**	**Xpress**	**Optimality deviation (%)**
	7	872,883	901,335	3.157
	8	889,681	929,900	4.325
	9	889,681	929,980	4.333
	10	892,511	967,357	7.737
1.4	1	393,050	393,050	0.000
	2	662,036	662,036	0.000
	3	815,457	815,457	0.000
	4	844,022	844,318	0.035
	5	872,883	889,601	1.879
	6	872,883	929,900	6.132
	7	889,601	929,980	4.342
	8	889,601	967,357	8.038
	9	929,900	967,437	3.880
	10	929,980	987,477	5.823
1.6	1	393,050	393,050	0.000
	2	662,036	662,036	0.000
	3	815,457	815,457	0.000
	4	844,022	861,036	1.976
	5	872,883	889,601	1.879
	6	889,601	929,900	4.334
	7	913,182	967,357	5.600
	8	929,900	967,437	3.880
	9	950,639	987,477	3.731
	10	950,639	987,557	3.738

Table 2.6 summarizes empirical findings for San Jose. Of the 100 problem instances for San Jose using the ArcGIS heuristic, 75 instances are not optimal. The optimality deviation tends to be greater when selecting more facilities or when facility capacities are higher. This implies the CMCLP with higher p (number of facilities) and larger c_j (capacity) are more difficult to solve for the ArcGIS heuristic. One potential reason is

Table 2.6 (continued)

α	p	**ArcGIS**	**Xpress**	**Optimality deviation (%)**
1.8	1	393,050	393,050	0.000
	2	662,036	662,036	0.000
	3	815,457	815,457	0.000
	4	844,022	861,036	1.976
	5	872,883	901,335	3.157
	6	889,601	929,900	4.334
	7	913,182	967,357	5.600
	8	929,900	987,477	5.831
	9	950,639	987,557	3.738
	10	950,639	1,006,446	5.545
2.0	1	393,050	393,050	0.000
	2	662,036	662,036	0.000
	3	815,457	815,457	0.000
	4	844,022	861,036	1.976
	5	872,883	901,335	3.157
	6	913,182	929,900	1.798
	7	913,182	967,357	5.600
	8	929,900	987,477	5.831
	9	950,639	1,006,446	5.545
	10	950,639	1,006,446	5.545

that more facilities to site and higher facility capacities are associated with a larger feasibility region in solution search processes, making it harder for the heuristic to find an optimal solution. Relying on the ArcGIS heuristic for solving the CMCLP, 7.595% of the total demand, amounting to 77,751 people, would not be suitably covered compared to the optimal configuration in the worst case ($p = 8, \alpha = 1.4$). The median deviation from optimality is 1.839%. The optimality deviation revealed here for the CMCLP is generally much larger than that reported in Murray et al. (2019) for solution of the MCLP.

This implies greater computation complexity for solving the CMCLP, making it more challenging for heuristics to identify optimal or close to optimal solutions. Computationally, the average time for the ArcGIS heuristic remains stable (around 0.537 seconds) over different facility capacity levels. Generally, solution using Xpress requires more time compared to ArcGIS, especially for more restrictive cases (e.g., $\alpha = 0.4,\ 0.6,\ 0.8,\ 1.0$ and 1.2). This is due to Xpress using branch-and-bound when solving associated integer linear programming problems. To satisfy tighter capacity limits, the binary allocation decision variables tend to be fractional in relaxed linear programs, which ultimately makes the branch-and-bound tree very deep in the search for feasible integer solutions and requires more computational effort. Most notably, it takes about 153.042 seconds on average for Xpress to solve problems when $\alpha = 1.0$.

Reported in Table 2.7 are 42 instances found where service was withheld for demand located within the coverage standard out of 100 problems for San Jose. Such instances usually occur when facility capacities are low. The amount of unserved demand within the maximal service distance tends to decrease with increasing facility capacities. Due to higher capacity bounds, facilities are able to serve more demand, leaving less demand unserved within the service standard. The most conspicuous instance of withholding service is the case where $p = 10$ and facility capacities are the lowest, $\alpha = 0.2$. In this case, 85.932% of people (665,127) living within 5 miles from selected facilities are not served within capacity limits. The median percentage of withholding service is 37.813%.

In addition, there are 86 cases where an average of 223,030 people per case are not allocated to their closest facilities, necessitating longer access travel.

Table 2.7 Capacity variation impacts on unserved demand within the service standard for San Jose case study

p	α	# of unserved demand area within the standard	# of unserved demand amount within the standard	Percentage of withholding service (%)
1	0.2	7	279,213	78.083
	0.4	7	173,907	52.523
	0.6	3	157,260	40.010
	0.8	2	61,719	16.412
2	0.2	11	504,736	79.255
	0.4	7	325,934	55.180
	0.6	7	220,803	35.737
	0.8	4	107,967	16.953
3	0.2	9	325,690	66.826
	0.4	10	393,723	54.718
	0.6	5	201,397	29.181
	0.8	2	39,675	5.749
4	0.2	11	410,983	71.105
	0.4	8	276,878	44.603
	0.6	6	313,128	37.849
	0.8	2	79,014	10.328
5	0.2	10	457,130	73.641
	0.4	10	473,599	57.246
	0.6	7	226,137	29.557
	0.8	5	80,053	10.025
6	0.2	7	330,798	66.853
	0.4	11	427,777	53.568
	0.6	6	288,280	34.156
	0.8	3	103,801	12.298
7	0.2	13	565,296	80.804
	0.4	10	431,756	54.071

Table 2.7 (continued)

p	α	**# of unserved demand area within the standard**	**# of unserved demand amount within the standard**	**Percentage of withholding service (%)**
	0.6	5	224,060	28.060
	0.8	2	61,116	7.387
	1.0	1	25,401	2.855
	1.2	1	28,565	3.072
8	0.2	12	525,534	82.836
	0.4	12	559,130	60.128
	0.6	6	275,080	32.588
	0.8	3	69,379	8.386
9	0.2	14	619,548	85.051
	0.4	11	511,370	58.359
	0.6	7	270,031	31.990
	0.8	2	72,377	8.574
10	0.2	15	665,127	85.932
	0.4	12	568,896	61.717
	0.6	6	257,218	30.475
	0.8	1	45,528	5.653

Similar results have been found for Santa Barbara. There are 38 (47.5%) problem instances of 80 instances where the ArcGIS heuristics cannot provide optimal solutions (Table 2.8). When facility capacities increase, more instances appear with non-optimal solutions provided by ArcGIS. In the worst case ($p = 4, \alpha = 1.8$), the ArcGIS solution miss 19,941 (9.948%) people who should have been served. Reported in Table 2.9 are 36 instances where some demand is withheld service even though it is within the service standard, which usually occurs when capacities are tight. The maximal percentage of withholding service is up to 78.806% ($p = 2, \alpha = 0.8$) and the median is 37.262%. Of

the 80 instances, some demand is dispatched to a further away facility in 68 instances, travelling an average of 14,509.012 miles further for service.

Table 2.8 Capacity variation impacts on CMCLP results (total demand served) for Santa Barbara case study

α	p	**ArcGIS**	**Xpress**	**Optimality deviation (%)**
0.2	1	23,265	23,265	0.000
	2	34,886	34,886	0.000
	3	38,481	38,481	0.000
	4	39,880	39,880	0.000
	5	40,070	40,070	0.000
	6	40,086	40,086	0.000
	7	40,089	40,089	0.000
	8	40,088	40,088	0.000
0.4	1	46,531	46,531	0.000
	2	69,772	69,772	0.000
	3	76,962	76,962	0.000
	4	79,756	79,756	0.000
	5	80,135	80,135	0.000
	6	80,172	80,172	0.000
	7	80,171	80,171	0.000
	8	80,176	80,176	0.000
0.6	1	69,796	69,796	0.000
	2	104,660	104,660	0.000
	3	115,443	115,443	0.000
	4	119,636	119,636	0.000
	5	120,205	120,205	0.000
	6	120,252	120,252	0.000
	7	120,260	120,260	0.000
	8	120,256	120,256	0.000
0.8	1	93,062	93,062	0.000
	2	139,546	139,546	0.000
	3	153,924	153,924	0.000
	4	159,512	159,512	0.000
	5	160,270	160,270	0.000

Table 2.8 (continued)

α	p	**ArcGIS**	**Xpress**	**Optimality deviation (%)**
	6	160,338	160,338	0.000
	7	160,342	160,342	0.000
	8	160,344	160,344	0.000
1.0	1	116,327	116,327	0.000
	2	163,028	169,122	3.603
	3	170,144	182,749	6.897
	4	181,337	184,909	1.932
	5	184,158	185,376	0.657
	6	183,873	190,853	3.657
	7	188,110	190,853	1.437
	8	192,231	193,980	0.902
1.2	1	116,327	116,327	0.000
	2	174,223	174,432	0.120
	3	180,950	183,972	1.643
	4	182,470	196,307	7.049
	5	194,779	200,168	2.692
	6	199,327	200,367	0.519
	7	199,800	200,380	0.289
	8	200,167	200,426	0.129
1.4	1	116,327	116,327	0.000
	2	174,432	174,432	0.000
	3	179,444	189,513	5.313
	4	181,935	199,235	8.683
	5	195,145	200,340	2.593
	6	198,554	200,424	0.933
	7	200,167	200,426	0.129
	8	199,994	200,426	0.216

Table 2.8 (continued)

α	p	**ArcGIS**	**Xpress**	**Optimality deviation (%)**
1.6	1	116,327	116,327	0.000
	2	174,432	174,432	0.000
	3	179,444	192,406	6.737
	4	182,574	199,390	8.434
	5	196,502	200,340	1.916
	6	198,554	200,424	0.933
	7	198,554	200,434	0.938
	8	200,365	200,436	0.035
1.8	1	116,327	116,327	0.000
	2	174,432	174,432	0.000
	3	181,456	192,406	5.691
	4	179,558	199,394	9.948
	5	197,408	200,340	1.464
	6	198,554	200,424	0.933
	7	198,554	200,434	0.938
	8	200,365	200,436	0.035
2.0	1	116,327	116,327	0.000
	2	174,432	174,432	0.000
	3	181,456	192,406	5.691
	4	183,363	199,394	8.040
	5	195,145	200,340	2.593
	6	198,554	200,424	0.933
	7	198,890	200,434	0.770
	8	200,357	200,436	0.039

Table 2.9 Capacity variation impacts on unserved demand within the service standard for San Barbara case study

p	α	# of unserved demand area within the standard	# of unserved demand amount within the standard	Percentage of withholding service (%)
1	0.2	207	30,489	56.719
	0.4	531	35,375	43.190
	0.6	6	1,037	1.464
	0.8	394	14,002	13.078
2	0.2	818	84,409	70.757
	0.4	941	82,549	54.194
	0.6	1339	68,480	39.552
	0.8	492	23,009	14.155
	1.0	315	4,719	2.715
3	0.2	1626	136,646	78.027
	0.4	714	70,205	47.704
	0.6	308	9,521	7.619
	0.8	538	23,293	13.144
4	0.2	1444	11,7580	74.673
	0.4	863	81,456	50.527
	0.6	416	24,187	16.817
	0.8	451	22,121	12.179
	1.2	15	2,357	1.186
5	0.2	1610	136,432	77.298
	0.4	1544	112,133	58.321
	0.6	600	63,371	34.520
	0.8	531	23,968	13.009
	1.2	20	93	0.046
6	0.2	1636	140,150	77.759
	0.4	1459	96,877	54.718
	0.6	1329	69,861	36.747
	0.8	414	24,344	13.182
7	0.2	1634	141,656	77.942
	0.4	1622	103,741	56.408
	0.6	902	73,010	37.776

Table 2.9 (continued)

p	α	# of unserved demand area within the standard	# of unserved demand amount within the standard	Percentage of withholding service (%)
	0.8	516	29,877	15.707
8	0.2	1777	149,062	78.806
	0.4	1669	110,079	57.859
	0.6	1329	61,373	33.790
	0.8	925	29,469	15.525
	1.0	54	1,497	0.766

2.6 Conclusion

This chapter examined solution characteristics and allocation response of the Capacitated Maximal Covering Location Problem (CMCLP) because of its accessibility in GIS packages and the significant increase in reported applications (Table 2.1). A total of 180 application instances for San Jose and Santa Barbara were investigated. On one hand, GIS makes it accessible for general users to structure and solve a capacitated coverage problem. It provides an integrated environment for data acquisition, management, manipulation, analysis and display, which is generally not possible for exact solution approaches. Provided heuristics in ArcGIS did solve capacitated coverage problems with less computational effort than an exact solver. On the other hand, empirical results revealed that heuristics approaches in ArcGIS rarely found an optimal solution. There was a high possibility, around 62.778% of the application instances, that sub-optimal

solutions were identified. Optimality gaps were found to be as high as 9.948%. With the increasing usage of GIS for addressing capacitated covering applications in recent years (Table 2.1), these findings are significant. The implications for planning, management and decision making are largely unknown, and are dependent on context and situation. Nevertheless, exploration, understanding and insight are fundamentally important. The findings suggest that imposing capacity limits introduces computational complexity for solving the CMCLP and thus communication of heuristically obtained results needs to improve. Explicit technical description of internal algorithms provided by GIS is needed for better overall understanding. More user control for adjusting heuristic parameters in the future is also likely to be beneficial.

Allocation derived using the CMCLP offers a more balanced service system by preventing facility workloads from exceeding established limits. Theoretically, capacity limits impose a more restrictive upper bound of facility workload range. Empirically, without capacity limits, the facility workload range may vary substantially (see Table 2.3). However, there are many service implications. First, some demand may not be allocated for service, yet remains within the service standard. Second, some demand may be allocated to a further away facility, resulting in significantly more travel efforts for service providers or demand. Therefore, further investigation of allocation response provided by CMCLP solutions in application is essential.

Chapter 3

Service Allocation Equity in Maximal Covering

3.1 Introduction

Coverage is an important concept in location analytics, usually characterized as a maximum travel distance or time standard for a facility to suitably respond to a demand for service (Church and Murray, 2018). For example, firefighter's (and paramedic's) response within 8 minutes is critical to save property and lives, pizza (and food) delivery within 30 minutes ensures meals stay warm and fresh, and recycling centers within a half-mile of supermarkets facilitate cash redemption. These are all representative types of service systems where coverage is central. Underlying these examples is that service standards reflect acceptable access and accessibility. Based on the concept of coverage, various location models have been proposed and studied to mimic important service goals and strategies. A prominent approach is the location set covering problem (LSCP)

This chapter represents a revised version of a paper submitted to *European Journal of Operations Research*, co-authored with Dr. Alan T. Murray, Dr. Richard L. Church and Dr. Ran Wei.

that seeks the minimum number of facilities to serve all demand within a stipulated travel distance/time standard of a facility (Toregas et al., 1971). Given a limited budget and resources, the maximal covering location problem (MCLP) is another central approach, proposed to site a fixed number of facilities to optimize suitable service of demand (Church and ReVelle, 1974). The LSCP and MCLP have been broadly applied and extended in many ways (Chung, 1986; Church and Murray, 2018; Current and Storbeck, 1988; Daskin and Stern, 1981; Sorensen and Church, 2010; Murawski and Church, 2009).

A major limitation of traditional location coverage analytics is that they concentrate primarily on where to locate facilities, intentionally ignoring how demand is allocated to sited facilities. Of course, this makes it difficult to control the workloads of sited facilities (i.e., total demand served by a sited facility). The MCLP, for example, has an important underlying assumption that facilities being sited have unlimited capacities. That is, the workload of any sited facility can be as high as possible in order to serve all demand within the coverage standard. Alternatively, the workload of a sited facility could also be as excessively low. As a result, significant variation in facility workloads is not uncommon for service configurations identified using the MCLP, yet can be problematic and inequitable. Shown in Figure 3.1 are three sited clinics identified using the MCLP that maximize the total covered demand within a service distance of 5 miles in Santa Barbara, CA. The middle facility serves 102,925 people (51.3% of the total demand) while the facility on the right only serves 18,602 people (9.3% of the total demand). The workload variation exceeds a factor five. Such significant service allocation variation

represents a form of inequity that is not desirable in practice: on one hand, a facility that is overutilized may not be able to serve all allocated demand with timely and high quality service; on the other hand, it may not be economical to open a facility that is vastly underutilized. At a more practical level, workload imbalance along these lines can cause employee dissatisfaction, low morale, poor productivity, marginal economic returns and other negative effects as well. For these reasons, facility workload balance in coverage modeling remains an important research topic.

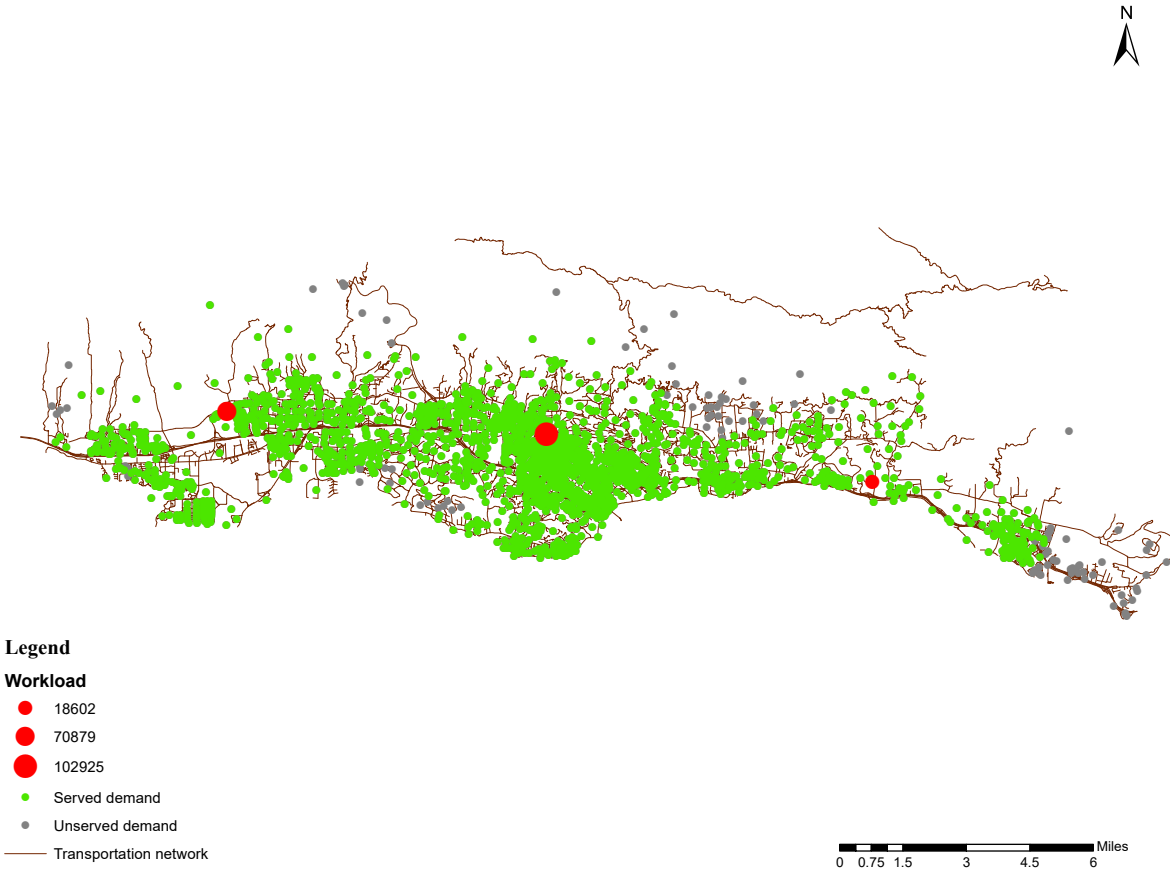

Figure 3.1 Workload imbalance for an MCLP solution

The prominent approach to deal with facility workload variation in coverage modeling has been to constrain total service provided by any one facility by adding capacity or/and threshold constraints to the model. The capacitated maximal covering location problem (CMCLP) was introduced along these lines, employing constraints to track and prevent the workload for each facility from exceeding an established limit (Chung et al., 1983; Church and Somogyi, 1985; Current and Storbeck, 1988; Elkady and Abdelsalam, 2016; Ferrari et al., 2018). Capacity limits can be employed and structured to shift demand from heavily burdened facilities to those with available resources, thus helping to balance facility workloads to some degree. Another approach is the McTHRESH (maximum coverage with thresholds) model, which maximizes the spatial market coverage and requires each facility to serve at least a given amount of demand (Balakrishnan and Storbeck, 1991). Threshold limits ensure that no open facility is underutilized, making each facility viable. There are also covering models with both capacity and threshold constraints, imposing upper and lower limits on facility workloads (Gerrard, 1995; Haghani, 1996). However, adding facility capacity and/or threshold constraints does not actually balance facility workloads in a direct manner. As a result, imbalanced facility workloads may still persist. In addition, the use of capacities and/or thresholds gives rise to several challenges, including appropriate specification of capacities and/or thresholds, uncertainty, undesirable allocation response and computational difficulty (Xu et al., 2020). There are, however, more direct ways to balance workloads that can be found in broader facility location work, including minimizing the maximum workload, workload range, and total

absolute difference. However, the capabilities of such approaches to address equity in coverage modeling remains largely unknown.

The purpose of this chapter is to study analytics that can be used to explicitly balance facility workloads in coverage modeling, and evaluate them in terms of solution quality and computational effort. A literature review related to this research is provided in Section 3.2. Then, five workload variation measures are studied and mathematical formulations for associated location cover analytics are given in Section 3.3. This is followed by a formalized evaluation approach proposed in Section 3.4. Empirical studies are carried out in Section 3.5 to assess and compare proposed models. Finally, this chapter ends with a discussion in Section 3.6 and conclusions in Section 3.7.

3.2 Background

Coverage models are prominent location optimization approaches (Church and Murray, 2018). Location covering primarily focuses on two basic objectives. One is to minimize the number of facilities necessary in order to ensure a sufficient level of coverage to each demand. Toregas et al. (1971) proposed the LSCP along these lines, seeking complete coverage using a minimal number of facilities. Considering limited budget and resource realities, it may be that not all demand can be covered. Thus, the other type of objective is to maximize coverage when siting a limited number of facilities. The MCLP introduced in Church and ReVelle (1974) was structured to address this, seeking to site a given num-

ber of facilities in such a manner that suitable coverage within the standard is provided to the most total demand possible. Many related studies have sought to extend the LSCP and MCLP in various ways. A number of extensions are based on the recognition of excessive facility workloads and uncertain availability of facilities. To consider possible busyness and unavailability of facilities, some research has focused on providing backup or redundant coverage (Daskin and Stern, 1981; Hogan and ReVelle, 1986; Bianchi and Church, 1988) while others have adopted probabilistic facility availability (Chapman and White, 1974; Daskin, 1983; Sorensen and Church, 2010). A prominent extension has been to track allocation and impose facility capacities and thresholds (Balakrishnan and Storbeck, 1991; Chung et al., 1983; Church and Somogyi, 1985; Current and Storbeck, 1988).

Imposing capacity and/or threshold constraints is popular for controlling facility workloads in coverage problems. Capacities represent an established maximum on demand that a sited facility can serve, thereby ensuring facility availability when service is needed. The CMCLP adds capacity limits to the MCLP. Chung et al. (1983) first formulated the CMCLP and proposed a substitution based heuristic to solve it. An alternative formulation of the CMCLP that allows partial assignments and multiple facilities at a site was proposed by Church and Somogyi (1985). The seminal work of Current and Storbeck (1988) discussed theoretical linkages between capacitated covering models, capacitated p-median problems and generalized assignment problems in addition to formulating the capacitated LSCP and CMCLP. Unfortunately, the application of the CMCLP is compli-

cated by challenges and issues. Pirkul and Schilling (1991) extended the basic CMCLP to allocate all demand regardless of proximity within the coverage standard through the use of an additional objective that minimizes total weighted distances for demand beyond the standard. Similar work can be found in Haghani (1996) and Yin and Mu (2012).

Thresholds impose a minimum amount of demand allocated to a sited facility, often done to ensure economic viability. Threshold constraints can be imposed to give lower bounds on facility workloads in coverage problems. Covering models with threshold constraints can be found in Church (1980), Current and Storbeck (1988), Carreras and Serra (1999), Drezner and Hamacher (2004) and Hong and Kuby (2016). A prominent example is the McTHRESH model which adds threshold constraints to the MCLP, ensuring a sufficient market to each facility Balakrishnan and Storbeck (1991). There are also studies considering both capacity and threshold constraints in covering models. For example, lower and upper limits on facility workloads were incorporated into the MCLP by Gerrard (1995) and Haghani (1996). Thresholds enable facility workloads to be controlled, similar to capacities. However, thresholds and capacities do not directly balance facility workloads, as noted previously.

More direct ways to balance facility workloads are found in work dealing with equity issues in location modeling. Starting with the early work of Mumphrey et al. (1971) and Savas (1978), facility equity has been incorporated in various location problems, especially in public sectors. Marsh and Schilling (1994) reviewed equity measures to quantify the effects of facility siting in location problems. They discussed 20 different

approaches that could be used to measure facility equity, such as range, variance, mean absolute deviation and others. Facility workloads are implicit in such concerns for equity, making them of broad general interest.

Minimizing maximum workload in a system is one way to achieve balance. For example, the maximum total demand assigned to a facility was minimized to reach equitable allocations by Berman et al. (2009) and Kim and Kim (2010). Similarly, Davoodi (2019) included two additional objectives to minimize the maximum clients a center serves and to minimize the range of workloads in a k-center problem. They also proposed an iterative algorithm based on a Voronoi diagram for solution. Workload range, which reflects dispersion, was utilized to account for balance. Marín (2011) located a fixed number of facilities to minimize the range of customer assignments between sited facilities, with restrictions that customers can only be assigned to their closest facilities. Weaver and Church (1981) minimized the range of assigned workloads while solving a vector assignment p-median problem with a special substitution heuristic. Daskin and Tucker (2018) minimized the range of assigned demand to obtain facility allocation equity as an extension to the p-median problem, proposing a genetic algorithm to solve the bi-objective problem. Similarly, D'Amico et al. (2002) restricted the ratio of the largest and smallest district areas from exceeding a prespecified upper bound in order to balance patrol car allocation in police command redistricting. Another approach appearing in the literature is to govern workload deviations from a system-wise average workload. Garfinkel and Nemhauser (1970) balanced workloads by imposing a constraint on population deviation

from the mean in districts. Zhu and McKnew (1993) developed a workload balancing model to control ambulances assigned to stations based on system-wise averages. A more detailed way is to control the total workload deviation between any two sited facilities, something used to equalize school utilization rates in district planning (Church and Murray, 1993). While researchers have considered facility workload balance in different contexts, the relative capabilities of approaches to address equity in the context of maximal coverage remain largely unknown. Thus, this chapter is focused on studying modeling approaches that can directly balance workloads in coverage modeling given significant observed increases in the application of capacitated coverage modeling (Xu et al., 2020). Such a trend was shown to be facilitated by access to capacitated coverage approaches in commercial geographic information system software, such as ArcGIS, where problems are solved using heuristics.

3.3 Methods

This section details existing and proposed analytic models that consider equitable facility workload balance in the context of maximal covering. The intent is to identify measures that more accurately account for workload balance, and can be readily incorporated in the maximal covering formulation. In this section, the MCLP is introduced first. Then, various workload variation measures are discussed, followed by model formulations that incorporate these different measures.

3.3.1 MCLP

Consider the following notation:

i = index of demand areas (I entire set)

j = index of potential facilities (J entire set)

d_{ij}= travel distance/cost/time between demand i and facility j

S = service coverage standard

$N_i = \{j|d_{ij} \leq S\}$, the set of facilities that can suitably cover demand i

$R_j = \{i|d_{ij} \leq S\}$, the set of demand that is within the coverage standard of facility j

a_i = amount of demand in area i

c_j = capacity of facility j

p = number of facilities to site

Decision variables:

$$X_j = \begin{cases} 1 & \text{if facility } j \text{ is sited} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{ij} = \begin{cases} 1 & \text{if demand } i \text{ is allocated to facility } j \\ 0 & \text{otherwise} \end{cases}$$

Using this notation, the MCLP can be structured using a location-allocation framework (Church and ReVelle, 1976; Church and Murray, 2018):

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{3.1}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{3.2}$$

$$Y_{ij} \leq X_j \quad \forall i \in I, j \in N_i \tag{3.3}$$

$$\Sigma_{j \in J} X_j = p \tag{3.4}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{3.5}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{3.6}$$

The objective (3.1) maximizes the total demand served. Constraints (3.2) require that each demand is allocated to at most one facility within the service coverage standard. Constraints (3.3) stipulate that a demand can only be allocated to a sited facility. These are often referred to as Balinski constraints, recognized as having integer-friendly solution properties (Church and Roberts, 1983; ReVelle, 1993; Gerrard, 1995; Church and Murray, 2009). Constraint (3.4) requires p facilities to be sited. Constraints (3.5) and (3.6) impose binary restrictions on location and allocation decision variables, respectively. This formulation contains $\Sigma_{i \in I} |N_i| + |J|$ decision variables and $2\Sigma_{i \in I} |N_i| + |I| + |J| + 1$ constraints, where $|\cdot|$ indicates the number of members in the associated set.

The introduction of allocation variables makes it possible to explicitly track facility workloads. Denote W_j the workload of facility j, with $W_j = \Sigma_i a_i Y_{ij}$. Accordingly, there are no restrictions on workloads in the MCLP as no total service allocation limits are imposed in Constraints (3.3). A potential consequence is that imbalanced facility workloads can result. That is, W_j can vary significantly across j in cases where $X_j = 1$. Therefore, consideration of facility workload balance in the context of maximal coverage

is critical in some application contexts, as noted in Current and Storbeck (1988); Pirkul and Schilling (1991); Haghani (1996); Xu et al. (2020), among others.

3.3.2 Workload Variation Measure

In order to consider workload balance, this section reviews quantitative measures of workload variation. Incorporation in the above modeling framework is then considered. As noted previously, Marsh and Schilling (1994) reviewed some twenty different approaches to measure facility equity, such as range, variance, mean absolute deviation and others. Five approaches are specifically noted here to characterize facility workload variation (Table 3.1). These five are used because they reflect the fundamental intent of equity and have been considered in various ways in location modeling literature. Additionally, these approaches can be incorporated in a manner that retains linearity, as is true for the MCLP.

The first measure in Table 3.1, total pairwise absolute workload difference, compares the workload of each two sited facilities, summing the absolute difference (see Church and Murray 1993). This measure explicitly characterizes any workload variation in the system. Other measures, such as those reviewed in Marsh and Schilling (1994) as well as others, can be viewed as approximations to the total pairwise measure for workload balancing. The total pairwise measure can be expanded as follows, given $W_j = \Sigma_i a_i Y_{ij}$

Table 3.1 Different measures to account for facility workload variation

Measures	Mathematical Expression	Reference	Linearity
Total pairwise absolute workload difference	$\Sigma_{j\in J^*}\Sigma_{j'\in J^*}\lvert W_j - W_{j'}\rvert$	Church and Murray (1993)	Piecewise linear
Total mean absolute workload deviation	$\Sigma_{j\in J^*}\lvert W_j - \overline{W}\rvert$	Zhu and McKnew (1993)	
Maximum mean absolute workload deviation	$\max_{j\in J^*}\lvert W_j - \overline{W}\rvert$	Garfinkel and Nemhauser (1970)	
Workload range	$\max_{j\in J^*} W_j - \min_{j\in J^*} W_j$	Marín (2011); Weaver and Church (1981); Davoodi (2019); Daskin and Tucker (2018)	Linear
Maximum workload	$\max_{j\in J^*} W_j$	Berman et al. (2009); Kim and Kim (2010); Davoodi (2019)	

Note: $W_j = \Sigma_{j\in J^*} a_i Y_{ij}$ here for the formulated MCLP. The notation W_j is used for simplicity of mathematical expressions.

and assuming a solution of $J^* = \{j_1, j_2, \cdots, j_p\}$:

$$\begin{aligned}
&\Sigma_{j\in J^*}\Sigma_{j'\in J^*}|W_j - W_{j'}| \\
&= \Sigma_{j\in J^*}\Sigma_{j'\in J^*}|\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}| \\
&= \Sigma_{j\in J^*}(|\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_1}| + |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_2}| + \cdots + \\
&\quad |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij}| + \cdots + |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_p}|)
\end{aligned} \tag{3.7}$$

The second measure in Table 3.1, total mean absolute deviation, sums each workload deviation from the system-wide average facility workload (see Zhu and McKnew 1993). This measure can be expanded as follows:

$$\begin{aligned}
&\Sigma_{j\in J^*}|W_j - \bar{W}| \\
&= \Sigma_{j\in J^*}|\Sigma_i a_i Y_{ij} - \frac{1}{p}\Sigma_{j\in J^*}\Sigma_i a_i Y_{ij}| \\
&= \frac{1}{p}\Sigma_{j\in J^*}|(\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_1}) + (\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_2}) + \cdots + \\
&\quad (\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij}) + \cdots + (\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_p})|
\end{aligned} \tag{3.8}$$

While the total mean absolute deviation does compare workload differences, it is done with respect to the mean workload. Some workload differences are positive and others are negative, so they will effectively cancel some differences. Thus, workload variation using total mean absolute deviation (3.8) overlooks the inherent differences that can occur, making it an approximation to the total pairwise absolute approach, (3.7). In addition, due to the triangle inequity, $|\alpha + \beta| \leq |\alpha| + |\beta|$ for any real numbers α, β, (3.8) is always less than or equal to (3.7) divided by p. That is, the total mean absolute deviation is always less than or equal to the total pairwise absolute workload difference

divided by p. The equality holds if and only if $\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_1}$, $\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_2}$, $\cdots$, $\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_p}$ are all non-negative or non-positive for all $j \in J^*$. This condition means that the workload of any sited facility j, W_j, needs to be the maximum or minimum workload, which only happens when $p = 2$ or workloads of all sited facilities are all equal.

The third measure in Table 3.1, maximum mean absolute deviation, tracks the maximum deviation of a facility workload from the average value (Garfinkel and Nemhauser, 1970). This measure is formulated as:

$$\begin{aligned} &\max_{j \in J^*} |W_j - \bar{W}| \\ &= \max_{j \in J^*} |\Sigma_i a_i Y_{ij} - \frac{1}{p} \Sigma_{j \in J^*} \Sigma_i a_i Y_{ij}| \end{aligned} \tag{3.9}$$

Worth noting is that (3.9) is a portion of the total mean absolute deviation, (3.8). Interestingly, these three measures, (3.7)-(3.9), can be considered piecewise linear. While they are technically non-linear with respect to decision variables defining W_j, simple linear transformations are possible.

The last two measures in Table 3.1, workload range and maximum workload, are linear functions with respect to decision variables defining W_j, making them straightforward to include in an extension of the MCLP. The workload range compares the difference between the maximum and minimum workloads of sited facilities (e.g., Daskin and Tucker

2018; Marín 2011). This measure is formulated as:

$$\begin{aligned} &\max_{j\in J^*} W_j - \min_{j\in J^*} W_j \\ &= \max_{j\in J^*} \Sigma_i a_i Y_{ij} - \min_{j\in J^*} \Sigma_i a_i Y_{ij} \\ &= \max_{j\in J^*} \max_{j'\in J^*} \{|\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|\} \end{aligned} \tag{3.10}$$

Thus, workload range (3.10) can be viewed as comparing sited facility workload pairs, but tracking only the maximum difference. Accordingly, it is an approximation of the total pairwise measure, (3.7). It is mathematically equivalent to the total pairwise measure only when p is 2.

The maximum workload measure (see Berman et al. 2009; Davoodi 2019) tracks only the maximum. This measure is formulated as:

$$\begin{aligned} &\max_{j\in J^*} W_j \\ &= \max_{j\in J^*} \Sigma_i a_i Y_{ij} \end{aligned} \tag{3.11}$$

Thus, maximum workload (3.11) is a portion of the workload range, (3.10), because it only accounts for the largest workload.

In summary, these five measures, (3.7)-(3.11), are not mathematically equivalent. The total pairwise absolute difference measure tracks workload variation most explicitly, with the other four measures serving as approximations. Therefore, an important research question is whether such proxies are effective and meaningful. Accordingly, it is essential to evaluate and compare how these workload variation measures behave when

incorporated in models, and in particular when utilized through the extension of the MCLP.

3.3.3 MCLP Extensions to Balance Workloads

It is possible to use the various equity oriented measures in Table 3.1, and detailed above, in some manner to extend the MCLP. This may not be straightforward in all cases. The primary approach taken here involves the use of multi-objectives, where the added objective incorporates the intent to minimize workload variability using one of the five equity metrics given above. The choice of using multi-objectives over including the equity measures in a constraint is intentional, enabling a user to derive the complete Pareto optimal frontier when considering both coverage and workload variation objectives simultaneously. Along with the objective are additional constraints needed to relate associated decision variables. The extension using the most explicit variation measure, the total pairwise workload difference, is formulated followed by extensions based on the other four approximation measures.

WBMCLP-TotPairDiff

The first model extension considers workload difference between each pair of sited facilities, seeking to minimize variability. Recall the total pairwise difference in (3.7). This compares sited facility workloads assuming they are already known. But W_j is unknown, a byproduct of model solution to identify workload assignments. Thus, (3.7) cannot be

readily used in a model formulation. To deal with this, the absolute workload difference between any two potentially sited facilities j and $j^{'}$ is structured using additional variables, $D_{jj'}$. Consider three situations regarding the siting of two facilities j and $j^{'}$: 1) both are sited, i.e. $X_j = X_{j'} = 1$; 2) both are not sited, i.e. $X_j = X_{j'} = 0$; and, 3) one is sited and the other not, i.e., $X_j = 1$ and $X_{j'} = 0$ (or vice versa). Constraints were introduced in Church and Murray (1993) to track these situations in a spatial optimization model applied to school districting. Consider the first situation where $W_j \geq W_{j'}$. The smallest value of $D_{jj'}$ satisfying the following accounts for the actual variability in workloads:

$$\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'} \leq D_{jj'} \tag{3.12}$$

Since $W_j \geq W_{j'}$, (3.12) is mathematically equivalent to $D_{jj'} \geq |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|$, making $D_{jj'}$ at least the absolute workload difference between j and $j^{'}$. This necessitates a second objective to ensure that the minimum value of $D_{jj'}$ is obtained in order to accurately account for workload difference, as reflected in (3.7). A workload balancing model that minimizes the total pairwise workload difference (WBMCLP-TotPairDiff) through extension of the MCLP is formulated as follows:

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{3.13}$$

$$\text{Minimize} \quad \Sigma_j \Sigma_{j' > j} D_{jj'} \tag{3.14}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{3.15}$$

$$Y_{ij} \leq X_j \quad \forall i \in I, j \in N_i \tag{3.16}$$

$$\Sigma_{j\in J} X_j = p \tag{3.17}$$

$$\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'} - M(1 - X_{j'}) \leq D_{jj'} \quad \forall j, j^{'} \in J \& j^{'} > j \tag{3.18}$$

$$\Sigma_i a_i Y_{ij'} - \Sigma_i a_i Y_{ij} - M(1 - X_j) \leq D_{jj'} \quad \forall j, j^{'} \in J \& j^{'} > j \tag{3.19}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{3.20}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{3.21}$$

$$D_{jj'} \geq 0 \quad \forall j, j^{'} \in J \& j^{'} > j \tag{3.22}$$

Objective (3.13) maximizes the total demand allocated. Objective (3.14) minimizes the total pairwise absolute workload difference of sited facilities. Constraints (3.15)-(3.17), (3.20) and (3.21) are those associated with the MCLP. Note this dissertation considers only single source problems where no fractional allocation is allowed. Assume that M is a very large positive number[1], Constraints (3.18) and (3.19) track the absolute workload difference between facilities j and $j^{'}$ in the three noted situations. In the case that both are sited, then Constraints (3.18) and (3.19) become $\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'} \leq D_{jj'}$ and $\Sigma_i a_i Y_{ij'} - \Sigma_i a_i Y_{ij} \leq D_{jj'}$, thus $D_{jj'} \geq |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|$. For the situation that both are not sited, then Constraints (3.18) and (3.19) become $-M \leq D_{jj'}$. Thus, $D_{jj'}$ is not required to be greater than zero since M is a very large number. Objective (3.14) ensures that $D_{jj'}$ will be zero in value. In the case that only one is sited (e.g., j is sited and $j^{'}$ is not), then Constraints (3.18) and (3.19) would become $\Sigma_i a_i Y_{ij} - M \leq D_{jj'}$ and

[1]In practice, M should be as small as possible for computational efficiency. Here, M is set to $\Sigma_{i\in R_j} a_i$ and $\Sigma_{i\in R_{j'}} a_i$ respectively in Constraints (3.18) and (3.19).

$-\Sigma_i a_i Y_{ij} \leq D_{jj'}$. Similarly, $D_{jj'}$ is not required to be greater than zero since M is a very large positive number. In addition, $D_{jj'}$ is non-negative by Constraints (3.22). So $D_{jj'} \geq |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|$ for pairs of sited facilities and $D_{jj'} \geq 0$ for other paired outcomes. Since $\Sigma_j \Sigma_{j'>j} D_{jj'}$ is minimized by objective (3.14), $D_{jj'}$ will seek to be the smallest value possible, which is the pairwise absolute difference $|\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|$ or 0. This model includes $\Sigma_{i \in I}|N_i| + \frac{|J|(|J|+1)}{2}$ decision variables and $\frac{3}{2}|J|^2 + 2\Sigma_{i \in I}|N_i| + |I| - \frac{1}{2}|J| + 1$ constraints, which is significantly larger in size than the MCLP.

WBMCLP-TotMeanDiff

Approximate approaches to WBMCLP-TotPairDiff, (3.13)-(3.22), are possible. One alternative is the minimization of the total mean absolute workload deviation. This would relate sited facility workloads to the average workload. The measure of total mean absolute deviation in (3.8) cannot be directly added to a model because it assumes facility workloads are known. To incorporate (3.8) into the model, D_j is introduced to represent the deviation of facility j's workload from the average workload. Consider an example where facility j is sited and W_j is greater than the average workload $\frac{\Sigma_j \Sigma_i a_i Y_{ij}}{p}$. The minimum value of D_j in the following condition is the mean absolute workload deviation of facility j.

$$\Sigma_i a_i Y_{ij} - \Sigma_i \Sigma_j a_i Y_{ij}/p \leq D_j \tag{3.23}$$

To make D_j take the minimum value, a second objective is necessary. An extension of the MCLP that minimizes the total mean absolute workload deviation (WBMCLP-

TotMeanDiff) is formulated as follows:

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{3.24}$$

$$\text{Minimize} \quad \Sigma_{j \in J} D_j \tag{3.25}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{3.26}$$

$$Y_{ij} \leq X_j \quad \forall i \in I, j \in N_i \tag{3.27}$$

$$\Sigma_{j \in J} X_j = p \tag{3.28}$$

$$\Sigma_i a_i Y_{ij} - \Sigma_i \Sigma_j a_i Y_{ij}/p \leq D_j \quad \forall j \in J \tag{3.29}$$

$$\Sigma_i \Sigma_j a_i Y_{ij}/p - \Sigma_i a_i Y_{ij} - M(1 - X_j) \leq D_j \quad \forall j \in J \tag{3.30}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{3.31}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{3.32}$$

$$D_j \geq 0 \quad \forall j \in J \tag{3.33}$$

Objective (3.24) maximizes the total demand allocated and objective (3.25) minimizes the total mean absolute workload deviation. Constraints (3.26)-(3.28), (3.31) and (3.32) reflect those in the MCLP. Constraints (3.29) and (3.30) track the mean absolute workload deviation considering two situations for facility j: 1) sited, i.e. $X_j = 1$; and, 2) not sited, i.e. $X_j = 0$. If facility j is sited, then Constraints (3.29) and (3.30) become $\Sigma_i a_i Y_{ij} - \Sigma_i \Sigma_j a_i Y_{ij}/p \leq D_j$ and $\Sigma_i \Sigma_j a_i Y_{ij}/p - \Sigma_i a_i Y_{ij} \leq D_j$. This means $D_j \geq |\Sigma_i a_i Y_{ij} - \Sigma_i \Sigma_j a_i Y_{ij}/p|$ for a sited facility j. If facility j is not sited, then Constraints (3.29) and (3.30) become $-\Sigma_i \Sigma_j a_i Y_{ij}/p \leq D_j$ and $\Sigma_i \Sigma_j a_i Y_{ij}/p - M \leq D_j$, not

imposing any effective constraint on D_j because M is a very large number[2]. Thus, $D_j \geq 0$ results in the case where a facility is not sited given (3.33). Since $\Sigma_{j \in J} D_j$ is minimized by objective (3.25), any D_j will seek to be its lower bound, $|\Sigma_i a_i Y_{ij} - \Sigma_i \Sigma_j a_i Y_{ij}/p|$ or 0, depending on whether facility j is sited or not. Thus, the total mean absolute deviation formulated in (3.8) is accurately reflected in (3.24)-(3.33). This model has $\Sigma_{i \in I}|N_i| + 2|J|$ decision variables and $2\Sigma_{i \in I}|N_i| + |I| + 4|J| + 1$ constraints, smaller in size compared to the WBMCLP-TotPairDiff.

WBMCLP-MaxMeanDiff

Another approximation approach is to minimize the maximum mean absolute deviation, (3.9). Since the maximum mean workload deviation is minimized, accounting for other deviation is implicit. Facility workloads would tend to approach the average workload, and are therefore balanced in theory. Unfortunately, the maximum mean absolute deviation (3.9) cannot be readily incorporated into a model formulation. To address this, a decision variable D is introduced to track the maximum absolute deviation of sited facility workload from the average. For example, assume that a facility j is sited and its workload is greater than the average, then the following inequality restricts D to be at least the mean absolute workload deviation:

$$\Sigma_i a_i Y_{ij} - \Sigma_i \Sigma_j a_i Y_{ij}/p \leq D \tag{3.34}$$

[2] M is set to $\Sigma_{i \in I} a_i/p$ in Constraints (3.30) and (3.41).

An objective is therefore needed to ensure D is the minimum value, making it the mean absolute workload deviation of facility j. A workload balancing model that minimizes the maximum mean absolute workload deviation in the context of the MCLP (WBMCLP-MaxMeanDiff) is formulated as follows:

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{3.35}$$

$$\text{Minimize} \quad D \tag{3.36}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{3.37}$$

$$Y_{ij} \leq X_j \quad \forall i \in I, j \in N_i \tag{3.38}$$

$$\Sigma_{j \in J} X_j = p \tag{3.39}$$

$$\Sigma_i a_i Y_{ij} - \Sigma_i \Sigma_j a_i Y_{ij}/p \leq D \quad \forall j \in J \tag{3.40}$$

$$\Sigma_i \Sigma_j a_i Y_{ij}/p - \Sigma_i a_i Y_{ij} - M(1 - X_j) \leq D \quad \forall j \in J \tag{3.41}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{3.42}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{3.43}$$

$$D \geq 0 \tag{3.44}$$

The first objective (3.35) maximizes the total demand allocated and the second objective (3.36) minimizes the maximum mean absolute deviation. Constraints (3.37)-(3.39), (3.42) and (3.43) reflect the MCLP. Constraints (3.40) and (3.41) track the maximum mean absolute deviation, with D considering situations when facility j is sited and not

sited. For the case that facility j is sited, $D \geq |\Sigma_i a_i Y_{ij} - \Sigma_i \Sigma_j a_i Y_{ij}/p|$ in Constraints (3.40) and (3.41). For the case that facility j is not sited, $D \geq 0$ due to Constraint (3.44). Collectively, $D \geq \max_j |\Sigma_i a_i Y_{ij} - \Sigma_i \Sigma_j a_i Y_{ij}/p|$. Combined with objective (3.36), D seeks to be the smallest value possible, which is exactly the maximum mean absolute deviation of sited facilities (3.9). This model has $\Sigma_{i \in I}|N_i| + |J| + 1$ decision variables and $2\Sigma_{i \in I}|N_i| + |I| + 3|J| + 2$ constraints, significantly smaller in size compared to the WBMCLP-TotPairDiff.

WBMCLP-Range

Another approximation to the WBMCLP-TotPairDiff is comparing maximum and minimum workloads. By minimizing the range, (3.10), facility workloads are driven to less variability. To incorporate the range into a model, decision variables U and L are introduced to track the maximum facility workload and the minimum workload of sited facilities, respectively. A workload balancing model that extends the MCLP by minimizing the workload range (WBMCLP-Range) is formulated as follows:

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{3.45}$$

$$\text{Minimize} \quad U - L \tag{3.46}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{3.47}$$

$$Y_{ij} \leq X_j \quad \forall i \in I, j \in N_i \tag{3.48}$$

$$\Sigma_{j \in J} X_j = p \tag{3.49}$$

$$\Sigma_i a_i Y_{ij} \leq U \quad \forall j \in J \tag{3.50}$$

$$\Sigma_i a_i Y_{ij} + M(1 - X_j) \geq L \quad \forall j \in J \tag{3.51}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{3.52}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{3.53}$$

$$L \geq 0 \tag{3.54}$$

Objective (3.45) maximizes the total demand allocated and objective (3.46) minimizes the workload range of sited facilities. Constraints (3.47)-(3.49), (3.52) and (3.53) reflect the MCLP. Constraints (3.50) ensure U is at least the maximum facility workload. That is $U \geq max_j W_j$, which can be re-written as $U \geq max_{j \in J^*} W_j$ because $W_j = 0$ for $j \notin J^*$. Constraints (3.51) consider two situations, whether facility j is sited or is not sited. If facility j is sited, then $0 \leq L \leq \Sigma_i a_i Y_{ij}$. If facility j is not sited, then $0 \leq L \leq M$ where M is a very large number[3]. Thus, $0 \leq L \leq \min_{j \in J^*} W_j$. Combined with objective (3.46), U will be its lower bound, $max_{j \in J^*} W_j$, and L will be its upper bound, $min_{j \in J^*} W_j$. Thus, the minimization of $U - L$ is equivalent to the minimization of the workload range (3.10). This model contains $\Sigma_{i \in I} |N_i| + |J| + 2$ decision variables and $2\Sigma_{i \in I} |N_i| + |I| + 3|J| + 2$ constraints, significantly smaller in size compared to the WBMCLP-TotPairDiff.

WBMCLP-Max

Related to WBMCLP-Range is a focus on minimizing the maximum workload, (3.11). If the maximum workload is minimized, other sited facility workloads are controlled

[3] M is set to $\max_j \Sigma_{i \in R_j} a_i$ in Constraints (3.51).

implicitly. Since the maximal covering objective would reward sited facility to cover as much as possible, workloads would tend to balance to some extent, in theory. An extension of the MCLP that minimizes the maximum workload (WBMCLP-Max) is formulated as follows:

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{3.55}$$

$$\text{Minimize} \quad U \tag{3.56}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{3.57}$$

$$Y_{ij} \leq X_j \quad \forall i \in I, j \in N_i \tag{3.58}$$

$$\Sigma_{j \in J} X_j = p \tag{3.59}$$

$$\Sigma_i a_i Y_{ij} \leq U \quad \forall j \in J \tag{3.60}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{3.61}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{3.62}$$

The first objective (3.55) maximizes the total demand allocated. The second objective (3.56) minimizes the maximum workload. Constraints (3.57)-(3.59), (3.61) and (3.62) address MCLP considerations. Constraints (3.60) restrict U to be greater than or equal to any observed facility workload, i.e. $U \geq max_j W_j$. Since U is minimized in objective (3.56), it will be the smallest value possible, which is exactly the maximum workload (3.11). This model has $\Sigma_{i \in I} |N_i| + |J| + 1$ decision variables and $2\Sigma_{i \in I} |N_i| + |I| + 2|J| + 1$ constraints, significantly smaller in size compared to the WBMCLP-TotPairDiff.

3.3.4 Forcing Assignment Constraints

A byproduct of the workload balancing objectives reflected in the above models is that some demand may simply be unassigned even though it is within the service coverage standard from a sited facility. In a sense this is akin to a strategic denial of service in order to better balance workloads. However, such avoidance of allocation is highly problematic in practice since it is obvious that response within the standard is both possible and likely, yet the model is not reflecting this due to the desire to balance workloads. Operationally speaking, it is difficult to imagine a situation where a public agency or private company of any sort would actually keep people from accessing service when response is possible. Later in the chaper we will return to this issue, demonstrating that it happens and discussing the implications. To address this situation in the context of workload balancing, it is possible to structure and impose additional constraints that force service assignment when demand is actually covered. Consider an additional binary decision variable C_i, where it is 1 if demand i is within the service coverage standard of a sited facility and 0 otherwise. The following constraints can be included in each of the above models to ensure that demand is allocated/served if it is within the service coverage standard from a sited facility:

$$\Sigma_{j \in N_i} X_j \leq pC_i \quad \forall i \in I \tag{3.63}$$

$$C_i \leq \Sigma_{j \in N_i} Y_{ij} \quad \forall i \in I \tag{3.64}$$

$$C_i = \{0, 1\} \quad \forall i \in I \tag{3.65}$$

Constraints (3.63) require C_i to be 1 if at least one facility that can suitably cover demand i is sited. Constraints (3.64) associate the coverage with the forced allocation: if demand i can be suitably covered by a sited facility, i.e. $C_i = 1$, then demand i must be allocated because $\Sigma_{j \in N_i} Y_{ij} \geq 1$. Finally, Constraints (3.65) give binary integer restrictions. Constraints (3.63)-(3.65) can be added to proposed workload balancing models if they are suitable for a specific application context. Each of the workload balancing models detailed above impose these constraints to force assignment when demand is within the service coverage standard, unless indicated otherwise.

3.4 Assessment

A critical question is whether the theoretical measures of variability in Table 3.1 that are formalized in the derived models effectively address workload balance in the context of maximal covering. This is important because the MCLP is known to produce facility workloads that can vary considerably, as noted by Current and Storbeck (1988); Pirkul and Schilling (1991); Haghani (1996); Xu et al. (2020). This section proposes an evaluation approach to comparatively assess capabilities of proposed approaches to address equity in maximal covering. The WBMCLP-TotPairDiff is treated as the benchmark model providing optimal solutions for evaluation. The reason for this is that the total pairwise measure characterizes workload variation explicitly and accurately, as discussed

in Section 3.3.2. The other four approximation approaches are evaluated relative to WBMCLP-TotPairDiff.

Each proposed workload balancing model formulated in the previous section is a discrete bi-objective problem, so the Pareto optimal set defines the best tradeoff between objectives. This chapter uses the constraint method (Cohon, 1978) to derive the complete Pareto optimal set. The task is then to compare obtained Pareto optimal sets for each approximation model with the Pareto optimal set of the benchmark model. This is not necessarily an easy task because solution sets are obtained in different objective spaces. In addition, every solution is associated with two objective attributes, total demand covered and the workload variation measure. To deal with these difficulties, the Pareto optimal sets of approximation models are mapped to the objective space of the WBMCLP-TotPairDiff, which is denoted as $Z = \mathbb{R}^2 : Z_1 \times Z_2$ where Z_1 and Z_2 represent the total demand amount allocated and the total pairwise absolute workload difference, respectively. This can be done by computing total pairwise measure values of solutions for non-benchmark models. Evaluation is then possible in the objective space Z. Three quantitative measures are defined to help assess a solution set in the bi-objective space: completeness, inferiority and maximum gap.

Definition 1. Let P be the Pareto optimal set in objective space Z, Q be a set of solutions in the same objective space Z, Q^* be the set of Pareto optimal solutions that exists in Q, i.e., $Q^* = P \cap Q$, then completeness is defined by $|Q^*|/|P|$.

Definition 2. Let P be the Pareto optimal set in an objective space Z, Q be a set of solutions also in Z, $Q^{'}$ be the set of solutions in Q that is inferior to at least one solution in P, that is $Q^{'} = \{t : t \prec s | t \in Q, \exists s \in P\}$ where the strict order $\prec$ denotes Pareto dominance and $t \prec s$ means solution s dominates solution t, then inferiority is defined by $|Q^{'}|/|Q|$.

Definition 3. With the notation in Definition 2, if $P \setminus Q \neq \emptyset$, let t be a solution in $P \setminus Q$, $t^{'}$ be the closest solution to t in Q, Z_1^s, Z_2^s be the total demand coverage and the workload variation value of a solution s, ϵ a very small number (e.g., 10e-6) to avoid zero denominator, then the maximum gap is defined by (gap_1, gap_2) where:

$$gap_1 = \max_{t \in P \setminus Q} \frac{|Z_1^t - Z_1^{t^{'}}|}{Z_1^t + \epsilon} * 100\% \tag{3.66}$$

$$gap_2 = \max_{t \in P \setminus Q} \frac{|Z_2^t - Z_2^{t^{'}}|}{Z_2^t + \epsilon} * 100\% \tag{3.67}$$

Definitions 1-3 are proposed to measure similarity/difference between a set of solutions and the Pareto optimal set in the objective space $Z \in \mathbb{R}^2$. Completeness measures how completely Pareto optimal solutions can be obtained in the evaluated solution set. Thus, the larger the completeness, the more Pareto optimal solutions found by the evaluated solution set. Inferiority characterizes how many solutions in the evaluated solution set are inferior solutions. Thus, the lower the inferiority, the smaller the proportion of inferior solutions in the evaluated solution set. If any optimal solution is missed, the maximum gap measures the maximum percentage deviation of the closest solution from the missed

Pareto optimal solution in objective space. So the smaller the maximum gap, the closer the given solution set is to the Pareto optimal set in the objective space.

The evaluation approach is summarized as follows:

Step 1: Derive the complete Pareto optimal set for each workload balancing model in its own objective space. This is done by changing the total demand coverage maximization objective to a constraint that restricts the total demand coverage to be no less than a threshold, and is systematically adjusted and then re-solved as a single-objective problem. The Pareto optimal set of the WBMCLP-TotPairDiff is denoted as P.

Step 2: Select an approximation approach, map its Pareto optimal solutions obtained in its original objective space to the objective space Z of the WBMCLP-TotPairDiff for evaluation, generating a solution set Q in Z. This is done by calculating the total pairwise absolute workload difference associated with each Pareto optimal solution of the evaluated model.

Step 3: Evaluate the obtained solution set (Q) against the Pareto optimal set (P) in Z by computing completeness, inferiority and maximum gaps.

Step 4: Repeat Steps 2-3 until every approximation model is evaluated.

3.5 Application Results

Empirical assessment was carried out utilizing two planning applications to evaluate and compare the five detailed workload balancing approaches in terms of solution quality and

computational effort. All workload balancing models with forcing assignment constraints, (3.63)-(3.65), are solved to obtain the entire set of Pareto optimal solutions. This is accomplished using the constraint method, as noted previously. An exact solver, Gurobi (version 9.0.1), is relied on for solution, employing the simplex algorithm combined with branch-and-bound. The solution qualities of the various models are evaluated against the benchmark, WBMCLP-TotPairDiff. Computational time is reported for a MacBook Pro (2.9 GHz Intel Core i5 processor and 8 GB memory).

3.5.1 San Jose Study

The first planning application involves postal delivery to representative points of 32 ZIP Code Tabulation Areas in San Jose, CA, as reported in Xu et al. (2020). The centroids of the areas are used to represent demand locations and potential postal service facilities. Travel distances between demand and potential facilities are computed based on the street network for San Jose. The service coverage standard (S) is set to a network distance of 3 miles. The number of postal service facilities to site (p) ranges from 2 to 17.

The complete set of Pareto optimal solutions is derived for each model. Figure 3.2 illustrates the trade-off between the total demand coverage and the pairwise absolute workload difference, WBMCLP-TotPairDiff, when p is 4. The horizontal axis shows demand coverage and the vertical axis indicates total pairwise absolute workload difference, both in thousands of people. There are 9 Pareto optimal solutions in this case. The maximum coverage possible by siting four facilities is 619 (in units of thousands

of people). The associated facility workloads are highly imbalanced when covering 619, with a large total pairwise absolute workload difference of 389. Figure 3.3(a) gives the spatial configuration of this solution where two facilities serves 195 and 196 and the other two serves 83 and 145. When $p = 4$, it is found that significantly more balanced facility workloads could be achieved by covering 10% less demand (Figure 3.2). The workload variation (in total pairwise absolute workload differences) can be reduced from 389 to 59 when serving 561. Figure 3.3(b) presents the spatial configuration of this more balanced solution where workloads of four sited facilities are 133, 137, 139 and 152. It can be seen that sited facilities are shifted towards the area with less dense demand. The lowest total pairwise measure value possible with four facilities is 9 when serving 421.

Figure 3.4 presents Pareto optimal solutions of the four approximate (non-benchmark) models when p is 4. There are 10 Pareto optimal solutions for the WBMCLP-TotMeanDiff. The total demand covered ranges from 421 to 619 and total mean absolute workload deviation ranges from 3.5 to 163. Most sited facility configurations are the same as those identified using the WBMCLP-TotPairDiff except one. When using the WBMCLP-MaxMeanDiff, there are 11 Pareto optimal solutions. The maximum absolute workload deviation ranges from 1.75 to 71.75. Some difference in facility configurations is witnessed compared to those of the WBMCLP-TotPairDiff. There are 9 Pareto optimal solutions for the WBMCLP-Range, whose facility configurations are mostly the same with those of the WBMCLP-TotPairDiff. The workload range varies from 3 to 113. When the WBMCLP-Max is applied, 39 Pareto optimal solutions are found, with the total demand

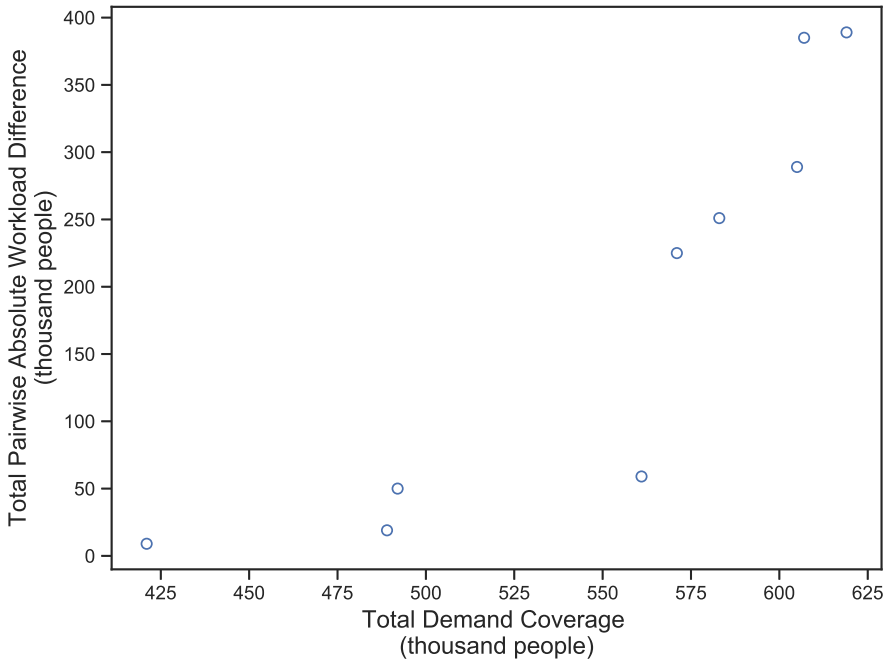

Figure 3.2 Pareto optimal solutions for WBMCLP-TotPairDiff (San Jose, $p = 4$)

covered ranging from 19 to 619 and the maximum workload ranging from 10 to 196. The facility configurations identified are more varied than those found using the other models.

The above comparison is not straightforward (Figures 3.2 and 3.4) because Pareto optimal solutions are shown in different objective spaces. Thus, solutions of approximate (non-benchmark) models are mapped to the objective space of the WBMCLP-TotPairDiff for comparison (Figure 3.5). This is done by calculating the total pairwise measure of each solution, then identifying and keeping non-dominated ones. Accordingly, 8 out of 9 solutions in the Pareto optimal set for the WBMCLP-TotPairDiff are found using the WBMCLP-TotMeanDiff. One solution given by the WBMCLP-TotMeanDiff is inferior.

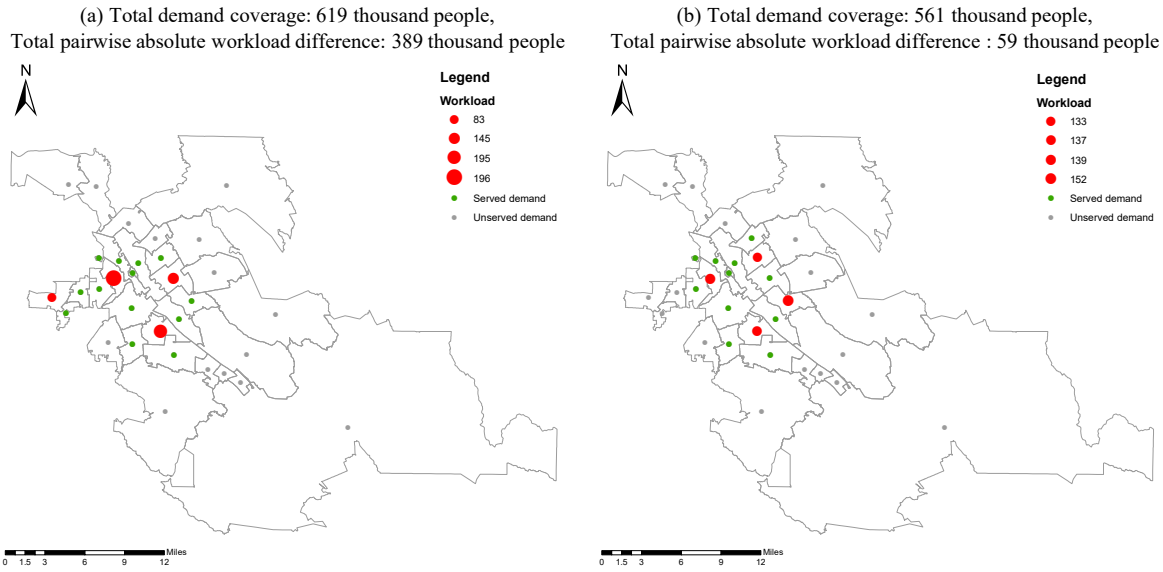

Figure 3.3 Two Pareto optimal solution configurations for WBMCLP-TotPairDiff (San Jose, $p = 4$)

In other words, the completeness is 88.9% and the inferiority is 11.1% according to Definitions 1 and 2. The missed optimal solution has objective values of $Z_1 = 561, Z_2 = 59$ and its closest WBMCLP-TotMeanDiff solution is $Z_1' = 561, Z_2' = 63$. Thus, the maximum gap is $gap_1 = 0, gap_2 = 6.8\%$ according to Definition 3. The obtained completeness, inferiority, and maximum gap values show that the WBMCLP-TotMeanDiff is able to identify a solution set close to the actual optimal set in this case. Similarly, the other models can be compared. The WBMCLP-Range also finds 8 solutions in the WBMCLP-TotPairDiff Pareto optimal set. The WBMCLP-MaxMeanDiff and WBMCLP-Max both find 5 Pareto optimal solutions.

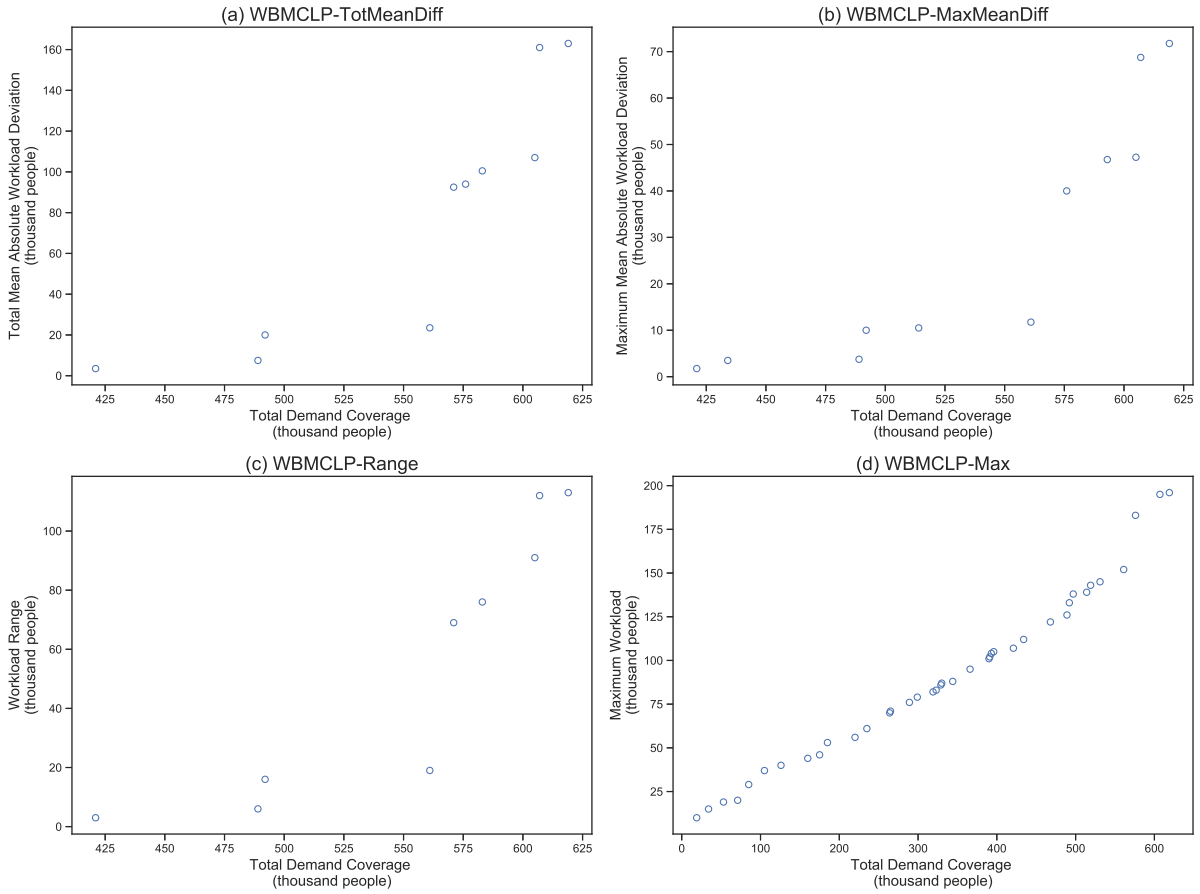

Figure 3.4 Pareto optimal solutions of non-benchmark workload balancing models (San Jose, $p = 4$)

The above analysis demonstrates that for $p = 4$, Pareto optimality is only approximated using non-benchmark workload balancing approaches. Problem instances for other values of p are analyzed similarly and summarized in Table 3.2. First, the four non-benchmark models all have relatively small completeness values and large inferiority and maximum gap values, especially with larger p. When $p \geq 6$, more than 50% of the Pareto optimal solutions are often missed. Also, in most cases, more than 50% of solu-

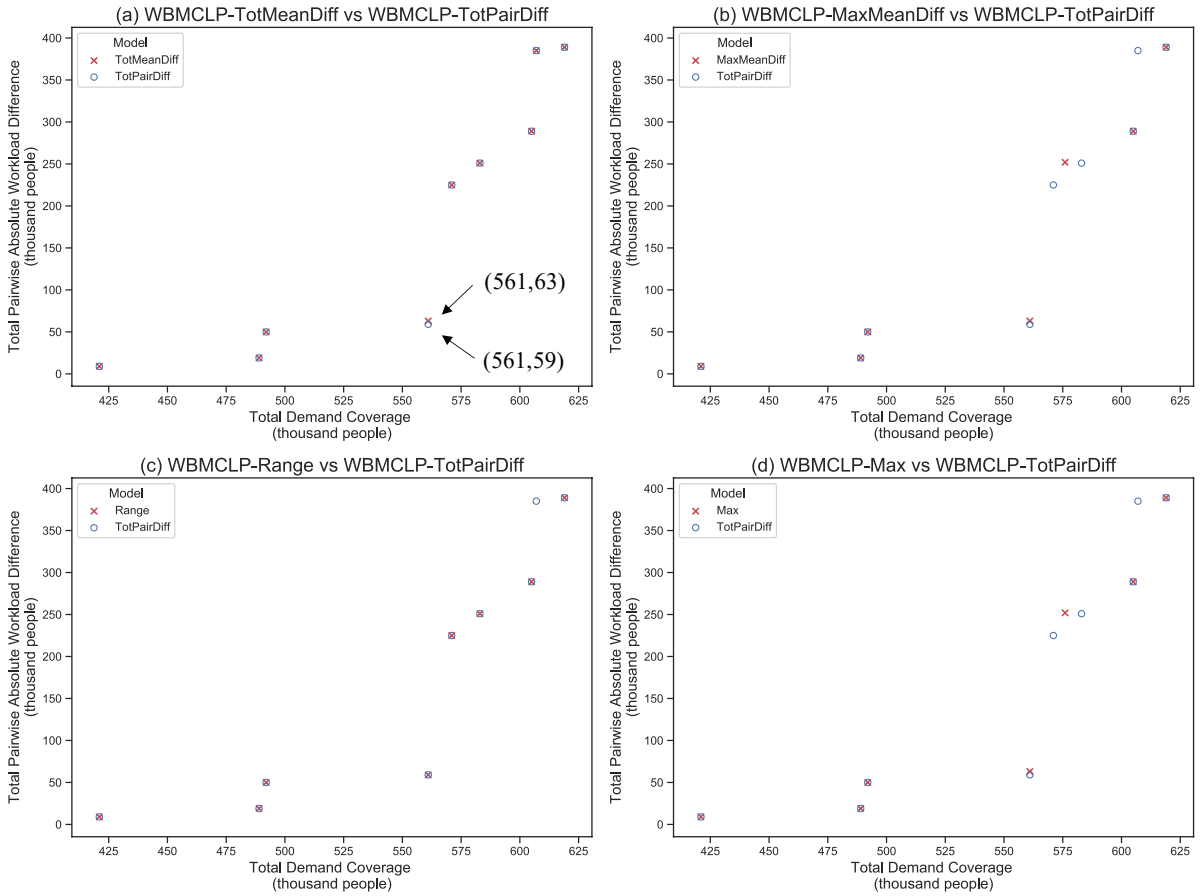

Figure 3.5 Pareto optimal solutions of approximate workload balancing models mapped to the objective space of WBMCLP-TotPairDiff (San Jose, $p = 4$)

tions are actually inferior, with the maximum gap reaching 5.4% in total coverage and 15.4% in workload variation. For example, when p= 17 using the WBMCLP-Max, only 4.5% of Pareto optimal solutions are identified. Among its solutions, 93.3% are actually inferior. These results highlight that the WBMCLP-TotPairDiff approach is not approximated well by the other four models. Therefore, if a model does not address workload

balance explicitly, solution optimality is likely to be questionable and deviation could be large. In addition, the WBMCLP-TotMeanDiff and WBMCLP-Range models have relatively larger completeness, and smaller inferiority and maximum gap compared to WBMCLP-MaxMeanDiff and WBMCLP-Max. This further confirms that more model simplification leads to larger optimality deviations. There is a trend that completeness decreases, and inferiority and maximum gaps increase as p becomes larger. This implies that for problems that are more difficult to solve due to workload balance, the optimality deviation of a non-benchmark approach becomes larger.

Computational times for obtaining the Pareto optimal set for each model using an exact solver is also compared (Table 3.2). First, it takes the exact solver considerable time to solve the WBMCLP-TotPairDiff when $p \geq 8$. For example, when $p = 17$, some 7879.519 seconds ($\sim$2.2 hours) are needed to find the complete Pareto optimal set. Second, the WBMCLP-TotPairDiff requires significantly more time compared to non-benchmark approaches. Clearly the larger the model in terms of decision variables and constraints, the more time needed for deriving the Pareto optimal set. For example, when $p = 17$, the WBMCLP-TotMeanDiff needs 318 seconds, yet the smallest size model, WBMCLP-Max, only requires 2.818 seconds. Third, in general, all workload balancing models require more time to solve problems as p increases.

Table 3.2 Solution quality and computational time comparison among workload balancing models (San Jose)

p	Model	Completeness (%)	Inferiority (%)	Maximum Gaps(%)	Solution Time (Seconds)
2	**TotPairDiff(2)**	**100**	**0**	**(0.0, 0.0)**	0.503
	TotMeanDiff	100	0	(0.0, 0.0)	0.379
	MaxMeanDiff	100	0	(0.0, 0.0)	0.247
	Range	100	0	(0.0, 0.0)	1.813
	Max	100	0	(0.0, 0.0)	4.098
3	**TotPairDiff(9)**	**100**	**0**	**(0.0, 0.0)**	7.891
	TotMeanDiff	100	0	(0.0, 0.0)	3.223
	MaxMeanDiff	100	0	(0.0, 0.0)	1.835
	Range	100	0	(0.0, 0.0)	5.705
	Max	100	0	(0.0, 0.0)	5.58
4	**TotPairDiff(9)**	**100**	**0**	**(0.0, 0.0)**	7.338
	TotMeanDiff	88.9	11.1	(0.0, 6.8)	3.175
	MaxMeanDiff	55.6	28.6	(1.0, 5.1)	2.415
	Range	88.9	0	(2.0, 1.0)	4.559
	Max	55.6	28.6	(1.0, 5.1)	5.997
5	**TotPairDiff(20)**	**100**	**0**	**(0.0, 0.0)**	31.709
	TotMeanDiff	55	42.1	(1.0, 6.0)	13.837
	MaxMeanDiff	30	50	(3.2, 11.5)	7.542
	Range	75	11.8	(3.6, 6.7)	20.143
	Max	60	36.8	(1.5, 15.6)	7.394
6	**TotPairDiff(19)**	**100**	**0**	**(0.0, 0.0)**	66.893
	TotMeanDiff	73.7	12.5	(3.1, 4.9)	18.077
	MaxMeanDiff	31.6	57.1	(1.2, 15.4)	7.159
	Range	42.1	27.3	(1.8, 15.0)	28.049
	Max	47.4	30.8	(2.5, 6.8)	7.394

Table 3.2 (continued)

p	Model	Completeness (%)	Inferiority (%)	Maximum Gaps(%)	Solution Time (Seconds)
7	**TotPairDiff(13)**	**100**	**0**	**(0.0, 0.0)**	52.748
	TotMeanDiff	61.5	33.3	(0.6, 1.6)	16.876
	MaxMeanDiff	30.8	66.7	(0.9, 10.4)	5.542
	Range	38.5	54.5	(1.3, 10.6)	28.161
	Max	30.8	69.2	(1.4, 9. 8)	5.783
8	**TotPairDiff(16)**	**100**	**0**	**(0.0, 0.0)**	108.634
	TotMeanDiff	37.5	57.1	(0.2, 3.1)	23.663
	MaxMeanDiff	18.8	81.3	(2.0, 14.9)	9.792
	Range	25	69.2	(1.0, 6.7)	23.49
	Max	25	71.4	(0.9, 10.0)	6.52
9	**TotPairDiff(25)**	**100**	**0**	**(0.0, 0.0)**	776.002
	TotMeanDiff	32	60	(0.2, 3.1)	112.536
	MaxMeanDiff	12	86.4	(4.5, 14.3)	16.161
	Range	16	71.4	(4.2, 13.3)	42.432
	Max	20	64.3	(2.2, 5.8)	5.624
10	**TotPairDiff(28)**	**100**	**0**	**(0.0, 0.0)**	1,270.63
	TotMeanDiff	32.1	60.9	(1.2, 3.9)	221.457
	MaxMeanDiff	0	100	(3.5, 13.3)	14.927
	Range	10.7	75	(3.6, 8.9)	47.174
	Max	7.1	88.9	(3.1, 6.8)	5.439
11	**TotPairDiff(21)**	**100**	**0**	**(0.0, 0.0)**	1,422.25
	TotMeanDiff	4.8	94.7	(1.0, 3.8)	216.559
	MaxMeanDiff	4.8	92.9	(3.8, 5.1)	10.613
	Range	19	66.7	(3.9, 5.3)	74.783
	Max	4.8	93.3	(3.6, 3.3)	5.648
12	**TotPairDiff(21)**	**100**	**0**	**(0.0, 0.0)**	2,269.48
	TotMeanDiff	9.5	90.5	(1.3, 2.6)	155.262
	MaxMeanDiff	0	100	(5.4, 3.6)	11.38
	Range	14.3	78.6	(2.6, 3.6)	68.217
	Max	4.8	94.2	(2.6, 3.3)	5.607

Table 3.2 (continued)

p	Model	Completeness (%)	Inferiority (%)	Maximum Gaps(%)	Solution Time (Seconds)
13	**TotPairDiff(22)**	**100**	**0**	**(0.0, 0.0)**	3,190.57
	TotMeanDiff	0	100	(1.2, 2.8)	155.356
	MaxMeanDiff	0	100	(3.9, 5.4)	11.968
	Range	9.1	87.5	(1.6, 2.4)	77.466
	Max	4.5	94.4	(2.1, 2.8)	4.244
14	**TotPairDiff(19)**	**100**	**0**	**(0.0, 0.0)**	2,876.58
	TotMeanDiff	10.5	88.9	(0.8, 3.8)	136.39
	MaxMeanDiff	5.3	94.7	(1.9, 3.3)	9.869
	Range	10.5	83.3	(1.2, 2.7)	75.817
	Max	10.5	85.7	(1.4, 3.7)	4.036
15	**TotPairDiff(20)**	**100**	**0**	**(0.0, 0.0)**	3,374.39
	TotMeanDiff	20	80	(0.9, 3.1)	141.455
	MaxMeanDiff	5	95	(1.8, 4.9)	15.596
	Range	10	86.7	(2.0, 2.7)	89.943
	Max	10	86.7	(0.9, 2.5)	3.394
16	**TotPairDiff(23)**	**100**	**0**	**(0.0, 0.0)**	6,647.06
	TotMeanDiff	30.4	63.2	(1.7, 3.2)	234.961
	MaxMeanDiff	4.3	93.3	(2.1, 4.1)	11.955
	Range	8.7	87.5	(1.1, 2.8)	101.821
	Max	8.7	88.2	(2.0, 2.6)	2.996
17	**TotPairDiff(22)**	**100**	**0**	**(0.0, 0.0)**	7,897.52
	TotMeanDiff	45.5	54.5	(0.4, 2.9)	318.347
	MaxMeanDiff	0	100	(2.0, 3.3)	16.113
	Range	4.5	92.9	(1.9, 3.7)	61.906
	Max	4.5	93.3	(1.6, 3.6)	2.818

3.5.2 Santa Barbara Study

The second case study investigates fire response to 80 block groups in Santa Barbara, CA. The centroids of these areas are used to represent both demand and potential fire stations. The demand is based on population in the area. The local street network is used to construct the transportation network, from which travel distance between demand and potential stations is derived. The service coverage standard is a network distance of 1.5 miles. The number of fire stations to site (p) ranges from 2 to 6.

Table 3.3 summarizes solution quality and computational time for each workload balancing model. When p is 2, there is only one Pareto optimal solution of the WBMCLP-TotPairDiff that the four non-benchmark approaches find. When $p = 3$, WBMCLP-Max and WBMCLP-Range have the same performance as the WBMCLP-TotPairDiff while the WBMCLP-MaxMeanDiff and WBMCLP-TotMeanDiff only find 33.3% of the optimal solutions, as well as a few dominated solutions. When p is 4, the WBMCLP-TotPairDiff has 27 Pareto optimal solutions (Figure 3.6). The maximum coverage possible is 719 with a maximum total pairwise absolute workload difference of 321. The workload variation can be reduced to half of the maximum value if giving up 8% demand coverage. A totally balanced system (with 0 workload variation) can be obtained when the total coverage is 632. The WBMCLP-Range approximates the WBMCLP-TotPairDiff best, finding 88.9% Pareto optimal solutions while it take 94,639.084 seconds (more than one day) to obtain these solutions. When p is 5 or 6, most proportions of found Pareto optimal solutions

reduce to below 50%, indicating the WBMCLP-TotPairDiff is not approximated well. The WBMCLP-Max identifies a large number of solutions, but only approximate the Pareto optimal front of the WBMCLP-to TotPairDiff. The other models generate fewer solutions. For example, when $p = 3$, there are 163 Pareto optimal solutions for the WBMCLP-Max while only around 10 exist for the other models. The large number of solutions also explains why it takes more time to solve the WBMCLP-Max in the Santa Barbara case. In sum, when approximating the benchmark model, optimality deviation is again observed, particularly as p increases. Computational efficiency remains a concern, with some workload balancing models not able confirm optimality after a week of computer time.

Table 3.3 Solution quality and computational time comparison among workload balancing models (Santa Barbara)

p	Model	Completeness (%)	Inferiority (%)	Maximum Gaps(%)	Solution Time (Seconds)
2	**TotPairDiff(1)**	**100**	**0**	**(0.0, 0.0)**	**10.938**
	TotMeanDiff	100	0	(0.0, 0.0)	1.953
	MaxMeanDiff	100	0	(0.0, 0.0)	0.969
	Range	100	0	(0.0, 0.0)	2.026
	Max	100	0	(0.0, 0.0)	1,135.06
3	**TotPairDiff(6)**	**100**	**0**	**(0.0, 0.0)**	**177.043**
	TotMeanDiff	33.3	60	(1.7, 31.4)	52.187
	MaxMeanDiff	33.3	50	(1.7, 33.3)	26.86
	Range	100	0	(0.0, 0.0)	19.505
	Max	100	0	(0.0, 0.0)	2,240.56
4	**TotPairDiff(27)**	**100**	**0**	**(0.0, 0.0)**	**5,195.20**
	TotMeanDiff	7.4	86.7	(3.7, 20.0)	1,313.74
	MaxMeanDiff	7.4	83.3	(6.0, 29.4)	72.625
	Range	88.9	0	(0.1, 0.5)	94,639.08
	Max	81.5	15.4	(0.6, 26.3)	2,103.97
5	**TotPairDiff(28)**	**100**	**0**	**(0.0, 0.0)**	**23,448.22**
	TotMeanDiff	10.7	87	(2.0, 21.1)	6,216.86
	MaxMeanDiff	7.1	88.9	(3.3, 36.4)	274.529
	Range[1]	78.6	0.00	(0.7, 8.3)	45,935.46
	Max	82.1	0	(0.7, 8.3)	6,869.47
6	**TotPairDiff(18)**	**100**	**0**	**(0.0, 0.0)**	**32,194.57**
	TotMeanDiff	44.4	38.5	(1.0, 22.2)	989.137
	MaxMeanDiff	27.8	70.6	(1.0, 30.0)	112.891
	Range[1]	33.3	33.3	(1.5, 41.2)	43,424.77
	Max	50	35.7	(1.1, 20.0)	104,737.23

[1] Best solution used if no optimal solution found after 12-hour processing for the single-objective problem.

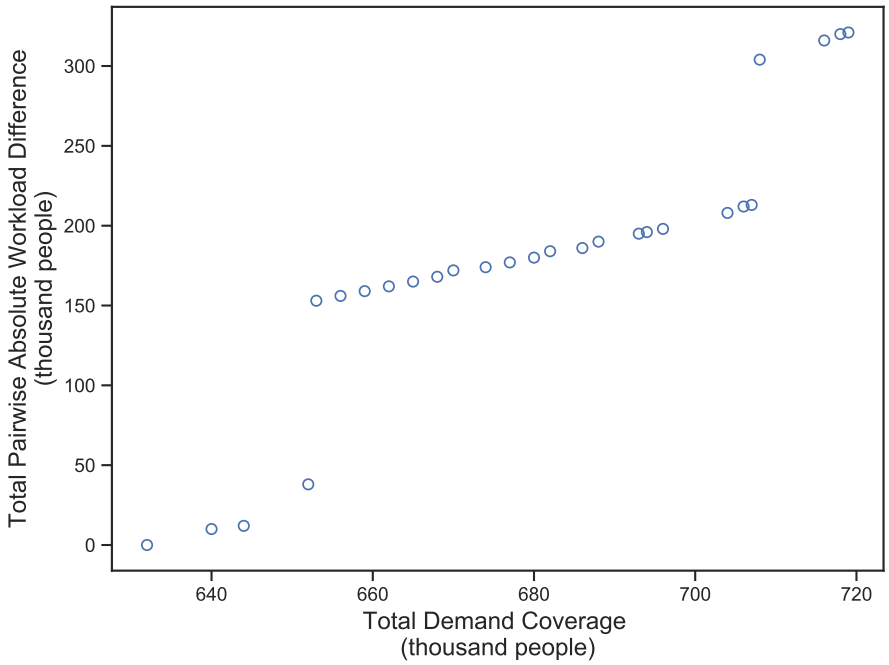

Figure 3.6 Pareto optimal solutions for WBMCLP-TotPairDiff (Santa Barbara, $p = 4$)

3.6 Discussion

There are a number of observations worth discussion associated with workload balancing. One is the issue of withholding of service. Another is how the detailed models compare to often relied upon capacitated approaches.

3.6.1 Withholding Service

As detailed previously, forcing assignment constraints (3.63)-(3.65) were imposed in order to avoid the situation of not assigning demand within the service coverage standard

of sited facilities. Models seeking to balance workload could inadvertently opt to not serve demand simply to balance workloads even though they are likely to require service because they are within a coverage standard of a sited facility. In order to illustrate this phenomenon, results are now noted for the case where no forcing of assignment is imposed using the WBMCLP-TotPairDiff. Figure 3.7 shows the trade-off between the total coverage and the pairwise absolute workload difference when $p = 4$ (San Jose). There are 40 Pareto optimal solutions in this case. This can be contrasted to the 9 optimal solutions in Figure 3.2 (WBMCLP-TotPairDiff with forcing assignment constraints). Why the difference? This is because there is more flexibility in service allocation. What is noteworthy is that the spatial configuration of facilities never changes, only the allocation of demand. For example, the facility configuration with total demand coverage of 616 and total pairwise workload variation of 384 is shown in Figure 3.7. The configuration (Figure 3.8) is the same as that of the solution given in Figure 3.3(a) (total coverage of 619 and workload variation of 389). The smaller workload variation in Figure 3.7 is achieved simply by not serving a portion of demand that is within the coverage standard. The unassigned demand is circled in Figure 3.8, effectively reducing the workload of a more heavily utilized facility from 196 to 193. This withholding of service phenomenon is not uncommon due to the workload balancing objective, which leads to the linear-looking point clusters in the trade-off plot (Figure 3.7). As discussed in Section 3.3.4, this withholding of service is problematic in practice. Worth mentioning as well is that computational effort actually increases. The extended allocation possibilities make work-

load balancing even harder to solve and require more time compared to those that force allocation. As an example, it takes 6.737 seconds to obtain the complete set of Pareto optimal solutions in Figure 3.2 while 50.499 seconds for those in Figure 3.7. But as noted previously, the artificial withholding of demand is problematic to begin with.

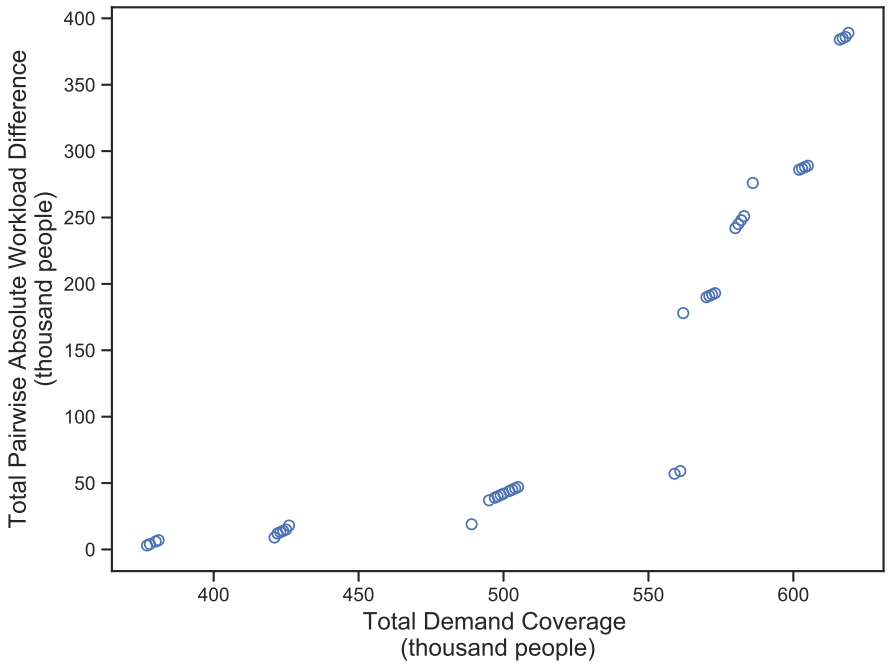

Figure 3.7 Pareto optimal solutions for WBMCLP-TotPairDiff with no forcing assignment constraints (Santa Barbara, $p = 4$)

3.6.2 Capacitated Comparison

Noted previously was that capacitated and threshold models are often relied upon to deal with workload balancing in covering problems, with the expectation that they ensure equity between facilities through explicit lower and upper limits on allocated demand. A

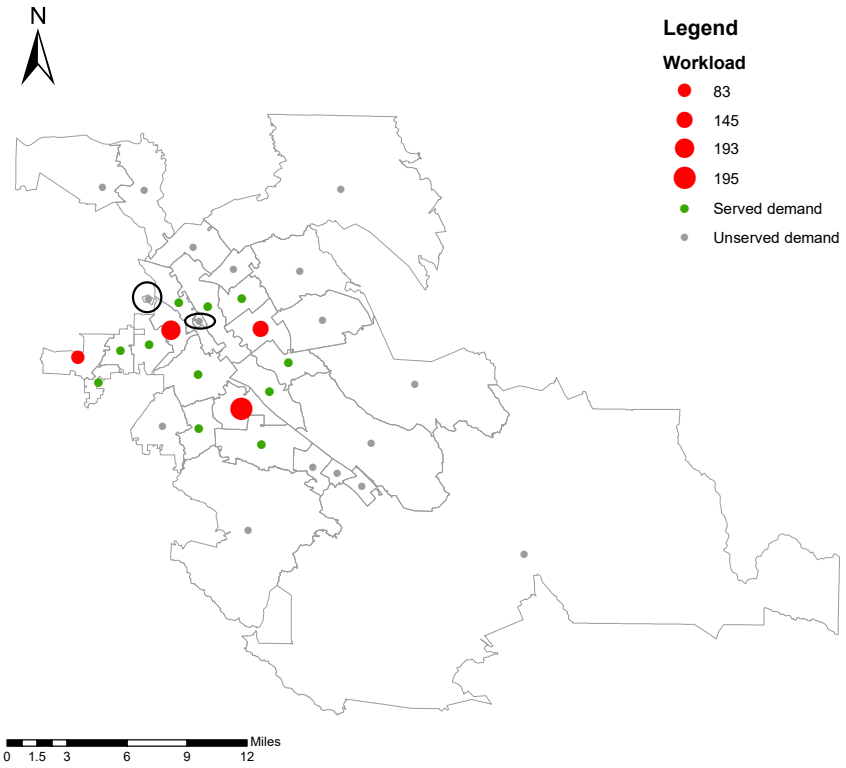

Figure 3.8 A Pareto optimal solution configuration for WBMCLP-TotPairDiff (no forcing assignment constraints)

prominent model is the CMCLP (Chung et al., 1983; Church and Somogyi, 1985), and is available for general application through ArcGIS (Xu et al., 2020). How do the reported findings compare to the CMCLP for workload balancing? Denote c_j the capacity of facility j. The CMCLP can be formulated as follows:

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{3.68}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{3.69}$$

$$\Sigma_{j \in J} X_j = p \tag{3.70}$$

$$\Sigma_i a_i Y_{ij} \leq c_j X_j \quad \forall j \in J \tag{3.71}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{3.72}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{3.73}$$

Capacity limits are imposed in Constraints (3.71), keeping facility j's workload from exceeding an upper limit c_j. There are two roles served by Constraints (3.71): first, the facility workload of a facility cannot be greater than its upper limit if it is sited; second, a demand can never be allocated to a facility that is not sited. Objective (3.68) and other constraints reflect the MCLP. This formulation contains $\Sigma_{i \in I}|N_i| + |J|$ decision variables and $\Sigma_{i \in I}|N_i| + |I| + 2|J| + 1$ constraints, significantly smaller than the WBMCLP-TotPairDiff in size. Note that the withholding service phenomena also happen when the CMCLP is applied due to capacity limits. Thus, the addition of forcing assignment constraints, (3.63)-(3.65), is necessary in the context of the previously applied models. However, forcing assignment does raise the likelihood that the resulting capacitated problem is infeasible.

The CMCLP is applied to the San Jose application and compared with the WBMCLP-TotPairDiff. The c_j is established to take a very wide range of values in order to fully explore capabilities of the CMCLP. Problem instances when siting four postal service facilities (i.e., $p = 4$) are solved and compared. Note that there is one capacitated instance that is infeasible due to forcing assignment constraints. The CMCLP only finds 11.1% of the solutions in the Pareto optimal solution set of the WBMCLP-TotPairDiff

and 40% of the solutions are inferior. The maximum gap in the demand coverage is 1.2% and in the total pairwise measure is 47.5%. These results indicate that the capacitated model does not balance facility workloads well in maximal covering by imposing upper limits on facility workloads. However, the computational time for solving the CMCLP (1.254 seconds when $p = 4$) is much less compared to that of workload balancing models.

3.7 Conclusion

This chpater proposed five workload balancing approaches in the context of maximal covering. In order to incorporate the goal of workload balance, five variation measures were detailed, including total pairwise absolute workload difference, total mean absolute workload deviation, maximum mean absolute workload deviation, workload range, and maximum workload. It is mathematically shown that the total pairwise measure tracks the workload variation most explicitly and other measures can be viewed as approximations. To incorporate the measures in a model, a second objective of minimizing the workload variation and associated constraints were proposed to extend the MCLP for workload balancing. As a result, five workload balancing models were formulated, including the WBMCLP-TotPairDiff that explicitly minimizes total pairwise absolute workload difference (benchmark) and along with the other four approximation models (non-benchmark). Additional constraints that force demand service allocation were also discussed and imposed. In order to assess capabilities of the proposed models to ad-

dress equity in maximal covering, an evaluation approach was formalized. This used the WBMCLP-TotPairDiff as the benchmark model and three quantitative evaluation measures, completeness, inferiority and maximum gaps, to evaluate the other four non-benchmark models.

Empirical studies involving postal service in San Jose and fire response in Santa Barbara were used to evaluate these models in terms of solution quality and computational effort. Results indicated that if a model does not appropriately reflect workload balance explicitly, optimality is not likely and large deviations are possible. When the number of facilities to site is small, the deviation from optimality was found to be slight. However, deviation from optimality increases significantly when many facilities are being located. The sacrifice is that the WBMCLP-TotPairDiff requires significantly more computational effort to solve. Computational efficiency by exact solvers for all proposed models is a major concern, limiting the size of practical applications that can be addressed. The computational time could be more than a week for selecting 5 facilities involving 80 demand nodes. The traditionally used capacitated method was also compared with proposed workload balancing approaches. Results showed that the capacitated MCLP does not balance facility workloads well by imposing workload upper limits though it is much easier to solve by an exact solver.

The implications of this chapter are many. Most importantly, modeling approaches based on approximation measures should be thoroughly understood. It is clear that in the context of maximal coverage that meaningful Pareto optimal results will not likely

be found unless the explicit approach, minimizing pairwise absolute workload difference, is used. Another critical finding is that significant computational challenges remain for supporting planning and decision making contexts. Problem application size is limited for exact solution. Future research is essential for developing more efficient solution techniques capable of addressing workload balancing in the context of maximal coverage.

Chapter 4

A Heuristic Algorithm for Balancing Workloads in Coverage Modeling

4.1 Introduction

Location covering models have been important spatial analytic approaches, used to support strategic planning, management and decision making in public and private sector contexts. The coverage concept is often related to service provision, acknowledging that response criteria like access and accessibility are fundamental. Church and Murray (2018) characterized coverage as a maximum distance or travel time standard for personnel at a facility to respond to a demand for service, detailing many examples of coverage standards. One of the most prominent coverage planning approaches is the maximal covering location problem (MCLP), formulated by Church and ReVelle (1974) to identify a set of facilities that can serve the most demand possible. It is considered NP complete with respect to computational complexity (Garey and Johnson, 1979; Megiddo et al., 1983),

This chapter represents a revised version of a paper submitted to *Computers, Environment and Urban Systems*, co-authored with Dr. Alan T. Murray, Dr. Richard L. Church and Dr. Ran Wei.

suggesting that specific problem instances can be difficult and challenging to solve. As a result, both exact and heuristic techniques are critical in solving the MCLP (Church and ReVelle, 1974; Murray and Church, 1996; Galvão and ReVelle, 1996; Zarandi et al., 2011; Church and Murray, 2018). The MCLP has been widely applied and extended in various ways (Murray and Tong, 2007; Lee and Murray, 2010; Tong and Church, 2012; Wei and Murray, 2015; Church and Li, 2016; Murray, 2016; Tong and Wei, 2017; Murray et al., 2019), reflecting its broad utility and applicability.

One of the underlying assumptions of the classic MCLP is that facilities can serve/cover an unlimited amount of demand. This simple assumption often leads to identified facility configurations with severely imbalanced workloads, where some facilities are overutilized while others are under-utilized. Significant workload variation in a system, however, can result in inequities, and may be unsustainable in many ways. Research seeking to control facility workload variation in coverage modeling includes imposing capacity limits (Chung et al., 1983; Church and Somogyi, 1985; Current and Storbeck, 1988; Elkady and Abdelsalam, 2016; Ferrari et al., 2018) and/or adding threshold limits that are workload lower bounds for sited facilities (Balakrishnan and Storbeck, 1991; Gerrard, 1995; Haghani, 1996; Church, 1980). Capacitated coverage problems are usually much more computationally complex and decidedly harder to be solved by exact approaches compared to the MCLP. This is because the introduction of facility capacities and/or thresholds result in allocation variables that tend to be fractional in a relaxed linear programming problem, necessitating the use of branching based methods to identify feasible integer solutions (Xu

et al., 2020). This is precisely why many heuristics have been developed for solving capacitated models (Chung et al., 1983; Church and Somogyi, 1985; Hogan and ReVelle, 1985; Shariff et al., 2012). Although capacitated models can control workloads to some extent, they are not direct ways to govern variation in covering problems. An explicit approach involves the use of multi-objectives, where demand coverage and workload variation can be simultaneously addressed. However, multi-objective problems can be computationally challenging as well, particularly associated with the MCLP, with practical limitations encountered for exact methods. Therefore, effective heuristics are essential, yet no research to date has developed approaches capable of solving multi-objective coverage models that focus on workload variability.

This chapter details a heuristic algorithm for balancing workloads in coverage modeling. Related research is reviewed in Section 4.2. Then, the mathematical formulation of an explicit workload balancing coverage model is given in Section 4.3. The design of a heuristic algorithm for solving this problem is derived in Section 4.4. Application results are presented in Section 4.5 followed by a discussion in Section 4.6. Finally, this chapter ends with concluding observations and comments.

4.2 Background

The intent of this chapter is to extend the MCLP in order to account for workload balancing. Thus, the discussion starts with solving the MCLP. Church and ReVelle

(1974) introduced the MCLP, as noted previously, along with two heuristic algorithms for solving it: greedy adding and greedy adding with substitution. The greedy adding with substitution heuristic seeks to improve a solution by exchanging or substituting an available facility location with an already selected facility location following the initial greedy procedure. Weaver and Church (1983) used Lagrangean relaxation with subgradient optimization to solve several forms of the MCLP. Adenso-Diaz and Rodriguez (1997) employed a tabu search metaheuristic to generate solutions for the MCLP. Galvão and ReVelle (1996) proposed a Lagrangean heuristic for solving the MCLP, producing an upper bound by a vertex addition and substitution heuristic; and lower bound by a subgradient optimization algorithm. Downs and Camm (1996) developed a dual-based heuristic with branch-and-bound for the MCLP. Pirkul and Schilling (1989) also utilized Lagrangean relaxation to solve a capacitated coverage maximization problem with backup service. Murray and Church (1996) designed a simulated annealing algorithm for solving the MCLP, providing comparisons to substitution/interchange approaches. Jaramillo et al. (2002) applied a genetic algorithm to solve several location problems, including maximal covering problems, and compared its performance against well-known heuristics using publicly available data sets. Tong et al. (2009) introduced a genetic algorithm for solving an extension of the MCLP, and a genetic algorithm was also used by Zarandi et al. (2011) to solve large-scale MCLP instances.

The above methods can be categorized as direct approaches for solving the MCLP. In contrast, there is another group of methods that is regarded as indirect. Church

and ReVelle (1976) showed how a maximal covering problem could be structured as an equivalent p-median problem. This was achieved by a distance matrix transformation. Because of this seminal work, it is possible to utilize algorithms developed for solving the p-median problem to solve the MCLP. Church and ReVelle (1976) discuss two p-median oriented heuristics for the MCLP, the well-known substitution/interchange algorithm of Teitz and Bart (1968) and the other a neighborhood search algorithm of Maranzana (1964). Among various algorithms designed for the p-median problem, a prominent approach is the global regional interchange algorithm. It attempts to speed up the interchange algorithm through data structure modifications in order to solve large problem instances (Densham and Rushton, 1992). The global regional interchange algorithm was subsequently integrated into ARC/INFO, a commercial geographic information system software package, providing capabilities to heuristically solve the MCLP (Church and Sorensen, 1996; Church, 2002). Church and Sorensen (1996) also discussed the possibility of integrating the Greedy Randomized Adaptive Search Procedure into a geographic information system for solving general location-allocation problems, including covering problems, representing much of how the heuristic in the Location-Allocation module of ArcGIS now works (Murray et al., 2019). While ArcGIS is effective in solving covering problems, Murray et al. (2019) found that optimal solutions are unlikely to be identified, with deviations from optimality being rather large in some instances.

Balancing workloads is a practical objective in location modeling, but also is at the heart of equity concerns (Mumphrey et al., 1971; Savas, 1978). Although facility equity

has indirectly been considered through the imposition of capacities and/or thresholds (Chung et al., 1983; Church and Somogyi, 1985; Current and Storbeck, 1988; Balakrishnan and Storbeck, 1991; Gerrard, 1995; Xu et al., 2020), there are direct attempts to account for equity in location-allocation problems. Weaver and Church (1981) solved a bi-objective p-median problem that involved minimizing weighted distance and balancing facility workloads. Murray and Gerrard (1997) proposed a capacitated regionally constrained p-median problem that accounted for equity issues. They solved this problem by Lagrangean relaxation. More recent work is that of Daskin and Tucker (2018), with a second objective introduced to minimize the range of assigned demand to sited facilities in order to obtain facility equity. A genetic algorithm was designed to solve this bi-objective problem. Workloads were also explicitly balanced in center problems by Davoodi (2019) through the addition of two additional objectives that minimize the maximum demand a center serves and the range of workloads. Davoodi (2019) proposed an iterative algorithm based on a Voronoi diagram to solve the problem. Beyond this research, there are many studies that sought to balance facility workloads in other types of location-allocation problems. For example, Berman et al. (2009) detailed heuristics to locate p facilities such that the maximum weights attracted to facilities is minimized. Kim and Kim (2010), Marín (2011), and Mišković and Stanimirović (2015) presented other attempts for balancing workload in location-allocation problems.

Workload balancing has also been considered in allocation problems. Zhu and McKnew (1993) developed a workload balancing model to allocate a fixed number of ambu-

lances in defined service locations. Utilization of ambulances at different locations was balanced through the minimization of the workload deviation from a system-wide average. Huang et al. (2006) proposed a stochastic model that minimized the total pairwise workload difference between two airline terminals over different time periods. They also developed a Benders decomposition algorithm to accelerate computation. Storage space allocation problems have also been addressed, where the objective is to balance container terminal workloads (Zhang et al., 2003; Bazzazi et al., 2009). Closely related are districting problems as well, because a typical objective is to minimize population variation among different districts in order to ensure political fairness or balance administrative loads. For example, Garfinkel and Nemhauser (1970) controlled the maximum population deviation from the average using a two-phase procedure. Church and Murray (1993) minimized the total pairwise absolute difference in utilization associated with school districting. D'Amico et al. (2002) constrained the ratio of the largest and smallest districts from exceeding a specified upper bound in order to balance patrol car allocation using a simulated annealing approach.

The above research demonstrates the necessity of an efficient solution method for maximal covering problems, and more generally location problems, that consider workload balance. The focus of this chapter is on a maximal covering problem extension that considers facility workload balance. This problem has proven to be computationally challenging. A heuristic is therefore proposed to derive high-quality solutions for this multi-objective problem.

4.3 Addressing Workload Balancing

The previous sections have highlighted the significance of balancing workloads as well as the utility of the MCLP to address a range of important planning contexts. Unfortunately the MCLP and other coverage models generally do not enable workload balance to be addressed explicitly. This section offers a formulation of the workload balancing MCLP as an extension that incorporates an additional objective to explicitly minimize total pairwise absolute workload differences (WBMCLP-TotPairDiff) between sited facilities.

Consider the following notation:

i = index of demand areas (I entire set)

j = index of potential facilities (J entire set)

d_{ij}= travel distance/cost/time between demand i and facility j

S = service coverage standard

$N_i = \{j|d_{ij} \leq S\}$, the set of facilities that can suitably cover demand i

a_i = amount of demand in area i

c_j = capacity of facility j

p = number of facilities to site

M = a very large positive number

Decision variables:

$$X_j = \begin{cases} 1 & \text{if facility } j \text{ is sited} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{ij} = \begin{cases} 1 & \text{if demand } i \text{ is allocated to facility } j \\ 0 & \text{otherwise} \end{cases}$$

$D_{jj'}$ = the absolute workload difference between any two potentially sited facilities j and $j^{'}$

$$C_i = \begin{cases} 1 & \text{if demand } i \text{ is within the service coverage standard of a sited facility} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Maximize} \quad \Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \tag{4.1}$$

$$\text{Minimize} \quad \Sigma_j \Sigma_{j' > j} D_{jj'} \tag{4.2}$$

$$\text{Subject to} \quad \Sigma_{j \in N_i} Y_{ij} \leq 1 \quad \forall i \in I \tag{4.3}$$

$$Y_{ij} \leq X_j \quad \forall i \in I, j \in N_i \tag{4.4}$$

$$\Sigma_{j \in J} X_j = p \tag{4.5}$$

$$\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'} - M(1 - X_{j'}) \leq D_{jj'} \quad \forall j, j^{'} \in J \& j^{'} > j \tag{4.6}$$

$$\Sigma_i a_i Y_{ij'} - \Sigma_i a_i Y_{ij} - M(1 - X_j) \leq D_{jj'} \quad \forall j, j^{'} \in J \& j^{'} > j \tag{4.7}$$

$$\Sigma_{j \in N_i} X_j \leq min\{p, |N_i|\} C_i \quad \forall i \in I \tag{4.8}$$

$$C_i \leq \Sigma_{j \in N_i} Y_{ij} \quad \forall i \in I \tag{4.9}$$

$$X_j = \{0, 1\} \quad \forall j \in J \tag{4.10}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in I, j \in N_i \tag{4.11}$$

$$D_{jj'} \geq 0 \quad \forall j, j^{'} \in J \& j^{'} > j \tag{4.12}$$

$$C_i = \{0, 1\} \quad \forall i \in I \tag{4.13}$$

The WBMCLP-TotPairDiff has two objectives. Objective (4.1) maximizes the total demand suitably covered. Objective (4.2) minimizes the absolute workload difference between sited facilities. Constraints (4.3)-(4.5) structure an equivalent MCLP as a location-allocation problem. A demand can only be allocated to a facility at most once in Constraints (4.3). Constraints (4.4) prohibit allocation of demand unless a facility is sited. Exactly p facilities are to be sited in Constraint (4.5). Constraints (4.6) and (4.7) track the workload difference between facilities. When considering the siting of two facilities j and $j^{'}$, there are three situations: 1) facilities j and $j^{'}$ are sited; 2) facilities j and $j^{'}$ are not sited; and, 3) either j or $j^{'}$ is sited. Accordingly, Constraints (4.6) and (4.7) must accurately track differences between the siting of j and $j^{'}$ for these three situations. In the case that both are sited, which is the only situation of interest with respect to workload difference, Constraints (4.6) and (4.7) would become $\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'} \leq D_{jj'}$ and $\Sigma_i a_i Y_{ij'} - \Sigma_i a_i Y_{ij} \leq D_{jj'}$, respectively. Thus, the outcome is effectively $D_{jj'} \geq |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|$, as intended. For the situation that both are not

sited, Constraints (4.6) and (4.7) become $-M \leq D_{jj'}$. Coupled with non-negativity conditions in Constraints (4.12), the outcome is $D_{jj'} \geq 0$, as intended. In the third case where only one facility is sited (e.g., j is sited and $j^{'}$ is not), then Constraints (4.6) and (4.7) result in $\Sigma_i a_i Y_{ij} - M \leq D_{jj'}$ and $-\Sigma_i a_i Y_{ij} \leq D_{jj'}$, respectively. As $D_{jj'}$ is non-negative in Constraints (4.12), then it must be no less than 0, as intended. Collectively, the outcome of Constraints (4.6) and (4.7) is the measurement that $D_{jj'} \geq |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|$ for pairs of sited facilities and $D_{jj'} \geq 0$ for other paired outcomes. Since $\Sigma_j \Sigma_{j'>j} D_{jj'}$ is minimized in objective (4.2), $D_{jj'}$ will seek to be the smallest value possible, which is the pairwise absolute difference $|\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|$ or 0. Note that the value of M affects solution efficiency and ideally a smallest sufficient M is preferred for a more efficient computation. Here the M in Constraints (4.6) is set to $\Sigma_{i \in R_j} a_i$ and in Constraints (4.7) is set to $\Sigma_{i \in R'_j} a_i$ in implementation. Constraints (4.8) and (4.9) force assignment/allocation for any demand that is within the coverage standard of a sited facility, necessary to avoid artificial withholding of service. Constraints (4.8) require C_i to be 1 if at least one facility that can suitably cover demand i is sited, where the $min\{p, |N_i|\}$ is also used to tighten the model. If demand i can be suitably covered by a sited facility, i.e. $C_i = 1$, then demand i must be allocated because $\Sigma_{j \in N_i} Y_{ij} \geq 1$ by Constraints (4.9). Finally, Constraints (4.10), (4.11), (4.12) and (4.13) specify binary integer and non-negativity restrictions.

There are a number of WBMCLP-TotPairDiff features that make it challenging and difficult to solve. The workload balancing objective, (4.2), is among the most significant. A PARTITION problem can be reduced to the WBMCP-TotPairDiff with fixed facili-

ties. Assume p sited facilities are known. Due to constraints (4.8) and (4.9) that force assignment (or partitioning) of covered demand, such coverage would be fixed once sited facilities are determined. Thus, the WBMCLP-TotPairDiff becomes an allocation (partition) problem seeking to minimize the total pairwise absolute workload difference. The allocation is trivial for facilities that are not sited (i.e., when $X_j = 0$) as well as demand that is beyond the service coverage standard of any sited facility. Assignment variables become zero in this case, i.e., $Y_{ij} = 0$, because an un-sited facility cannot serve demand and demand cannot be allocated to a facility beyond the cover standard. This leaves the $Y_{ij} = 0$ variables associated with sited facilities (i.e., when $X_j = 1$) and demand that are within the coverage standard of one or more sited facilities needing to be resolved. Define J^* as the set of sited facilities, i.e., $J^* = \{j|X_j = 1\}$; R_j the set of demand that can be suitably covered by facility j, where $R_j = \{i|d_{ij} \leq S\}$; and, R the union of R_j for all $j \in J^*$, i.e., the set of demand that can be suitably served. The allocation problem can be formulated as follows.

$$\text{Minimize} \quad \Sigma_j \Sigma_{j'>j} D_{jj'} \tag{4.14}$$

$$\Sigma_{i \in J^* \cap N_i} Y_{ij} = 1 \quad \forall i \in R \tag{4.15}$$

$$\Sigma_{i \in R_j} a_i Y_{ij} - \Sigma_{i \in R_{j'}} a_i Y_{ij'} \leq D_{jj'} \quad \forall j, j^{'} \in J^* \& j^{'} > j \tag{4.16}$$

$$\Sigma_{i \in R_{j'}} a_i Y_{ij'} - \Sigma_{i \in R_j} a_i Y_{ij} \leq D_{jj'} \quad \forall j, j^{'} \in J^* \& j^{'} > j \tag{4.17}$$

$$Y_{ij} = \{0, 1\} \quad \forall i \in R_j, j \in J^* \tag{4.18}$$

Objective (4.14) minimizes the total pairwise absolute workload difference between sited facilities, noted in (4.2). Constraints (4.15) stipulate that a demand within the standard of a sited facility must be allocated to one and only one facility. Thus, Constraints (4.15) are a combination of Constraints (4.3), (4.8) and (4.9). Constraints (4.16) and (4.17), which are simplified versions of Constraints (4.6) and (4.7), track the absolute workload difference between two sited facilities j and $j^{'}$. Constraints (4.18) define binary allocation decision variables Y_{ij} associated with sited facilities and demand that can be covered. Again, this assumes that the set of sited facilities, J^*, is known, which is not the case for the WBMCLP-TotPairDiff, making it even more difficult to solve.

Assume there is an instance of a PARTITION problem with $|R|$ members. Set the demand a_i where $i \in R$ to be equal to the $|R|$ members to be partitioned. Then, set $p = 2$ (therefore $|J^*| = 2$) and $N_i = J$ for each i. Clearly, the PARTITION instance has a feasible solution if and only if the constructed instance of the allocation problem (4.14)-(4.18) has an optimal solution with a value of 0. Thus, (4.14)-(4.18) is also NP complete (see Garey and Johnson 1979), suggesting it is indeed difficult to be optimally solved and a polynomial time algorithm is not possible unless $P = NP$. It is also evident that the allocation variable Y_{ij} tends to be fractional due to objective (4.14), which makes finding an integer solution more difficult during a branch-and-bound procedure in an exact algorithm. Thus, the computational time using exact methods for this allocation problem theoretically becomes very large as problem size grows. Therefore, it is even more computationally challenging to solve the bi-objective WBMCLP-TotPairDiff involving

facility location and demand allocation simultaneously, as (4.14)-(4.18) is a special case. Accordingly, exact methods are not likely to be computationally viable for problems encountered in practical application, as highlighted in the results that follow. This means that the development of an efficient solution approach is needed for any application of the WBMCLP-TotPairDiff.

4.4 Heuristic Algorithm

Since exact capabilities for solving the WBMCLP-TotPairDiff are limited, a heuristic algorithm is proposed that is capable of identifying high quality solutions in an efficient manner. Figure 4.1 gives a flowchart describing the developed heuristic. There are three major stages of the proposed approach: initialization, subproblem delineation and local search. In the initialization stage, the algorithm begins with an empty set Υ for tracking solutions, pre-defined threshold coverage bounds Ψ_k for subproblems, and a group of simulated annealing parameters (initial temperature T, stopping temperature T_{min}, cooling factor β, etc.) for local search. The algorithm then identifies a set of feasible facility configurations using an initial hybrid facility site selection procedure. Each obtained facility configuration is used as a starting point in the subproblem delineation stage.

Given an initial facility configuration, the next stage in Figure 4.1 decomposes the WBMCLP-TotPairDiff into a set of subproblems using the constraint method. This is a multi-objective solution approach that is broadly applied, the details of which can be

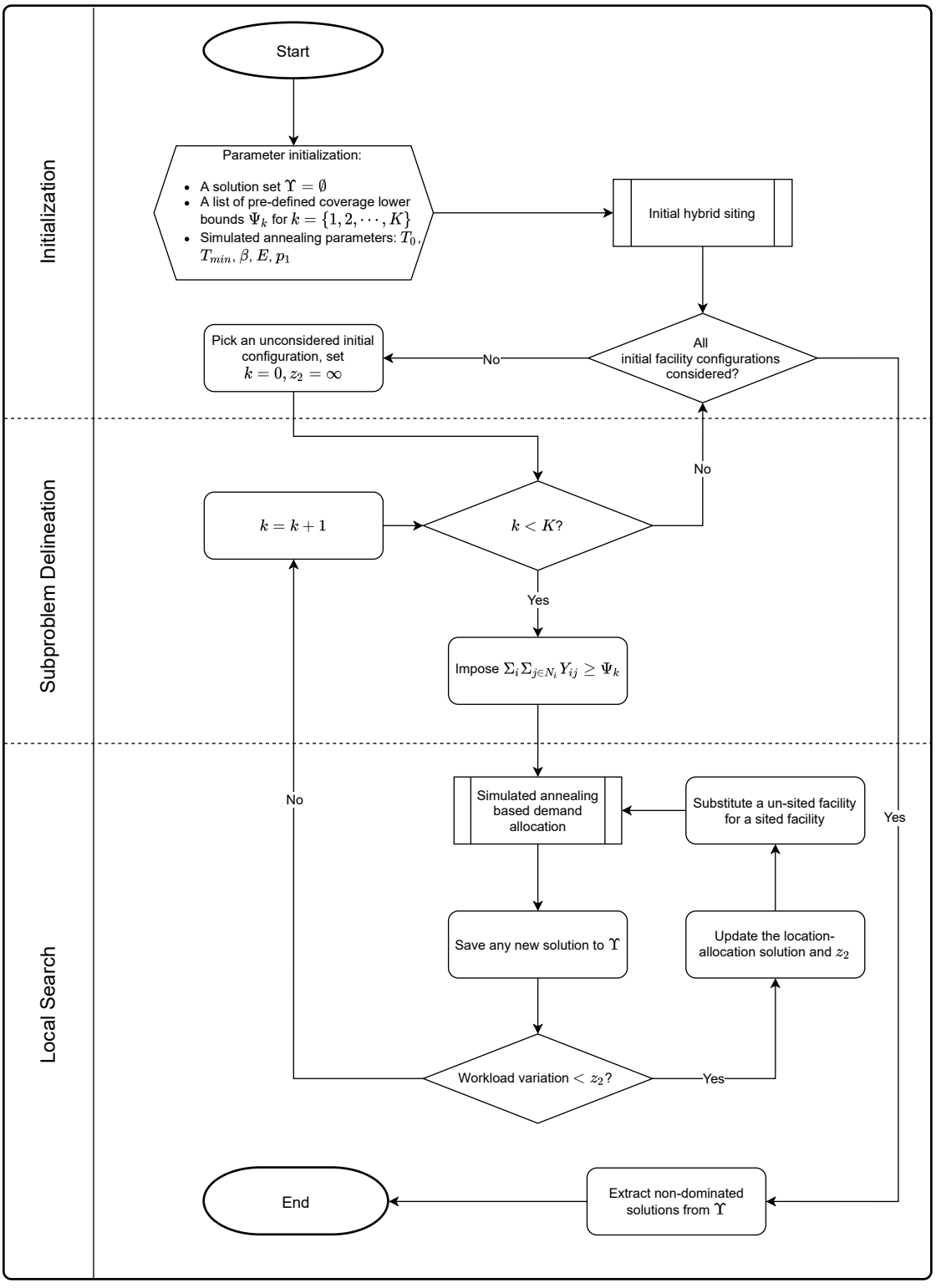

Figure 4.1 Heuristic solution algorithm

found in Cohon (1978). This is accomplished by making objective (4.1) a constraint, imposing that the total demand coverage is no less than the bound. Specifically, for a facility configuration obtained in the initialization stage, the algorithm considers successive coverage lower bounds Ψ_k for $k = 1, 2, \cdots, K$, with objective (4.1) becoming a constraint of the following form:

$$\Sigma_i \Sigma_{j \in N_i} a_i Y_{ij} \geq \Psi_k \tag{4.19}$$

The subproblem, consisting of (4.2)-(4.13) and (4.19), now minimizes the workload variation while ensuring the total demand coverage is at least Ψ_k for a given facility configuration.

The local search stage in Figure 4.1 then follows, seeking to improve a subproblem solution. There has been a long history of successful heuristics focused on separating facility location and demand allocation decision making when solving location-allocation problems (Maranzana, 1964; Teitz and Bart, 1968; Haghani, 1996; Daskin and Tucker, 2018). The local search is similarly conceived in that a facility location is the central focus of substitution, with demand allocation carried out using a simulated annealing approach. Specifically, the local search stage begins with a facility configuration, and then employs a simulated annealing approach to obtain the allocation. The algorithm then considers substitutions of a facility site that are not part of the current facility configuration; such a substitution then triggers the need to update demand allocations, subsequently achieved via the simulated annealing approach. Improvements to the location-allocation

configuration are accepted if they produce a decrease in workload variation while maintaining the stipulated demand coverage. Local search stops when no improvement in the workload variation can be obtained for the current subproblem. Any new solution found during the local search is added to Υ. This enhances the likelihood that non-dominated/high-quality solutions are identified as well as avoids redundant computation associated with already identified solutions.

With the updated location-allocation solution, the algorithm in Figure 4.1 then returns to define another subproblem for a different threshold coverage bound Ψ_k for $k = \{1, 2, \cdots, K\}$ followed by local search. This process is repeated until all subproblems have been defined and solved. Once all initial facility configurations are considered in the above process, the heuristic ends by extracting and returning non-dominated solutions in the set Υ. Specifics associated with initial hybrid siting and simulated annealing allocation are now detailed.

4.4.1 Initial Hybrid Siting

The performance of a heuristic algorithm in location problems is usually sensitive to the quality of initial solutions (Adenso-Diaz and Rodriguez, 1997; Wei and Murray, 2014). Indeed, there are a number of ways to identify initial (feasible) solutions, including specification, randomization, greedy or hybrid approaches (e.g., Maranzana 1964; Church and ReVelle 1974; Murray and Church 1996; Daskin and Tucker 2018; Murray et al. 2019). Since the WBMCLP-TotPairDiff is a bi-objective optimization problem, the "goodness"

of the initial solution set depends on both objective values (i.e. the total demand coverage and the total pairwise workload variation) associated with every single solution and diversity of solutions (or distribution of solutions in the objective space). Wei and Murray (2014) proposed a hybrid way to obtain an initial population using a genetic algorithm for solving their bi-objective problem, which includes solving relaxed problems by ignoring some uncertain conflict constraints and random initialization. The superiority of their hybrid initialization compared to three other initialization approaches is essentially due to the consideration of solutions with high and low objective values. Similarly, in order to generate good solutions that enable complete exploration of the objective space, a hybrid facility siting procedure is taken here.

The initial hybrid siting procedure is given in Figure 4.2, which has three parallel components. First, the procedure relaxes the workload balancing constraints and objective, then solves the resulting MCLP by exact or heuristic approaches. Here, Gurobi is used to solve the MCLP and multiple optimal or close-to-optimal facility configurations are obtained by forcing 1 sited facility not to be sited each time from an initial MCLP optimal solution. These facility siting solutions are associated with large demand coverage (possibly large workload variation as well). There is some likelihood they are good starting points for solution search. Second, the procedure computes the maximum possible demand coverage by each potential facility, i.e., $\Sigma_{i \in R_j} a_i$. The procedure then inspects whether there are exactly p facilities that have the same $\Sigma_{i \in R_j} a_i$ and their R_j do not intersect (here the strict zero intersection condition is used, but one can set some

threshold to allow small intersected coverage). If feasible, such a p-facility configuration is kept. These facility location solutions may already have balanced workloads. Such cases are rare but do occur and help to find good starting solutions when p is small (2 or 3). The last component is to quickly derive initial facility configurations that span different demand coverage levels in order to achieve solution diversity. With each coverage bound Ψ_k, the procedure picks a group of p facilities having the smallest $\Sigma_{i \in R_j} a_i$ that is no less than an average workload Ψ_k/p. Therefore, there will be at least K groups of p-facilities generated by the last component. This hybrid initialization is expected to enhance performance of the algorithm compared to random initialization.

4.4.2 Simulated Annealing Based Demand Allocation

An important component of the local search stage is to allocate demand to sited facilities. This is accomplished in the context of solving an allocation problem, (14)-(18). As discussed above, this allocation problem is NP complete and requires an efficient method to solve it. Simulated annealing is a systematic method for perturbing an incumbent solution while allowing degradation of solution quality under some conditions in order to avoid the objective value becoming trapped in local optima (Karmarkar and Karp, 1982). It has been widely applied to many types of optimization problems, including p-median (Murray and Church, 1996; Golden and Skiscim, 1986), quadratic assignment (Burkard and Rendl, 1984; Sharpe and Marksjö, 1986; Wilhelm and Ward, 1987), partitioning

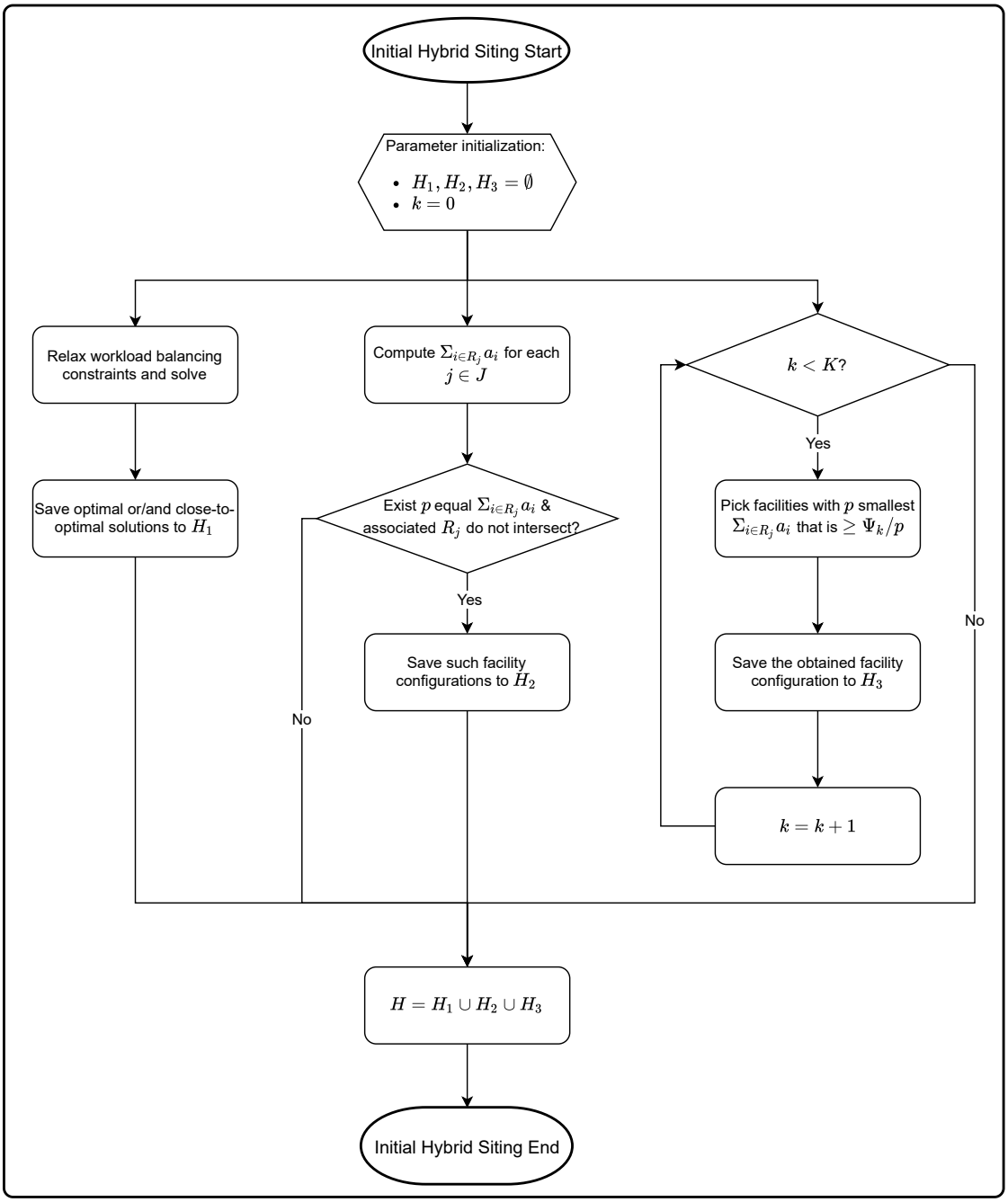

Figure 4.2 Initial hybrid siting configuration construction

(Johnson et al., 1991), police districting (D'Amico et al., 2002) and others. A simulated annealing approach for demand allocation is now detailed.

Figure 4.3 depicts the simulated annealing demand allocation approach. Parameters in this procedure include a cooling schedule (current temperature T, stopping temperature T_{min} and cooling factor $\beta \in (0,1)$), an equilibrium state threshold E, and a probability $p_1 \in [0,1]$. The approach starts with an initial allocation solution s obtained from a greedy allocation. Specifically, demand that is within the coverage standard of only one sited facility is allocated, then the facility having the minimum current workload is prioritized for serving more demand until all demand in R is allocated. At temperature T, a search to a neighbor solution s' is attempted. If such a move reduces the total pairwise workload variation, it is accepted and the current solution s is updated. Otherwise, the move is accepted with a probability that depends on the current temperature T and the degradation of the workload variation Δ. Specifically, the acceptance probability is defined by $e^{\frac{-\Delta}{T}}$ (see Kirkpatrick et al. 1983), and decreases as the temperature cools down. At a particular temperature T, many attempts of a neighbor solution are explored until an equilibrium state is reached. To reach an equilibrium state at each T, a sufficient number of neighbor solutions have to be searched, which is defined as at least E neighbor solutions searched where E is a proportion of the neighborhood size. Once the equilibrium state is reached, the temperature T is decreased according to a geometric cooling process, namely $T = \beta \times T$ where $\beta \in (0,1)$ (Talbi, 2009). These steps

iterate until $T < T_{min}$. The best solution found during the process is stored as the final allocation solution for the given facility configuration.

Although the basic structure of simulated annealing is similar in many different applications for solving an optimization problem, the effectiveness of a simulated annealing approach relies on how a neighbor solution of an incumbent solution is identified (Koulamas et al., 1994; Talbi, 2009). Here, a neighbor allocation solution is searched by re-allocating demand. There are many different ways to re-allocate, such as to re-allocate a single demand, making a one-to-one allocation exchange, making an allocation exchange among three demand, etc. Re-allocating a single demand to another sited facility is the most basic approach. A neighbor solution for an incumbent solution obtained from this approach is illustrated in Figure 4.4(a). This is done by a spatial search process for a sited facility $j^{'}$ that can cover a demand, replacing allocation from facility j. Re-allocating a single demand is very flexible, but doing so for a relatively balanced solution often increases total workload variation. As a result, such a move is unlikely to be accepted, making it an ineffective neighborhood search approach in this case. Another demand re-allocation strategy is a one-to-one exchange, leading to a neighbor solution similar to that illustrated in Figure 4.4(b). This type of solution can be obtained by a spatial search for a pair of demand i and $i^{'}$ such that i could be re-allocated to facility $j^{'}$ and $i^{'}$ could be re-allocated to facility j. Compared to the single demand re-allocation, the one-to-one exchange is more likely to reduce workload variation, making it a potentially effective attempt. For both types of re-allocation, if more than one re-allocation option

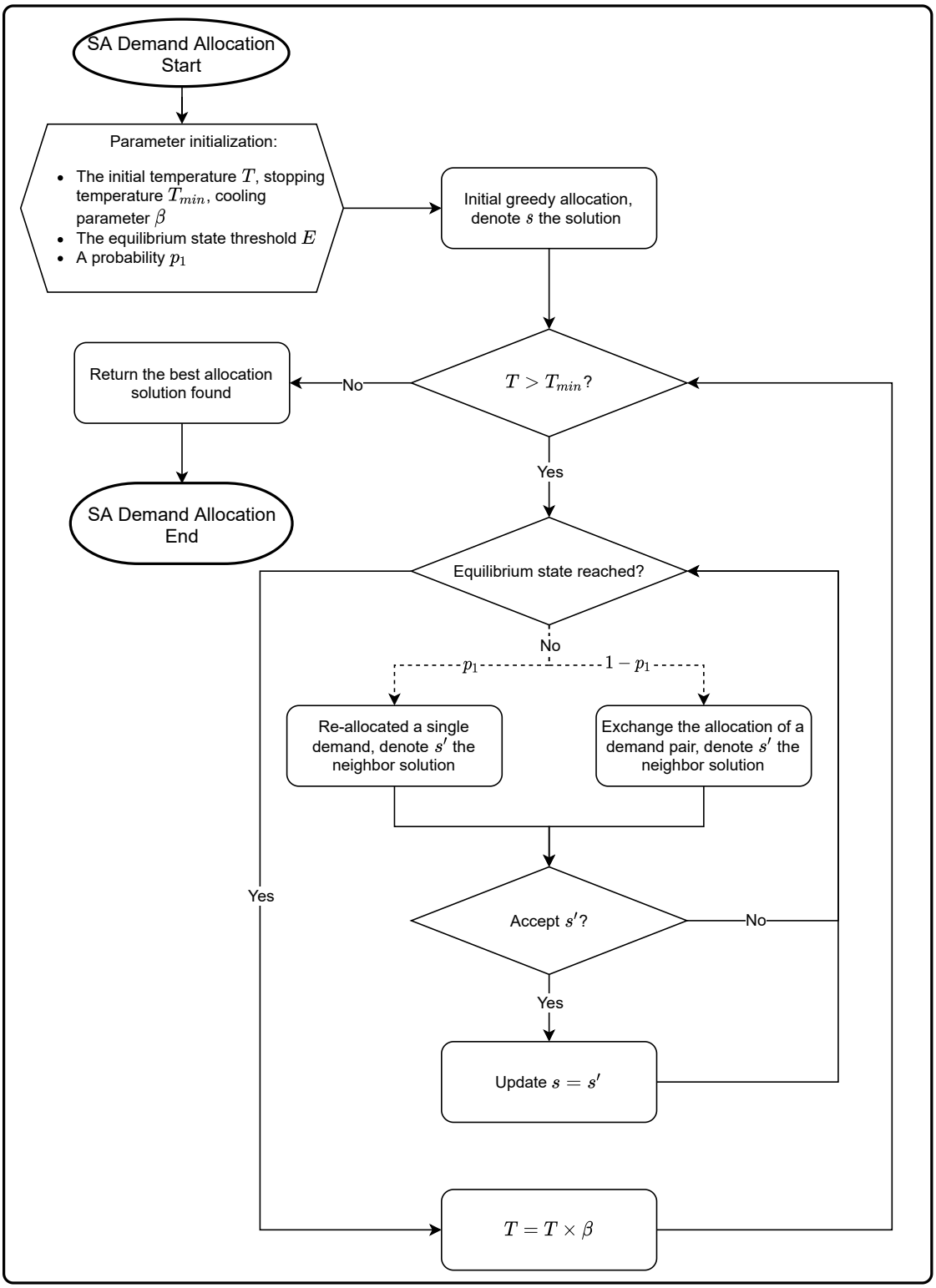

Figure 4.3 Simulated annealing demand allocation process

available, a random one is picked. The neighbor solution search employed here is designed to combine single demand and one-to-one re-allocation by randomly choosing between the two processes with a probability of p_1 for the single demand re-allocation and $1 - p_1$ for a one-to-one exchange.

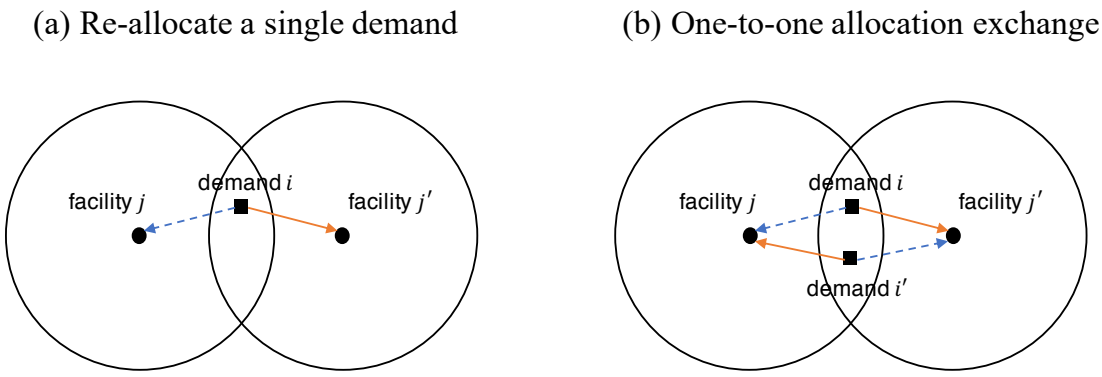

Figure 4.4 Two potential types of neighbor allocation

4.5 Application Results

The proposed algorithm for solving the WBMCLP-TotPairDiff was programmed in Python (version 3.6) and executed on a personal MacBook Pro (with 2.9 GHz Intel Core i5 processor and 8 GB memory). Four different case studies were used to evaluate the effectiveness of the proposed heuristic algorithm. The initial temperature T was set to make the median initial acceptance rate to be no less than 50%, $T = -median(\{\Delta\})/lg(0.5)$ where $\{\Delta\}$ is a set of workload variation values by 1,000 random move experiments. The stopping temperature T_{min} and cooling parameters β were then set accordingly to ensure

enough iterations in the simulated annealing. To establish benchmarks, the WMBCLP-TotPairDiff was solved optimally when possible using the constraint method (Cohon, 1978) and Gurobi (version 9.0.1). This then enables an assessment of heuristic performance, at least in the cases where the exact approach is successful. Three quantitative measures, completeness, maximum gaps and average gaps, are used for algorithm evaluation. Let P be the Pareto optimal set, Q be a set of solutions found by the algorithm, completeness is defined as $|P \cap Q|/|P|$, the ratio of the number of found Pareto optimal solutions to the number of all optimal solutions. A higher completeness means more Pareto optimal solutions are found by the algorithm. Denote Z_1 and Z_2 the total demand amount allocated and the total pairwise absolute workload difference of a solution to the WBMCLP-TotPairDiff, respectively. Let t be a missed optimal solution, i.e., $t \in P \setminus Q$, $t^{'}$ be the closest solution to t in Q, the maximum gaps are defined as $(\max_{t \in P \setminus Q} \frac{|Z_1^t - Z_1^{t^{'}}|}{Z_1^t}, \max_{t \in P \setminus Q} \frac{|Z_2^t - Z_2^{t^{'}}|}{Z_2^t})$, the maximum percentage deviation of optimal solutions from their closest solutions by the algorithm in both objective values. Similarly, the average gaps are defined as $(\frac{1}{|P \setminus Q|} \Sigma_{t \in P \setminus Q} \frac{|Z_1^t - Z_1^{t^{'}}|}{Z_1^t}, \frac{1}{|P \setminus Q|} \Sigma_{t \in P \setminus Q} \frac{|Z_2^t - Z_2^{t^{'}}|}{Z_2^t})$. The smaller maximum (average) gaps, the better obtained solutions approximate to the Pareto optimal front. Computational time is also noted.

4.5.1 San Jose Postal Service

The first case study is associated with postal service in the city of San Jose (Xu et al., 2020). The goal is to maximize the total expected demand (population) served while

balancing postal facility workloads simultaneously. There are 32 demand areas (ZIP Code Tabulation Areas) with a total demand of 1,023 (in thousands of people). The demand area centroids are used to represent both demand locations and potential postal service facilities. Travel distances between demand and potential facilities are computed based on the street network for the region (U.S. Census Bureau, 2018a). The service coverage standard is a distance of 3 miles. The number of facilities considered ranges from 2 to 15.

Parameters of the heuristic algorithm were set as follows. The total demand coverage bound Ψ_k is a percentage, e.g., $\{1.00, 0.98, 0.96, \cdots, 0.02\}$, of maximum total demand that can be covered for a given number of facilities, p. The initial temperature T is 300, the minimal temperature T_{min} is 10 and the cooling parameter β is 0.9. Finally, the equilibrium state threshold E is set to 10% of the neighborhood size (i.e. a rough number of possible neighbor solutions of the mentioned two types), and the probability of re-allocating a single demand p_1 is 50%.

Table 4.1 summarizes the computational results of the heuristic algorithm. In general, the heuristic is able to identify high-quality non-dominated solutions and the computational time needed is significantly faster than the exact solver, especially as problem difficulty increases as p increases. The heuristic is able to find more than 70% of Pareto optimal solutions in all cases and the average gap is usually less than 10% in the total demand coverage with ranges of 0-17% in the total pairwise absolute workload difference. When siting two or three facilities, the heuristic successfully finds all Pareto optimal

solutions, requiring less computational effort. As p becomes larger, the algorithm performance is slightly degraded but the computational efficiency of the heuristic is also more evident. When siting 15 facilities, there are 20 Pareto optimal solutions and 16 (80%) are found by the heuristic. Only 3 solutions are dominated. The average gap in total demand coverage is 1.5% and the workload variation is 3.6%. This problem is solved in around 5 minutes by the heuristic compared to 1 hour using the exact solver. The degraded effectiveness of the algorithm is expected as p increases, because (1) there are more combinatory siting options, which greatly increases the difficulty of finding the right p-facility combination in the facility substitution process; (2) there are more allocation options for a demand ($|R_i|$ becomes larger) and more overlapped coverage among sited facilities (a demand could be in multiple N_j for $j \in J^*$), which makes it hard to get the allocation swaps that lead to the optima in the simulated annealing process.

The trade-off between total demand coverage and workload variation can be visualized. Figure 4.5 shows the true Pareto optimal fronts derived by the exact solver and solutions found using the heuristic for both small and large values of p. When p is 3, there are 9 Pareto optimal solutions with total demand coverage ranging from 96 to 536 thousand people and the total pairwise deviation ranging from 6 to 102 thousand people (Figure 4.5(a)). In this case, all nine optimal solutions are found. The spatial configuration of the highly imbalanced location-allocation solution covers 536 thousand people and has a workload variation of 102 thousand people is given in Figure 4.6. The three sited facilities have workloads of 145, 195 and 196 thousand people, respectively.

Table 4.1 Computational results for proposed algorithm (San Jose postal service delivery, $S = 3$)

p	# of Pareto optimal solutions	# of Pareto optimal solutions found (Completeness)	Maximum Gaps (%, %)	Average Gaps (%, %)	Solution Time - Algorithm (Seconds)	Solution Time - Gurobi (Seconds)
3	9	9 (100.0)	(0.0, 0.0)	(0.0, 0.0)	0.63	7.89
4	9	9 (100.0)	(0.0, 0.0)	(0.0, 0.0)	0.90	7.34
5	20	17 (85.0)	(20.4, 23.5)	(8.8, 17. 4)	3.34	31.71
6	19	17 (89.5)	(1.7, 11.8)	(0. 9, 9.1)	12.05	66.89
7	13	11 (84.6)	(3.2, 13.0)	(1.6, 7.7)	15.88	52.75
8	16	12 (75.0)	(3.0, 3.8)	(1.3, 3.1)	19.54	108.63
9	25	18 (72.0)	(3.1, 8.6)	(1.4, 3.7)	56.52	776.00
10	28	22 (78.6)	(3.0, 18.7)	(1.8, 5.8)	64.00	1,270.63
11	21	15 (71.4)	(8.5, 8.0)	(4.2, 4.4)	62.21	1,422.25
12	21	16 (76.2)	(8.7, 3.9)	(3.6, 2.1)	67.99	2,269.48
13	22	16 (72.7)	(7.8, 4.6)	(3.0, 2. 3)	136.42	3,190.57
14	19	15 (78.9)	(5.7, 5.7)	(2.6, 3.2)	184.66	2,876.57
15	20	16 (80.0)	(1.8, 5.1)	(1.5, 3.6)	304.00	3,374.39

In contrast, a more balanced solution with three facilities serving 133, 138 and 139 thousand people is shown in Figure 4.7. This solution has a total demand coverage of 410 thousand people and a pairwise workload variation of 12 thousand people. When p is 15, the Pareto optimal front ranges from 775 to 1,006 in demand coverage and 918 to 5,326 in the workload variation (Figure 4.5(b)). A few Pareto optimal solutions were missed but they do span the frontier. As Figure 4.5 illustrates, the heuristic solutions are well distributed and cover the objective space of the Pareto optimal fronts. The deviations from the solutions on the front are small.

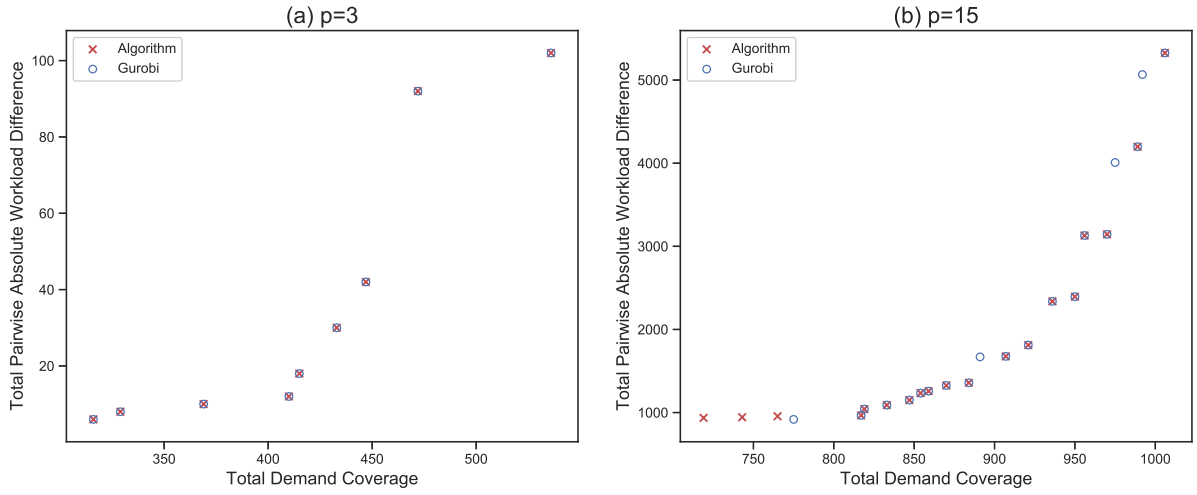

Figure 4.5 Pareto optimal front comparison (San Jose postal service delivery)

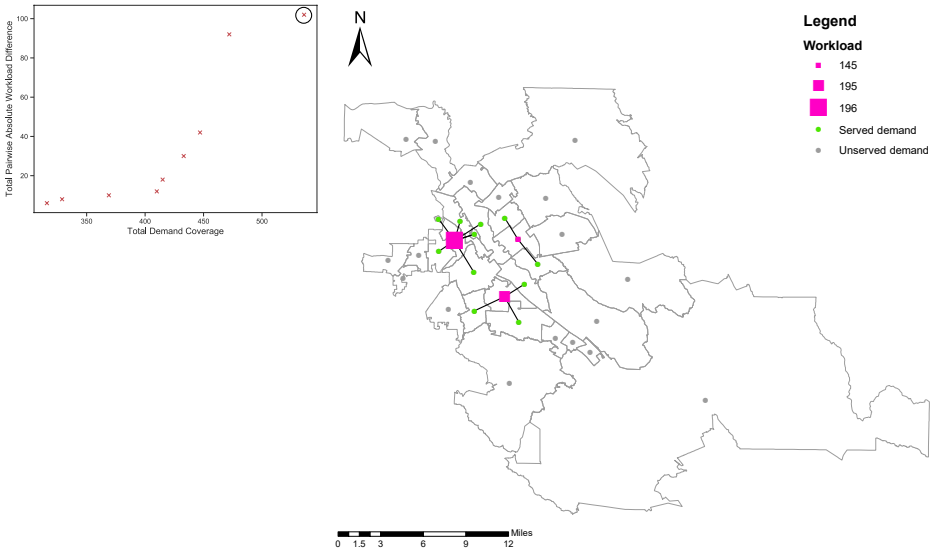

Figure 4.6 Workloads associated with an unbalanced location and allocation decision (San Jose postal service delivery, $p = 3$)

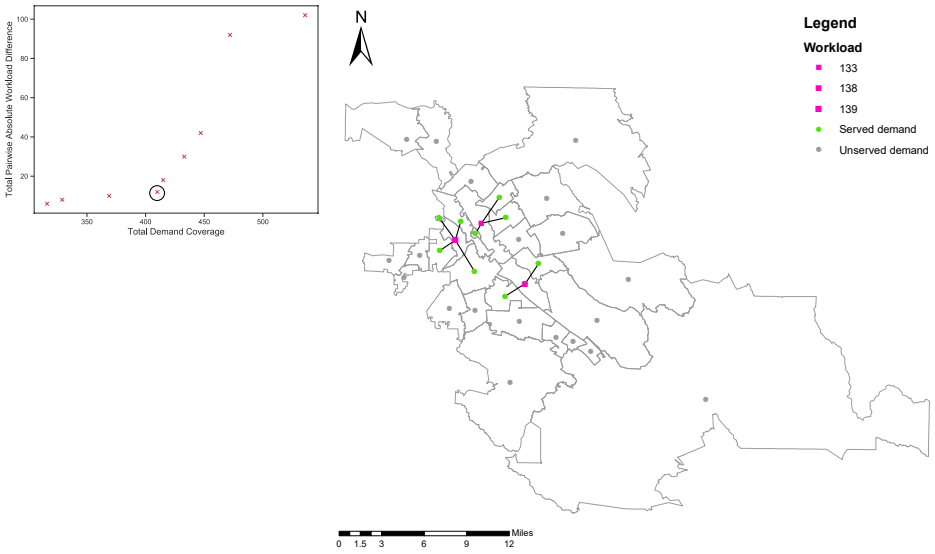

Figure 4.7 Workloads associated with a balanced location and allocation decision (San Jose postal service delivery, $p = 3$)

4.5.2 Santa Barbara Fire Response

The second case study investigates fire response to 80 block groups in downtown Santa Barbara, CA. The centroids of these areas are used to represent both demand and potential fire stations. The demand amount is based on population in the area. The local street network is used to construct the transport network (U.S. Census Bureau, 2018a) and calculate travel information between demand and potential stations. The service coverage standard is 1.5 miles. The number of fire stations to site (p) ranges from 2 to 6. Parameters are the same as those reported previously.

Table 4.2 gives the computational results when the heuristic is applied to Santa Barbara fire response. When p is 2 or 3, the heuristic is able to identify all Pareto optimal solutions. When p is 4, the heuristic finds 18 of 27 Pareto optimal solutions, with 5 dominated solutions, in 225.46 seconds. The trade-off curve between total demand coverage and total pairwise workload variation measure is given in Figure 4.8(a). The maximum coverage possible with four facilities is 719 hundred people, with a workload variation of 321 hundred people. When covering 632 hundred people, it is possible to reach a balanced system. The trade-off plot shows that solutions generated by the heuristic are very close to, if not coincide with, optimal non-dominated solutions. The average percentage gap for total demand coverage is 0.4% and workload variation is 1.5%. When siting six facilities, 12 out of 18 (66.7%) Pareto optimal solutions are found by the heuristic in 997 seconds compared to 1.5 hours by Gurobi. The average percentage gaps are 0.2% and 8.2% in total coverage and workload variation, respectively. The trade-off is illustrated in Figure 4.8(b). There are 18 Pareto optimal solutions, with total coverage ranging from 750 to 793 hundred people and workload variation ranging from 0 to 403 hundred people. Most of the solutions identified by the heuristic are or near optimal solutions. Overall, heuristic solutions are evenly distributed over the objective space, with little divergence from the Pareto front. Again, the computational time is much shorter for the heuristic compared to the exact approach.

Table 4.2 Computational results for proposed algorithm (Santa Barbara fire response, $S = 1.5$)

p	# of Pareto optimal solutions	# of Pareto optimal solutions found (Completeness %)	Maximum Gaps (%, %)	Average Gaps (%, %)	Solution Time - Algorithm (Seconds)	Solution Time - Gurobi (Seconds)
3	6	6 (100.0)	(0.0, 0.0)	(0.0, 0.0)	28.88	177.04
4	27	18 (66.7)	(1.0, 3.9)	(0.4, 1.5)	225.46	5,195.20
5	28	14 (50.0)	(1.5, 18.2)	(0.4, 4.1)	1,257.67	23,448.21
6	18	12 (66.7)	(0.3, 30.0)	(0.2, 8.2)	997.00	32,194.57

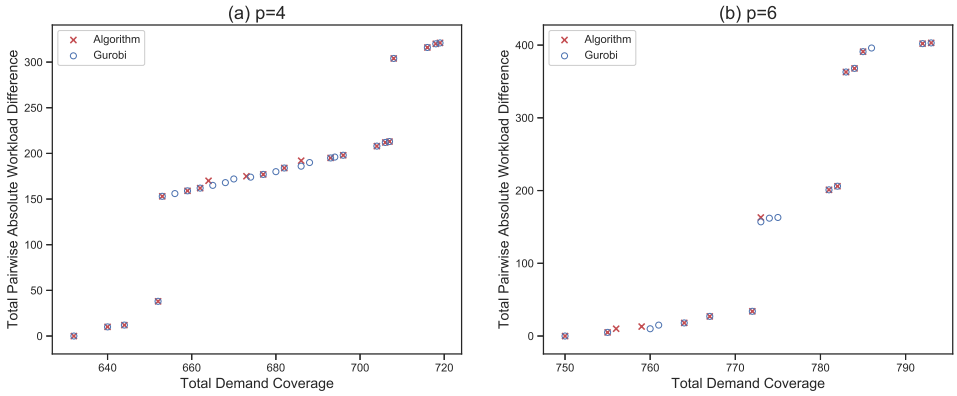

Figure 4.8 Pareto optimal front comparison (Santa Barbara fire response)

4.5.3 Boston Fire Response

The third case study investigates fire response in eight suburbs of Northwest Boston (Acton, Bedford, Carlisle, Concord, Lincoln, Maynard, Sudbury and Wayland). Murray and Tong (2009) provided context for this planning application. There are 511 reported

structure fires during 1990 to 2004, which are used to represent both demand and potential fire stations. The demand amount is one at each location. The local street network is used to construct the travel network and calculate travel information between demand and potential stations. The service coverage standard is 1.5 miles. The number of fire stations to site (p) ranges from 2 to 6. The initial temperature T is 10 and the minimum temperature T_{min} is 1. The remaining parameters are as defined above.

Table 4.3 summarizes computational results for this study. When p is 2, all 20 Pareto optimal solutions are found by the heuristic in 15 seconds compared to 29,241.13 seconds by the exact solver. When p is 3, the heuristic identifies 28 of 32 Pareto optimal solutions (completeness of 87.5%), with 2 inferior solutions. Solution time is about 33 seconds compared to around 11 days by the exact solver. Cases with larger p values were also attempted. Unfortunately, these problem instances cannot be solved optimally. For example, when $p \geq 4$, Gurobi was not successful after 2 weeks in verifying any optimal solutions. Since all Pareto optimal solutions could not be found, heuristic evaluation along these lines is not possible. However, the Pareto frontier identified by the heuristic is possible. When p is 4, the heuristic identified 35 solutions (shown in Figure 4.9) within 91 seconds. The maximum coverage is 226 with a workload variation of 282. The total pairwise workload difference can be reduced by 61.7% (i.e., to 108) by covering 2.6% less demand. A system with equivalent workloads can be obtained when covering 165 demand. For p equals 5 and 6, 42 and 35 solutions are identified by the heuristic, respectively.

Table 4.3 Computational results for proposed algorithm (Boston fire response, $S = 1.5$)

p	# of Pareto optimal solutions	# of Pareto optimal solutions found (Completeness %)	Maximum Gaps (%, %)	Average Gaps (%, %)	Solution Time - Algorithm (Seconds)	Solution Time - Gurobi (Seconds)[2]
2	20	20 (100.0)	(0.0, 0.0)	(0.0, 0.0)	15.27	29,241.13
3	32	28 (87.5)	(2.8, 2[1])	(1.6, 1.5[1])	32.98	989,708.29
4	–	35	–	–	90.77	–
5	–	42	–	–	92.82	–
6	–	35	–	–	184.03	–

[1] Absolute gaps are used due to zero denominator when computing percentage gaps.

[2] Unreported times indicate no feasible solution found after 2+ weeks of processing.

4.5.4 Santa Barbara Nutrition

The final case study considered involves the Special Supplemental Nutrition Program for Women, Infants, and Children in the larger Santa Barbara area (Goleta, Santa Barbara, and Carpinteria) (Xu et al., 2020). The program aims to provide healthcare, supplemental foods and nutrition education for eligible women, infants and children. There are 2,070 census blocks in the region, with a total population (2010) of 200,450 people. The population of each block is used as a proxy for service demand, represented using the centroid. In addition, there are 82 locations identified as potential facility sites. The road network is used to compute travel distance (U.S. Census Bureau, 2018a). A service

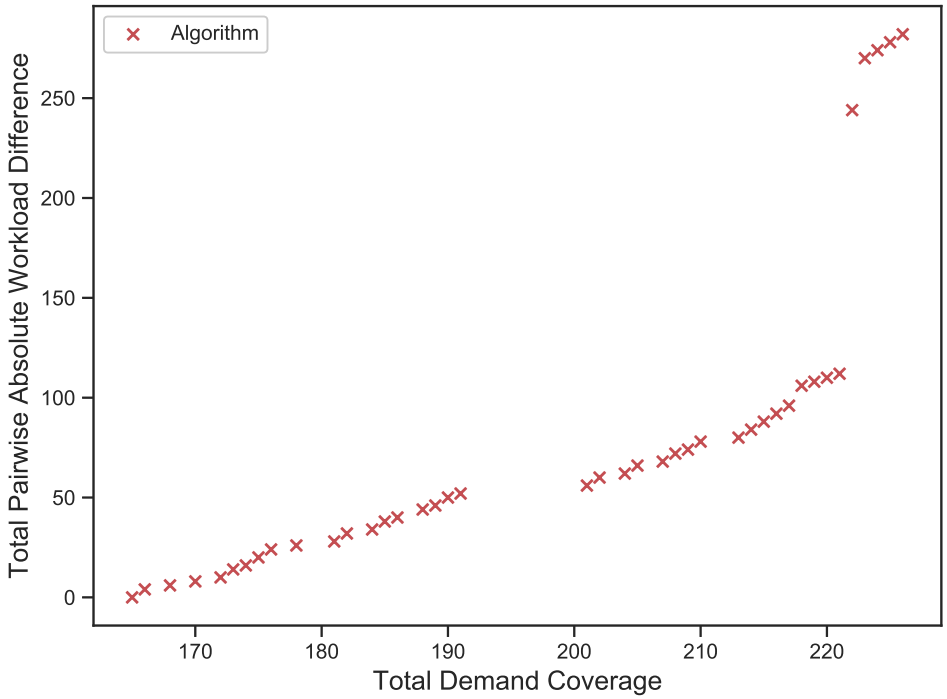

Figure 4.9 Non-dominated solutions identified by heuristic algorithm (Boston fire response, $p = 5$)

coverage standard of 5 miles is assumed. Siting 2 to 8 facilities while considering workload balance is investigated. The initial temperature T is 300, the minimal temperature T_{min} is 10 and the cooling parameter β is 0.9. The equilibrium state threshold E is 10% of the neighborhood size. The probability of re-allocating a single demand p_1 is 50%.

The computational results are summarized in Table 4.4. The coverage set N_i (or R_j) is very large due to the 5 mile coverage standard in this area. On average, a facility can cover 805 demand areas, which amounts to 74,056 people on average. Therefore, there are numerous demand allocation possibilities, which make associated problem instances very

challenging to solve. Unfortunately, no problem instances can be solved completely using the exact solver within 2 weeks, even for only two sited facilities. When p is 2, there are 34 solutions identified as non-dominated solutions by the heuristic. The maximum coverage is 174,432 people with a total pairwise workload variation of 30,318. When covering 159,122, the workloads of the two sited facilities are equal. When p is 5, the heuristic identifies 58 non-dominated solutions (Figure 4.10 inset), taking 532 seconds to solve. The maximum demand coverage possible is 200,340 people, with an associated total pairwise workload variation of 218,190 people. By sacrificing 0.2% population coverage, workload variation can be significantly reduced to 123,172 people (by about 44%). Covering 8.6% less population could further decrease the workload variation to 3,968 people, as shown in Figure 4.10. A system with five balanced facility workloads can serve 148,115 people. Each sited facility has a workload of 29,623 people. Finally, for p equals to 8, the heuristic identifies 56 non-dominated solutions in 1,430.85 seconds.

4.6 Discussion

There are a number of items worth further discussion. One is the importance of the initial hybrid facility siting approach. A second one is the performance of the simulated annealing algorithm on solving the allocation problem (4.14)-(4.18). The last one is the sensitivity of heuristic performance to different service coverage standards.

Table 4.4 Computational results for proposed algorithm (Santa Barbara nutrition, $S = 5$)

p	# of solutions identified	Solution Time - Algorithm (Seconds)	Solution Time - Gurobi (Seconds)[1]
3	44	40.89	–
4	69	132.84	–
5	58	532.26	–
6	111	870.58	–
7	73	1,179.66	–
8	56	1,430.84	–

[1] Unreported times indicate no feasible solution found after 2+ weeks of processing.

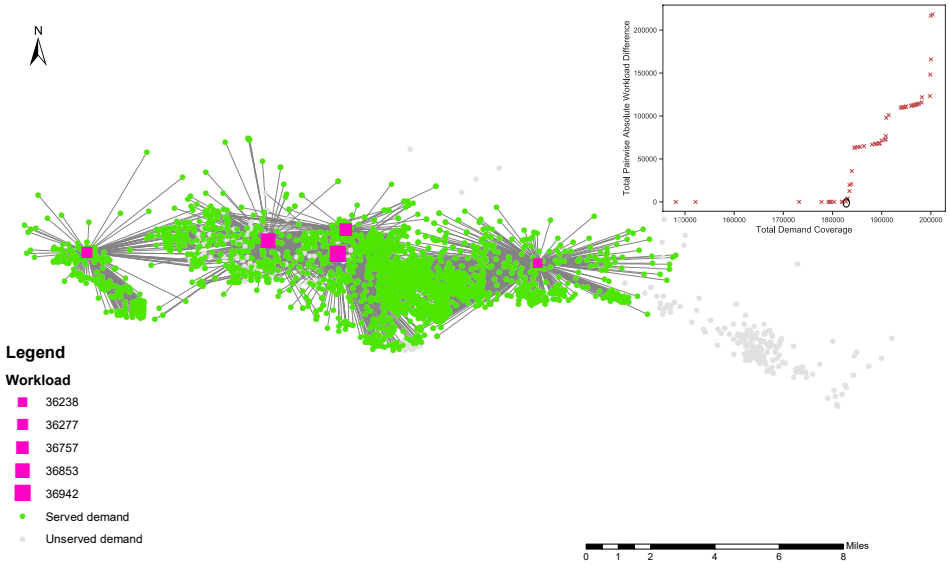

Figure 4.10 Workloads associated with a balanced location and allocation decision (Santa Barbara nutrition, $p = 5$)

4.6.1 Hybrid vs Random Siting

The initial hybrid siting approach (Figure 4.2) is critical to the success of the algorithm. To demonstrate this point, a random initialization is compared with the hybrid procedure. The random initialization generates the same amount of p-facility configurations randomly. Other steps and parameters of the algorithm remain the same. Table 4.5 summarizes the comparison between the hybrid procedure and the best results among five different runs with the random initialization using San Jose data. Results show that the hybrid initialization is able to find more optimal solutions compared to the random initialization. For example, when p is 7, the random initialization finds a maximum of 6 out of 13 Pareto optimal solutions among five different runs while the hybrid initialization finds 11 Pareto optimal solutions. Figure 4.11 shows the best non-dominated solutions identified by the algorithm with a random initialization and those by the algorithm with the hybrid initialization when p is 7. Although the random initialization provides good convergence in the middle part of the front, its solutions are not well uniformly distributed and miss optimal solutions with high and low demand coverage (or workload variation). It is also evident that solutions by the random initialization are dominated by the hybrid initialization in Figure 4.11. The better performance of the hybrid procedure is due to the consideration of solutions with extreme objective values and the greedy steps that enable a complete exploration of the objective space.

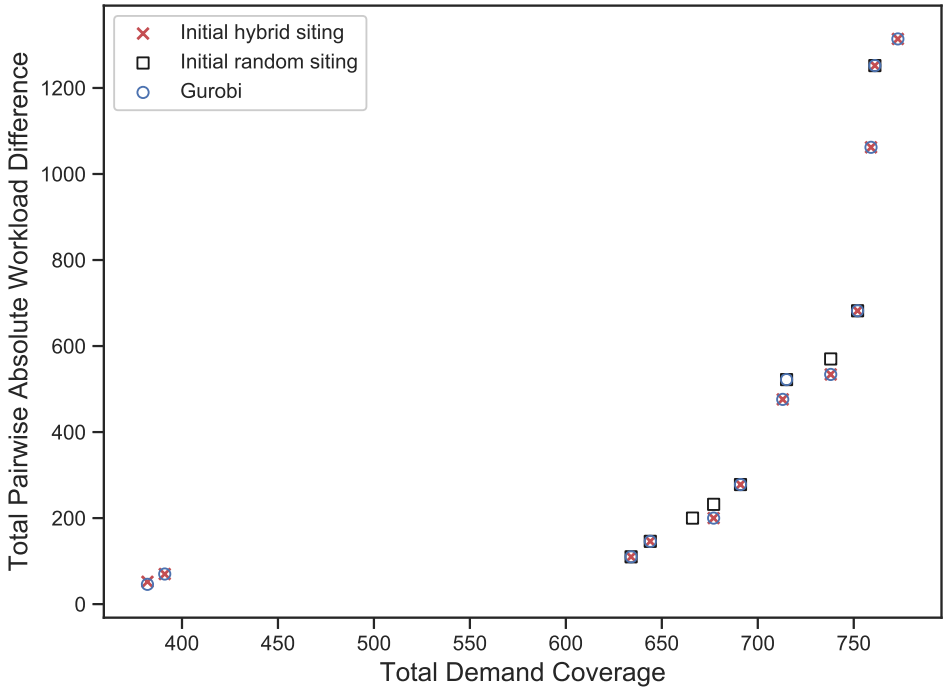

Figure 4.11 Non-dominated solutions using hybrid facility initialization vs random facility initialization (San Jose postal service delivery, $p = 7$)

Table 4.5 Computational results comparing initial hybrid siting vs random siting (San Jose, $S = 3$)

p	# of Pareto optimal solutions	# of Pareto optimal solutions found (Completeness %)	
		Hybrid	**Random[1]**
2	2	2 (100.0)	1 (50.0)
3	9	9 (100.0)	5 (66.7)
4	9	9 (100.0)	7 (77.8)
5	20	17 (85.0)	11 (55.5)
6	19	17 (89.5)	15 (78.9)
7	13	11 (84.6)	6 (46.2)
8	16	12 (75.0)	5 (37.5)
9	25	18 (72.0)	11 (44.0)
10	28	22 (78.6)	20 (71.4)
11	21	15 (71.4)	9 (42.9)
12	21	16 (76.2)	12 (57.1)
13	22	16 (72.7)	13 (59.1)
14	19	15 (78.9)	12 (63.2)
15	20	16 (80.0)	15 (75.0)

[1] Best results among five runs with random initialization.

4.6.2 Strength of Simulated Annealing Based Allocation

To evaluate the performance of the simulated annealing algorithm on solving the allocation problem, (4.14)-(4.18), it is compared with the initial greedy allocation (as explained in Section 4.4.2) and the Gurobi solver for a set of allocation problems with randomly generated facility configurations. The total pairwise workload variation (i.e., the objective value of the allocation problem) and computational time are compared. The gap of the workload variation is computed as the percentage deviation from the objective value

found by Gurobi. Computational results are summarized in Tables 4.6 and 4.7. Out of all 9 random instances (when $p = 3, 10, 15$) using the relatively easier San Jose case (Table 4.6), simulated annealing is able to find optimal solutions in a similar amount of time with Gurobi. While the greedy algorithm miss the optimal solution in 4 cases and the gap in the objective value can be as large as 125.6% though it is very fast. Out of the 20 random instances (when $p = 5, 8$) using the large Santa Barbara nutrition case (Table 4.7), simulated annealing is able to find the same solutions as Gurobi does in 8 cases and its generated objective values are usually within 2% from Gurobi results in cases that solutions do not coincide. This is done in a comparable amount of computational time, compared to Gurobi. It is noted that there are 5 instances that Gurobi is not able to converge the gaps between the upper and lower bounds for the objective value after 1 hour of processing, which highlights the inefficiency of solving this NP-hard allocation problem using the exact method. The greedy algorithm runs very fast but generates poor-quality solutions that have objective values much larger than Gurobi results. These results show the necessity and strength of the simulated annealing algorithm for solving the allocation problem when the sited facilities are fixed for the WBMCLP-TotPairDiff.

Table 4.6 Computational results of solving the allocation problem: simulated annealing vs greedy algorithm vs Gurobi (San Jose postal service, $S = 3$)

p	Sited facilities	Simulated Annealing		Greedy		Gurobi	
		Workload Variation (Gap%)	Time (Seconds)	Workload Variation (Gap%)	Time (Seconds)	Workload Variation	Time (Seconds)
3	[3, 17, 18, 22, 23]	78 (0.0)	0.05	176 (125.6)	0.00	78	0.06
	[3, 12, 18, 22, 29]	278 (0.0)	0.00	278 (0.0)	0.00	278	0.02
	[3, 8, 10, 18, 29]	764 (0.0)	0.00	764 (0.0)	0.00	764	0.02
10	[4, 8, 11, 14, 19, 21, 24, 28, 29, 30]	356 (0.0)	0.11	702 (97.2)	0.00	356	0.08
	[3, 7, 8, 10, 14, 17, 18, 22, 23, 30]	748 (0.0)	0.14	748 (0.0)	0.00	748	0.12
	[2, 3, 7, 8, 10, 13, 18, 22, 25, 29]	2,790 (0.0)	0.00	2,790 (0.0)	0.00	2,790	0.02
15	[3, 4, 7, 8, 13, 14, 17, 18, 19, 22, 23, 24, 25, 28, 32]	966 (0.0)	0.20	1,724 (78.5)	0.00	966	0.17
	[2, 3, 4, 7, 8, 11, 12, 13, 14, 18, 21, 22, 23, 25, 31]	1,676 (0.0)	0.26	1,988 (18.6)	0.00	1,676	0.29
	[1, 2, 3, 6, 7, 8, 10, 13, 15, 17, 18, 20, 22, 25, 29]	5,326 (0.0)	0.08	5,368 (0.8)	0.00	5,326	0.08

Table 4.7 Computational results of solving the allocation problem: simulated annealing vs greedy algorithm vs Gurobi (Santa Barbara nutrition)

p	Sited facilities	Simulated Annealing		Greedy		Gurobi	
		Workload Variation (Gap%)	Time (Seconds)	Workload Variation (Gap%)	Time (Seconds)	Workload Variation	Time (Seconds)
5	[33, 26, 43, 8, 1]	153,052 (0.0)	1.20	153,052 (0.0)	0.37	153,052	0.13
	[63, 82, 16, 69, 28]	6 (0.0)	36.98	5,678 (94,533.3)	0.50	6[1]	>3,600
	[18, 74, 39, 65, 30]	3,484 (0.0)	1.13	3,484 (0.0)	0.56	3,484	0.37
	[54, 12, 43, 53, 80]	103,666 (1.5)	1.40	108,748 (6.5)	0.47	102,140	0.40
	[74, 27, 53, 72, 39]	48,604 (0.3)	2.47	68,600 (41.5)	0.54	48,476	0.32
	[46, 60, 47, 68, 20]	4 (0.0)	0.97	4 (0.0)	0.48	4[1]	>3,600
	[76, 75, 64, 31, 1]	182,150 (0.1)	3.70	202,986 (11.6)	0.14	181,902	0.12
	[1, 81, 55, 23, 26]	91,510 (0.1)	1.18	93,050 (1.7)	0.47	91,450	0.09
	[4, 60, 7, 21, 72]	312,784 (0.01)	0.92	326,154 (4.3)	0.15	312,748	0.03
	[32, 30, 41, 38, 44]	4 (0.0)	1.05	4 (0.0)	0.38	4[1]	>3,600

[1] The best objective value found when the objective value gap is not converged after 1 hour of processing.

Table 4.7 (continued)

p	Sited facilities	Simulated Annealing		Greedy		Gurobi	
		Workload Variation (Gap%)	Time (Seconds)	Workload Variation (Gap%)	Time (Seconds)	Workload Variation	Time (Seconds)
8	[15, 51, 35, 17, 11, 10, 82, 26]	379,680 (1.1)	2.88	397,528 (5.9)	0.40	375,402	0.59
	[48, 69, 4, 23, 8, 70, 36, 78]	279,754 (1.8)	2.97	294,832 (7.3)	0.63	274,834	0.29
	[70, 81, 44, 72, 53, 17, 76, 6]	251,323 (0.3)	1.93	265,383 (5.9)	0.43	250,641	0.25
	[17, 11, 41, 58, 53, 64, 13, 79]	214,011 (0.0)	1.83	227,039 (6.1)	0.59	214,011	4.64
	[15, 46, 78, 26, 25, 24, 50, 22]	16 (0.0)	0.78	16 (0.0)	0.50	16[1]	>3600
	[76, 5, 61, 64, 31, 54, 13, 63]	103,212 (0.2)	5.89	130,042 (26.2)	0.58	103,026	2.69
	[47, 6, 69, 74, 38, 21, 66, 77]	120,378 (1.8)	6.97	153,852 (30.1)	0.56	118,216	0.52
	[82, 62, 59, 80, 69, 72, 15, 76]	124,692 (0.1)	4.77	146,602 (17.6)	0.41	124,622	0.32
	[1, 27, 2, 29, 8, 42, 33, 31]	267,849 (0.0)	2.21	267,849 (0.0)	0.41	267,849	0.58
	[78, 73, 82, 34, 62, 66, 19, 52]	41 (487.5)	4.42	7,565 (107,971.40)	0.67	7[1]	>3,600

[1] The best objective value found when the objective value gap is not converged after 1 hour of processing.

4.6.3 Sensitivity to Coverage Standard S

The service coverage standard S affects the complexity of a workload balancing problem instance. A large S would usually increase the number of facilities that can suitably cover each demand (i.e., $|N_i|$) and the number of demand that can be covered by each potential facility (i.e., $|R_j|$) as well. Thus, a large S tends to increase facility location and demand allocation possibilities, leading to a larger solution space for search. Therefore, the service coverage standard S is adjusted to 5 miles and 7 miles in San Jose study to test how sensitive the algorithm performance is to S. Tables 4.8 and 4.9 summarize computational results when S is 5 miles and 7 miles respectively. The number of facilities to be located, p, ranged from 2 to the maximum amount that optimal results could be obtained by an exact solver within 2 weeks of computational time. When S is 5 miles and p ranges from 2 to 10, 57%-100% Pareto optimal solutions are found by the proposed algorithm. More solutions are produced that are inferior solutions compared to when S is 3 miles but optimality gaps remain small for most instances. When S is 7 miles and p ranges from 2 to 5, the overall results are comparable to finding a Pareto front when S is 3 or 5 miles. Therefore, solution quality is not degraded significantly. However, the computational time of the exact solver significantly increases with larger S, which further demonstrates the computational complexity of the problem and the necessity of a heuristic algorithm. For example, it takes the exact solver about 203,162 seconds (56 hours) to obtain all Pareto optimal solutions when S is 5 miles and p is 10 compared to

1,270 seconds when S is 3 miles. In contrast, the solution time when using the proposed algorithm is still acceptable when S is large enough in the study area. It takes the algorithm about 2,935 seconds when S is 5 miles and p is 10.

Table 4.8 Computational results for proposed algorithm (San Jose, $S = 5$)

p	# of Pareto optimal solutions	# of Pareto optimal solutions found (Completeness %)	Maximum Gaps (%, %)	Average Gaps (%, %)	Solution Time - Algorithm (Seconds)	Solution Time - Gurobi (Seconds)
2	2	2 (100.0)	(0.0, 0.0)	(0.0, 0.0)	0.02	2.24
3	7	6 (85.7)	(0.0, 10.0)	(0.0, 10.0)	2.01	24.81
4	7	6 (85.7)	(2.0, 57.1)	(2.0, 57.1)	8.50	122.45
5	5	3 (60.0)	(8.1, 150.0[1])	(4.5, 77.1)	9.72	275.52
6	8	7 (87.5)	(0.0, 0.7)	(0.0, 0.7)	48.67	1,430.35
7	12	7 (58.3)	(10.9, 166.7[2])	(6.6, 43.7)	123.94	14,407.43
8	14	8 (57.1)	(28.6, 3.5)	(5.7, 1.1)	537.63	73,069.99
9	15	9 (60.0)	(4.4, 15.9)	(1.1, 3.1)	1,712.83	58,880.03
10	15	10 (66.7)	(1.3, 2.0)	(0.3, 0.6)	2,935.10	203,162.82

[1] The missed Pareto optimal solution has a workload variation of 4 and its closet solution by the algorithm has a workload variation of 10; (10-4)/4 leads to the 150% gap in workload variation.

[2] The missed Pareto optimal solution has a workload variation of 12 and its closet solution by the algorithm has a workload variation of 32; (32-12)/12 leads to the 166.7% gap in workload variation.

4.7 Conclusion

This chapter proposed a heuristic algorithm for solving the workload balancing MCLP, which is a bi-objective optimization problem that maximizes demand coverage and mini-

Table 4.9 Computational results for proposed algorithm (San Jose, $S = 7$)

p	# of Pareto optimal solutions	# of Pareto optimal solutions found (Completeness %)	Maximum Gaps (%, %)	Average Gaps (%, %)	Solution Time - Algorithm (Seconds)	Solution Time - Gurobi (Seconds)
2	1	1 (100.0)	(0.0, 0.0)	(0.0, 0.0)	0.19	0.31
3	6	6 (100.0)	(0.0, 0.0)	(0.0, 0.0)	0.77	41.25
4	12	10 (83.3)	(0.2, 0.3)	(0.1, 0.2)	54.10	505.85
5	20	16 (80.0)	(1.6, 2.6)	(0.9, 1.3)	253.94	260,919.42

mizes total pairwise absolute workload difference simultaneously. The proposed heuristic algorithm has three major stages. In the initialization stage, the heuristic identifies a set of initial facility siting solutions using a hybrid approach. In the subproblem delineation stage, there is a focus on allocation to minimize workload variation while ensuring total demand coverage is no less than a pre-defined coverage bound. With an initial facility configuration and a subproblem allocation, a local search stage follows. Facility siting is updated by substituting facilities that are not sited for sited facilities. Demand is allocated to sited facilities using a simulated annealing approach that incorporates two types of neighbor solution search approaches. Empirical studies involving four applications demonstrated that the heuristic was able to identify non-dominated solutions that approximate the Pareto optimal front well, yet required significantly less computational resources. Solutions identified by the heuristic were well distributed in the objective

space, covering extreme objective values. Finally, the heuristic proved capable of identifying solutions for larger problem instances for which the exact solver was not successful.

Chapter 5

Conclusions

This dissertation evaluated prominent modeling approaches to govern facility workloads in coverage problems. These approaches were explored and compared in different ways, focusing on resulting workloads and qualitative considerations. Solution approaches were developed and applied, with an efficient heuristic algorithm identified for obtaining trade-off solutions.

This chapter offers concluding comments to the dissertation. A summary of each chapter is detailed in the next section. This is followed by a discussion of theoretical contributions. The chapter ends with suggestions for future research.

5.1 Summary

Chapter 1 offered motivation for this research work, identified key problems, reviewed primary objectives, and spoke to the significance of various aspects of associated modeling. Further, an overview of this dissertation was provided.

Chapter 2 evaluated the CMCLP, the prominent method to govern workloads in coverage problems, from the perspective of workload balancing. This was done with respect to both solution characteristics and allocation response. First, this chapter showed why workloads could be balanced using the CMCLP from a theoretical and mathematical perspective. Second, the solution quality and computational time to solve the CMCLP using GIS packages were evaluated against the exact approach, recognizing that the CMCLP is accessible in GIS and increasingly relied upon in planning applications. GIS provides an integrated environment for data acquisition, management, manipulation, analysis, and display; and offers easy access to modeling for general users. Two empirical studies that involved 180 problem instances were carried out. Although the heuristics applied in GIS packages solved capacitated problems in less computational effort, some 63% of problems could not be solved optimally. The observed maximum optimality gap was approximately 10%. Third, service implications due to the use of capacities were highlighted. Some demand may not be allocated even if it is located within the service standard. In addition, some demand may be allocated to a further away facility. This means that a closer sited facility exists, yet capacity limitations dictate that some demand must seek service from a different facility. With all these limitations, approaches that can better balance facility workloads in coverage problems should be explored.

Chapter 3 proposed five workload balancing approaches in the context of maximal covering. This chapter first analyzed different equity measures including the total pairwise absolute workload difference, total mean absolute workload deviation, maximum

mean absolute workload deviation, workload range, and maximum workload. It proved that the total pairwise absolute workload difference captures the workload variation most explicitly. As a result, other measures can be regarded as an approximation to pairwise workload difference. The five different workload balancing maximal covering models incorporating equity measures were formulated through the use of an additional objective and associated constraints. Evaluation was conducted by treating the workload balancing model with the total pairwise measure as the benchmark, against which other models are compared. The empirical results demonstrated that it was very likely that approximation approaches would be suboptimal, often with significant optimality gaps. This suboptimality was especially prevalent when the number of facilities to site becomes large. Although the model with the total pairwise measure was able to identify optimal trade-offs, it required significantly more computational effort to be solved optimally. This called for the development of more efficient solution techniques to address workload balancing issues.

Chapter 4 proposed a heuristic algorithm for solving the workload balancing MCLP as a bi-objective optimization problem that maximizes demand coverage and minimizes total pairwise absolute workload difference simultaneously. The proposed algorithm starts with an initialization stage where a hybrid siting approach was designed to obtain facility configurations. Then, the algorithm splits the bi-objective problem into subproblems using the constraint method. Next, a local search stage was proposed to search for the best location-allocation solution for each subproblem. Facility siting is updated by

substituting a facility that is not sited with a sited facility, followed by demand allocation via simulated annealing. Evaluation with four different empirical datasets showed that the heuristic could find a set of solutions that approximates the Pareto optimal front well, but did so with much less computational effort. In addition, the algorithm was able to derive good-quality solutions when the exact method failed.

5.2 Theoretical Contributions

Facility workload balancing is an important topic in location problem. However, it has been ignored or not appropriately handled in covering problems. This dissertation provided a systematic evaluation of prominent approaches to control facility workloads in coverage problems, highlighting the implications of applying capacities in practice. This ultimately offered insights on opportunities for better approaches. This work enables a technical and practical understanding of different equity measures to control facility workloads. As a result, new modeling approaches that explicitly balance workloads were possible. The proposed models are not directed to any specific application, but rather are expected to have many potential applications. In addition, recognizing computational difficulty, this dissertation designed a heuristic method for obtaining solutions to the proposed workload balancing model. It proved to be efficient, making the model applicable for supporting planning.

5.2.1 Better Understanding of Capacitated Methods

The first major contribution of the dissertation is a better understanding of capacitated methods. The prominent way to control facility workloads in coverage problem is to introduce facility capacity c_j. Along these lines, the CMCLP was structured by imposing a constraint $\Sigma_{i \in I} a_i Y_{ij} \leq c_j X_j$ for each potential facility j in the context of maximal covering (MCLP). The MCLP can be regarded as a special case of CMCLP with $c_j = \infty$ for all j. The introduction of facility capacity has two major benefits. First, this relaxes the unrealistic assumption that a facility can serve an unlimited amount of demand within its service coverage standard. Second, from the perspective of workload balancing, facility capacities help to balance facility workloads to some extent. It has been shown that the upper bound of the workload range of a CMCLP solution is $UB^{CMCLP} = \max_j\{\min\{c_j, \Sigma_{i \in R_j} a_i\}\} - \min_j\{\min_{i \in R_j} a_i\}$ while the upper bound of the workload range of a corresponding optimal MCLP solution is $UB^{MCLP} = max_j\{\Sigma_{i \in R_j} a_i\} - \min_j\{\min_{i \in R_j} a_i\}$ by substituting $c_j = \infty$ for the MCLP, where R_j is the set of demand that is within the service coverage standard of facility j. Therefore, the workload range of an optimal CMCLP solution, UB^{CMCLP}, is bounded by the corresponding workload range of an optimal MCLP, UB^{MCLP}, that is $UB^{CMCLP} \leq UB^{MCLP}$. In another words, the CMCLP solution tends to be more balanced compared to a corresponding MCLP solution due to the introduction of facility capacities. In addition, the workload variation of an optimal CMCLP solution really de-

pends on the specific setup of facility capacities and the spatial configuration of regional demand.

However, there are several limitations associated with the use of capacities. The first is that most approaches for specifying appropriate facility capacities are subjective and require expert knowledge and/or historical information. Second, a facility's capacity is not necessarily a strict limit, so there may be some degree of flexibility. Thus, the use of strict capacities may introduce uncertainty, negatively impacting service system effectiveness and efficiency. Third, the facility capacity could fail to reflect the actual workloads that a facility undertakes. In the context where the service provider lacks control in preventing people from accessing services, especially in the public sector, facility capacity could be exceeded in practice.

Another challenge is that the addition of capacities often significantly increases computational processing in model solution, sometimes beyond computing capabilities. Assume sited facilities are known, and let $J^* = \{j|X_j = 1\}$ be the set of sited facilities. For demand that is beyond the service coverage standard of any sited facilities, i.e., $\{i|d_{ij} > S \quad \forall j \in J^*\}$, and un-sited facilities $\{j|j \notin J^*\}$, their associated $Y_{ij} = 0$. Then for the remaining demand and sited facilities, the CMCLP becomes an allocation problem that seeks to maximize the total service coverage while not exceeding facility capacities. For a certain facility $j \in J^*$, the allocation problem is essentially a 0-1 knapsack problem that is known as NP-hard. Because the capacity constraint $\Sigma_{i \in R_j} a_i Y_{ij} \leq c_j$ for a sited facility j would very likely lead to a fractional Y_{ij} in order to satisfy the constraint and

maximize total coverage. This can make the search explore deeply into the branch-and-bound tree before finding an integer solution for the 0-1 knapsack problem. Thus, the introduction of capacity significantly increases the computational complexity. Empirical experiments also demonstrated that substantial computational resources were needed for solving the CMCLP optimally.

The final limitation is the use of capacities can result in undesirable allocation response. A demand could be denied service even if it is within the service coverage standard from a sited facility. The withholding of service will happen if no workload capacity is available from a sited facility that can suitably serve the demand. However, it could be problematic in practice to restrict people from service when they are spatially close to a service provider. Another possible allocation response is that a demand may be dispatched to a further away facility when allocation to a closer sited facility violates the capacity limit and there is sufficient capacity to accommodate it elsewhere. Cases with such undesirable allocation response were observed in empirical studies. Although constraints that ensures mandatory assignment and/or closest assignment could be imposed, they might make the problem infeasible due to strict facility capacities.

5.2.2 Explicit Modeling Approaches for Addressing Workload Balancing

The second major contribution is the proposal of explicit modeling approaches that better address workload balancing in coverage problems. Five workload variation measures that are analytically tractable were formulated and compared mathematically. Denote W_j as the workload of facility j, $\bar{W}$ the average workload of sited facilities. It is found that the total pairwise absolute workload difference $\Sigma_{j\in J^*}\Sigma_{j'\in J^*}|W_j - W_{j'}|$ can be expanded as $\Sigma_{j\in J^*}(|\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_1}| + |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_2}| + \cdots + |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij}| + \cdots + |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_p}|)$ which tracks the workload variation most explicitly among the five measured studied. The total mean absolute workload deviation $\Sigma_{j\in J^*}|W_j - \bar{W}|$ can be re-written as $\frac{1}{p}\Sigma_{j\in J^*}|(\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_1}) + (\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_2}) + \cdots + (\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij}) + \cdots + (\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij_p})|$. Thus, the total mean absolute workload variation overlooks the inherent workload difference and can be regarded as an approximation to the total pairwise measure. In addition, the total mean absolute workload deviation is no greater than the total pairwise absolute workload difference divided by p according to the triangle inequality. They are equivalent only when $p = 2$ or workloads of sited facilities are all equal. The third measure, maximum mean absolute deviation $\max_{j\in J^*}|W_j - \bar{W}|$, is a portion of the total mean absolute deviation, thus, also an approximated approach. The workload range $\max_{j\in J^*} W_j - \min_{j\in J^*} W_j$ can be re-written as $\max_{j\in J^*}\max_{j'\in J^*}\{|\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|\}$, comparing pairwise workload difference but only tracking the maximum difference. So

the workload range is a simplified way compared to the total pairwise measure as well as the maximum workload $\max_{j \in J^*} W_j$.

With each workload variation measure, the dissertation proposed a bi-objective optimization model that maximizes the total coverage and minimizes the workload variation. Take the WBMCLP-TotPairDiff as an example. Denote $D_{jj'}$ the absolute workload difference between any two potentially sited facilities j and $j^{'}$. An additional objective and two sets of constraints were added to the MCLP. The second objective is to minimize the total pairwise absolute workload difference, $\Sigma_j \Sigma_{j'>j} D_{jj'}$. Added constraints are $\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'} - M(1 - X_{j'}) \leq D_{jj'}$ and $\Sigma_i a_i Y_{ij'} - \Sigma_i a_i Y_{ij} - M(1 - X_j) \leq D_{jj'}$ for all facilities j and $j^{'} > j$ where M is a very large positive number. These constraints result in $D_{jj'} \geq |\Sigma_i a_i Y_{ij} - \Sigma_i a_i Y_{ij'}|$ for pairs of sited facilities and $D_{jj'} \geq 0$ for other paired outcomes, hence only tracking pairwise absolute workload difference between sited facilities. Similarly, other four workload balancing models were formulated. In addition, constraints were also formulated to avoid the potential withholding service. Consider an additional decision variable, C_i, will be 1 if demand i is within the service coverage standard of a sited facility and 0 otherwise. Forcing assignment constraints are $\Sigma_{j \in N_i} X_j \leq p C_i$ and $C_i \leq \Sigma_{j \in N_i} Y_{ij}$ for all demand i. Also, three quantitative evaluation measures, completeness, inferiority and maximum gap, and an evaluation procedure were proposed to compare bi-objective problems in different objective spaces. Compared to traditional capacitated methods, the proposed modeling approaches have three advantages: a) more explicitly balance workloads; b) provide a complete trade-off between

coverage and workload balancing objectives, to better assist the decision making process; and c) can avoid withholding service allocation response.

5.2.3 Efficient Heuristic Solution Method

The third major contribution is an efficient heuristic solution method for workload balancing in coverage problems, making the new approach applicable in practice. The heuristic algorithm is comprised of three major components: initialization, subproblem delineation and local search. A hybrid initialization procedure was proposed to generate good quality starting solutions. Different from a random initialization, this hybrid method produces solutions that relatively uniformly cover the objective space and includes solutions with extreme objective values. This is particularly important for solving multi-objective problems because the ultimate goal is to derive the complete Pareto optimal front. The algorithm relaxes the workload balancing model and solves the resulting MCLP for solutions with large demand coverage; and seeks for relatively balanced facility configurations by a spatial search for solutions with small workload variation. Empirical evaluation has shown that the hybrid initialization outperformed the random method.

The heuristic method splits the problem into a group of single-objective problems by using the constraint method. Then for each subproblem, facility locations are updated by one-to-one facility substitution and demand allocation is achieved using a simulated annealing procedure. The simulated annealing based allocation procedure was designed because the WBMCLP-TotPairDiff proved to be NP-hard. Specifically, a PARTITION

problem that is known as NP-hard can be reduced to the WBMCLP-TotPairDiff when sited facilities are known. A key component in the simulated annealing approach is the neighbor solution search. This dissertation proposed a way that combines two types of demand re-allocation: single demand re-allocation and one-to-one exchange. The first one re-allocates a demand to another suitable sited facility, which is very flexible but may increase workload variation for a relatively balanced solution. The one-to-one allocation exchange, on the other hand, searches for two suitable pairs of facility-demand and swaps the allocation, which is more likely to generate a more balanced solution. Both are done by a spatial search to ensure the service coverage standard is not violated. The designed simulated annealing demand allocation uses a 50/50 combination of them, which has been shown more effective than the use of a single re-allocation approach or other combinations. The strength of the designed allocation method over a greedy approach and the exact solver was supported by evaluation results. Finally, through four empirical studies with real-world data, the proposed heuristic algorithm proved capable of deriving good quality solutions for small and medium size problems in a more efficient way; and identifying solutions for larger problem instances for which the exact solver was not successful.

5.3 Future Work

Four limitations and opportunities for future research are summarized as follows.

First, this research studied and evaluated five equity measures including the total pairwise absolute workload difference, the total mean absolute workload deviation, the maximum mean absolute workload deviation, the workload range and the maximum workload, and finally used the model with the total pairwise workload difference. These five measures are selected because they were frequently considered in location modeling literature and they are analytically tractable. And the model with the total pairwise measure was found to be the most explicit one balancing workloads. However, neither scale invariance nor the principle of transfers is satisfied in the total pairwise measure (Erkut, 1993; Eiselt and Laporte, 1995), thus may limit the usefulness of this measure. One potential future work direction is to structure alternative measures that are more complex in the modeling. This would require further evaluation as well, particularly with respect to empirical performance.

Second, this research considers only single source problems where allocation decision variables are binary, and fractional assignment is not allowed. The underlying assumption is that a demand unit can only be connected to one sited facility and this is suitable in many application contexts, such as postal delivery service, school districting, customer allocation in telecommunication, and others. However, there could be cases that a demand can be allocated to more than one sited facility and single source problems would not be appropriate. Also, the non-single source problem is generally less difficult to solve compare to the single source problem, because allocation decision variables are allowed to be continuous. It is also expected that there is more room to balance facility workloads

when fractional assignment is permitted. Therefore, one potential direction for the future work is to study the workload balancing maximal covering model that allows fractional assignment.

Third, an assumption here is no more than one facility is allowed to be sited at a location. Think about a case: there are two potential facilities A and B, A is located in an area with a large amount of demand within its service standard while B has few demand within its coverage. In this case, we should not expect or make the workload of these two facilities the same, because A can be a good candidate of a "super" service provider (e.g., a supermarket) that has larger capacity and can serve more demand. One way to incorporate this is to extend the workload balancing model by allowing multiple units a location, that is co-location. Co-location allows the facility workload to expand in area with large demand amount and reflects planning strategies in practice (Gerrard, 1995). If co-location is allowed, then the equity measure should be re-designed to reflect the workload variation appropriately, decision variables and some constraints need to be updated too. Also, it is expected that the model with co-location will be computationally harder to solve compared to the proposed model in this work, and efficient solution approaches may need to be developed.

Last, other efficient solution approaches could be explored for the proposed workload balancing MCLP. This dissertation designed a heuristic that delineates the bi-objective problem into a set of single-objective subproblems using the constraint method, then updates facility siting by one-to-one facility substitution and allocates demand by a

simulated annealing approach. Heuristics, such as genetic algorithm, have been found to be suitable for solving multi-objective problems, and therefore may prove valuable. As for the demand allocation, other heuristics, such as tabu search, GRASP, etc., may offer improved performance.

References

Adenso-Diaz, B. and Rodriguez, F. (1997). A simple search heuristic for the mclp: Application to the location of ambulance bases in a rural region. *Omega*, 25(2):181–187.

Alho, A. R., e Silva, J. d. A., de Sousa, J. P., and Blanco, E. (2018). Improving mobility by optimizing the number, location and usage of loading/unloading bays for urban freight vehicles. *Transportation Research Part D: Transport and Environment*, 61:3–18.

Alp, O., Erkut, E., and Drezner, Z. (2003). An efficient genetic algorithm for the p-median problem. *Annals of Operations research*, 122(1):21–42.

Anhorn, J. and Khazai, B. (2015). Open space suitability analysis for emergency shelter after an earthquake. *Natural Hazards and Earth System Sciences*, 15(4):789–803.

Balakrishnan, P. and Storbeck, J. E. (1991). Mcthresh: modeling maximum coverage with threshold constraints. *Environment and Planning B: Planning and Design*, 18(4):459–472.

Bazzazi, M., Safaei, N., and Javadian, N. (2009). A genetic algorithm to solve the storage space allocation problem in a container terminal. *Computers & Industrial Engineering*, 56(1):44–52.

Beltran, C., Tadonki, C., and Vial, J. P. (2006). Solving the p-median problem with a semi-lagrangian relaxation. *Computational Optimization and Applications*, 35(2):239–260.

Berman, O., Drezner, Z., Tamir, A., and Wesolowsky, G. O. (2009). Optimal location with equitable loads. *Annals of Operations Research*, 167(1):307–325.

Bianchi, G. and Church, R. L. (1988). A hybrid fleet model for emergency medical service system design. *Social science & medicine*, 26(1):163–171.

Bozkaya, B., Zhang, J., and Erkut, E. (2002). An efficient genetic algorithm for the p-median problem. *Facility location: Applications and theory*, pages 179–205.

Burciu, Ş., Ştefănică, C., Roşca, E., Dragu, V., and Ruscă, F. (2015). Location of an intermediate hub for port activities. In *IOP conference series: materials science and engineering*, volume 95, page 012064. IOP Publishing.

Burkard, R. E. and Rendl, F. (1984). A thermodynamically motivated simulation procedure for combinatorial optimization problems. *European Journal of Operational Research*, 17(2):169–174.

Caliper (2019). *TransCAD*. Version 10.5. Newton, Caliper Corporation. URL: https://desktop.arcgis.com/en/arcmap/10.5/get-started/main/get-started-with-arcmap.htm.

Carreras, M. and Serra, D. (1999). On optimal location with threshold requirements. *Socio-Economic Planning Sciences*, 33(2):91–103.

Chapman, S. and White, J. (1974). Probabilistic formulation of the emergency service facilities location problems. Presented at ORSA/TIMS National Meeting, San Juan, Puerto Rico.

Christaller, W. (1966). *Central places in Southern Germany*. Englewood Cliffs, N.J., Prentice-Hall.

Chung, C., Schilling, D., and Carbone, R. (1983). The capacitated maximal covering problem: A heuristic. In *Proceedings of Fourteenth Annual Pittsburgh Conference on Modeling and Simulation*, volume 1983, pages 1423–1428.

Chung, C.-H. (1986). Recent applications of the maximal covering location planning (mclp) model. *Journal of the Operational Research Society*, 37(8):735–746.

Church, R. and ReVelle, C. (1974). The maximal covering location problem. In *Papers of the regional science association*, volume 32, pages 101–118. Springer-Verlag.

Church, R. L. (1974). *Synthesis of a class of public faciltiy location models*. PhD thesis, The Johns Hopkins University, Baltimore, MD.

Church, R. L. (1980). Developing solid waste planning regions for the tennessee valley authority. In *Proceedings of the 11th annual Pittsburgh conference on modelling and simulation*, volume 11, pages 1611–1618.

Church, R. L. (2002). Geographical information systems and location science. *Computers & Operations Research*, 29(6):541–562.

Church, R. L. and Li, W. (2016). Estimating spatial efficiency using cyber search, gis, and spatial optimization: a case study of fire service deployment in los angeles county. *International Journal of Geographical Information Science*, 30(3):535–553.

Church, R. L. and Murray, A. T. (1993). Modeling school utilization and consolidation. *Journal of Urban Planning and Development*, 119(1):23–38.

Church, R. L. and Murray, A. T. (2009). *Business site selection, location analysis and GIS*. John Wiley & Sons Incorporated.

Church, R. L. and Murray, A. T. (2018). *Location Covering Models: History, Applications and Advancements*. Springer.

Church, R. L. and ReVelle, C. S. (1976). Theoretical and computational links between the p-median, location set-covering, and the maximal covering location problem. *Geographical Analysis*, 8(4):406–415.

Church, R. L. and Roberts, K. L. (1983). Generalized coverage models and public facility location. In *Papers of the regional science association*, volume 53, pages 117–135. Springer.

Church, R. L. and Somogyi, C. (1985). Optimizing service and access coverage. Presented at North American Meetings of the Regional Science Association, Philadelphia, PA.

Church, R. L. and Sorensen, P. (1996). Integrating normative location models into gis: Problems and prospects with the p-median model (94-5). *Spatial analysis: modelling in a GIS environment*, pages 179–190.

Cohon, J. L. (1978). *Multiobjective programming and planning*, volume 140. New York: Academic Press.

Current, J. R. and Storbeck, J. E. (1988). Capacitated covering models. *Environment and planning B: planning and Design*, 15(2):153–163.

D'Amico, S. J., Wang, S.-J., Batta, R., and Rump, C. M. (2002). A simulated annealing approach to police district design. *Computers & operations research*, 29(6):667–684.

Daskin, M. S. (1982). Application of an expected covering model to emergency medical service system design. *Decision Sciences*, 13(3):416–439.

Daskin, M. S. (1983). A maximum expected covering location model: formulation, properties and heuristic solution. *Transportation science*, 17(1):48–70.

Daskin, M. S. (2011). *Network and discrete location: models, algorithms, and applications*. John Wiley & Sons.

Daskin, M. S. and Stern, E. H. (1981). A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2):137–152.

Daskin, M. S. and Tucker, E. L. (2018). The trade-off between the median and range of assigned demand in facility location models. *International Journal of Production Research*, 56(1-2):97–119.

Davoodi, M. (2019). k-balanced center location problem: A new multi-objective facility location problem. *Computers & Operations Research*, 105:68–84.

Densham, P. J. and Rushton, G. (1992). A more efficient heuristic for solving largep-median problems. *Papers in Regional Science*, 71(3):307–329.

Downs, B. T. and Camm, J. D. (1996). An exact algorithm for the maximal covering problem. *Naval Research Logistics (NRL)*, 43(3):435–461.

Downs, J., Horner, M., Loraamm, R., Anderson, J., Kim, H., and Onorato, D. (2014). Strategically locating wildlife crossing structures for f lorida panthers using maximal covering approaches. *Transactions in GIS*, 18(1):46–65.

Drezner, Z. and Hamacher, H. W. (2004). *Facility location: applications and theory*. Springer Science & Business Media.

Efroymson, M. and Ray, T. (1966). A branch-bound algorithm for plant location. *Operations Research*, 14(3):361–368.

Eiselt, H. A. and Laporte, G. (1995). Objectives in location problems. *Facility location: a survey of applications and methods*.

Elkady, S. K. and Abdelsalam, H. M. (2016). A modified multi-objective particle swarm optimisation algorithm for healthcare facility planning. *International Journal of Business and Systems Research*, 10(1):1–22.

Erfani, S. M. H., Danesh, S., Karrabi, S. M., Gheibi, M., and Nemati, S. (2019). Statistical analysis of effective variables on the performance of waste storage service using geographical information system and response surface methodology. *Journal of environmental management*, 235:453–462.

Erfani, S. M. H., Danesh, S., Karrabi, S. M., Shad, R., and Nemati, S. (2018). Using applied operations research and geographical information systems to evaluate effective factors in storage service of municipal solid waste management systems. *Waste Management*, 79:346–355.

Erkut, E. (1993). Inequality measures for location problems. *Computers & Operations Research*.

Esri (2019). *ArcGIS*. version 10.5. Redlands, Esri. URL: https://desktop.arcgis.com/en/arcmap/10.5/get-started/main/get-started-with-arcmap.htm.

Ferrari, T., Camara, M. V. O., Nassi, C. D., Ribeiro, G. M., Costa Junior, R. R., Ribeiro Júnior, C., and Bilate, A. (2018). Analysis of the location of rescue ambulance dispatch bases: a case study in rio de janeiro, brazil. *Geographical Analysis*, 50(4):397–421.

Galvão, R. D. and ReVelle, C. (1996). A lagrangean heuristic for the maximal covering location problem. *European Journal of Operational Research*, 88(1):114–123.

Garey, M. R. and Johnson, D. S. (1979). *Computers and intractability*, volume 174. freeman San Francisco.

Garfinkel, R. S. and Nemhauser, G. L. (1970). Optimal political districting by implicit enumeration techniques. *Management Science*, 16(8):B–495.

Gary, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-completeness*. W.H. Freeman and Co.

Gerrard, R. A. (1995). *The location of service facilities using models sensitive to response distance, facility workload, and demand allocation*. PhD thesis, University of California, Santa Barbara, Santa Barbara, CA.

Gerrard, R. A., Church, R. L., Stoms, D. M., and Davis, F. W. (1997). Selecting conservation reserves using species-covering models: Adapting the arc/info gis. *Transactions in GIS*, 2(1):45–60.

Golden, B. L. and Skiscim, C. C. (1986). Using simulated annealing to solve routing and location problems. *Naval Research Logistics Quarterly*, 33(2):261–279.

Haghani, A. (1996). Capacitated maximum covering location models: Formulations and solution procedures. *Journal of advanced transportation*, 30(3):101–136.

Hakimi, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations research*, 12(3):450–459.

Helo, P., Rouzafzoon, J., Solvoll, G., Hanssen, T.-E. S., Westin, L., and Westin, J. (2018). Distribution center location analysis for nordic countries by using network optimization tools.

Hogan, K. and ReVelle, C. (1985). The capacitated maximal covering problem. Presented at ORSA/TIMS 1985, Boston, MA.

Hogan, K. and ReVelle, C. (1986). Concepts and applications of backup coverage. *Management science*, 32(11):1434–1444.

Hong, S. and Kuby, M. (2016). A threshold covering flow-based location model to build a critical mass of alternative-fuel stations. *Journal of Transport Geography*, 56:128–137.

Huang, H. C., Lee, C., and Xu, Z. (2006). The workload balancing problem at aircargo terminals. *Or Spectrum*, 28(4):705–727.

Jaramillo, J. H., Bhadury, J., and Batta, R. (2002). On the use of genetic algorithms to solve location problems. *Computers & Operations Research*, 29(6):761–779.

Johnson, D. S., Aragon, C. R., McGeoch, L. A., and Schevon, C. (1991). Optimization by simulated annealing: an experimental evaluation; part ii, graph coloring and number partitioning. *Operations research*, 39(3):378–406.

Karmarkar, N. and Karp, R. M. (1982). *The differencing method of set partitioning*. Computer Science Division (EECS), University of California Berkeley.

Kim, D.-G. and Kim, Y.-D. (2010). A branch and bound algorithm for determining locations of long-term care facilities. *European Journal of Operational Research*, 206(1):168–177.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, 220(4598):671–680.

Koulamas, C., Antony, S., and Jaen, R. (1994). A survey of simulated annealing applications to operations research problems. *Omega*, 22(1):41–56.

Lee, G. and Murray, A. T. (2010). Maximal covering with network survivability requirements in wireless mesh networks. *Computers, Environment and Urban Systems*, 34(1):49–57.

Lemire, P.-O., Delcroix, B., Audy, J.-F., Labelle, F., Mangin, P., and Barnabé, S. (2019). Gis method to design and assess the transportation performance of a decentralized biorefinery supply system and comparison with a centralized system: case study in southern quebec, canada. *Biofuels, Bioproducts and Biorefining*, 13(3):552–567.

Liao, K. and Guo, D. (2008). A clustering-based approach to the capacitated facility location problem 1. *Transactions in GIS*, 12(3):323–339.

Manliclic, K. S. (2016). *A Suite of Methodologies to Systematically Site Distributed Generation Technologies , such as Poly-Generation Fuel Cells , in Support of Alternative Transportation Infrastructure*. PhD thesis, University of California, Irvine, Irvine, CA.

Maranzana, F. (1964). On the location of supply points to minimize transport costs. *Journal of the Operational Research Society*, 15(3):261–270.

Marianov, V. and Serra, D. (1998). Probabilistic, maximal covering location—allocation models forcongested systems. *Journal of Regional Science*, 38(3):401–424.

Marín, A. (2011). The discrete facility location problem with balanced allocation of customers. *European Journal of Operational Research*, 210(1):27–38.

Marsh, M. T. and Schilling, D. A. (1994). Equity measurement in facility location analysis: A review and framework. *European journal of operational research*, 74(1):1–17.

Megiddo, N., Zemel, E., and Hakimi, S. L. (1983). The maximum coverage location problem. *SIAM Journal on Algebraic Discrete Methods*, 4(2):253–261.

Mišković, S. and Stanimirović, Z. (2015). Memetic algorithm for the balanced resource location problem with preferences. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–6. IEEE.

Mulvey, J. M. and Beck, M. P. (1984). Solving capacitated clustering problems. *European Journal of Operational Research*, 18(3):339–348.

Mumphrey, A. J., Seley, J. E., and Wolpert, J. (1971). A decision model for locating controversial facilities. *Journal of the American Institute of Planners*, 37(6):397–402.

Murawski, L. and Church, R. L. (2009). Improving accessibility to rural health services: The maximal covering network improvement problem. *Socio-Economic Planning Sciences*, 43(2):102–110.

Murray, A. T. (2016). Maximal coverage location problem: impacts, significance, and evolution. *International Regional Science Review*, 39(1):5–27.

Murray, A. T. and Church, R. L. (1996). Applying simulated annealing to location-planning models. *Journal of Heuristics*, 2(1):31–53.

Murray, A. T. and Gerrard, R. A. (1997). Capacitated service and regional constraints in location-allocation modeling. *Location science*, 5(2):103–118.

Murray, A. T., Matisziw, T. C., Wei, H., and Tong, D. (2008). A geocomputational heuristic for coverage maximization in service facility siting. *Transactions in GIS*, 12(6):757–773.

Murray, A. T. and O'Kelly, M. E. (2002). Assessing representation error in point-based coverage modeling. *Journal of Geographical Systems*, 4(2):171–191.

Murray, A. T. and Tong, D. (2007). Coverage optimization in continuous space facility siting. *International Journal of Geographical Information Science*, 21(7):757–776.

Murray, A. T. and Tong, D. (2009). Gis and spatial analysis in the media. *Applied geography*, 29(2):250–259.

Murray, A. T., Xu, J., Wang, Z., and Church, R. L. (2019). Commercial gis location analytics: capabilities and performance. *International Journal of Geographical Information Science*, 33(5):1106–1130.

Naharudin, N. (2014). Application of location/allocation models and gis to the location of national primary schools in rawang, malaysia.

Narasimhan, S., Pirkul, H., and Schilling, D. A. (1992). Capacitated emergency facility siting with multiple levels of backup. *Annals of Operations Research*, 40(1):323–337.

Narula, S. C., Ogbu, U. I., and Samuelsson, H. M. (1977). An algorithm for the p-median problem. *Operations Research*, 25(4):709–713.

Pirkul, H. (1987). Efficient algorithms for the capacitated concentrator location problem. *Computers & Operations Research*, 14(3):197–208.

Pirkul, H. and Schilling, D. (1989). The capacitated maximal covering location problem with backup service. *Annals of Operations Research*, 18(1):141–154.

Pirkul, H. and Schilling, D. A. (1988). The siting of emergency service facilities with workload capacities and backup service. *Management Science*, 34(7):896–908.

Pirkul, H. and Schilling, D. A. (1991). The maximal covering location problem with capacities on total workload. *Management Science*, 37(2):233–248.

ReVelle, C. (1993). Facility siting and integer-friendly programming. *European Journal of Operational Research*, 65(2):147–158.

ReVelle, C. S. and Swain, R. W. (1970). Central facilities location. *Geographical analysis*, 2(1):30–42.

Sánchez, J., Curt, M. D., Sanz, M., and Fernández, J. (2015). A proposal for pellet production from residual woody biomass in the island of majorca (spain). *AIMS Energy*, 3(3):480–504.

Savas, E. S. (1978). On equity in providing public services. *Management Science*, 24(8):800–808.

Shahid, A. and Mas Machuca, C. (2015). Enhanced dimensioning and comparative analysis of different protection schemes for hybrid pon converged access networks (hpcan).

Shariff, S. R., Moin, N. H., and Omar, M. (2012). Location allocation modeling for healthcare facility planning in malaysia. *Computers & Industrial Engineering*, 62(4):1000–1010.

Shariff, S. S. R., Moin, N. H., and Omar, M. (2013). An alternative heuristic for capacitated p-median problem (cpmp). In *2013 IEEE Business Engineering and Industrial Applications Colloquium (BEIAC)*, pages 916–921. IEEE.

Sharma, B., Clark, R., Hilliard, M. R., and Webb, E. G. (2018). Simulation modeling for reliable biomass supply chain design under operational disruptions. *Frontiers in Energy Research*, 6:100.

Sharpe, R. and Marksjö, B. S. (1986). Solution of the facilities layout problem by simulated annealing. *Computers, environment and urban systems*, 11(4):147–154.

Sorensen, P. and Church, R. (2010). Integrating expected coverage and local reliability for emergency medical services location problems. *Socio-Economic Planning Sciences*, 44(1):8–18.

Straitiff, S. L. and Cromley, R. G. (2010). Using gis and k= 3 central place lattices for efficient solutions to the location set-covering problem in a bounded plane. *Transactions in GIS*, 14(3):331–349.

Talbi, E.-G. (2009). *Metaheuristics: from design to implementation*, volume 74. John Wiley & Sons.

Teitz, M. B. and Bart, P. (1968). Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations research*, 16(5):955–961.

Teixeira, T. R., Ribeiro, C. A. A. S., dos Santos, A. R., Marcatti, G. E., Lorenzon, A. S., de Castro, N. L. M., Domingues, G. F., Leite, H. G., Mota, P. H. S., de Almeida Telles, L. A., et al. (2018). Forest biomass power plant installation scenarios. *Biomass and Bioenergy*, 108:35–47.

The City of San Jose (2019). Data downloads. Retrieved January 22, 2019, from http://www.sanjoseca.gov/index.aspx?NID=3308.

Tiggelaar, S. (2016). Very preterm infants in alberta: comparison of health technology service use, health outcomes and costs across five health zones.

Tong, D. and Church, R. L. (2012). Aggregation in continuous space coverage modeling. *International Journal of Geographical Information Science*, 26(5):795–816.

Tong, D., Murray, A., and Xiao, N. (2009). Heuristics in spatial analysis: a genetic algorithm for coverage maximization. *Annals of the Association of American Geographers*, 99(4):698–711.

Tong, D. and Wei, R. (2017). Regional coverage maximization: alternative geographical space abstraction and modeling. *Geographical Analysis*, 49(2):125–142.

Toregas, C., Swain, R., ReVelle, C., and Bergman, L. (1971). The location of emergency service facilities. *Operations research*, 19(6):1363–1373.

U.S. Census Bureau (2018a). Tiger/line® shapefiles and tiger/line® files. Retrieved July 14, 2018, from https://www.census.gov/geo/maps-data/data/tiger-line.html.

U.S. Census Bureau (2018b). Zip code tabulation areas (zctas). Retrieved September 11, 2018, from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html.

Weaver, J. and Church, R. (1981). Average response time and workload balance: two criteria for ambulance station location. *Systems Science in Health Care*, pages 975–983.

Weaver, J. and Church, R. (1983). A comparison of solution procedures for covering location problems. *Modeling and Simulation*, 14:1417.

Wei, R. and Murray, A. T. (2014). A multi-objective evolutionary algorithm for facility dispersion under conditions of spatial uncertainty. *Journal of the Operational Research Society*, 65(7):1133–1142.

Wei, R. and Murray, A. T. (2015). Continuous space maximal coverage: Insights, advances and challenges. *Computers & operations research*, 62:325–336.

Wei, R. and Murray, A. T. (2016). A parallel algorithm for coverage optimization on multi-core architectures. *International Journal of Geographical Information Science*, 30(3):432–450.

Wilhelm, M. R. and Ward, T. L. (1987). Solving quadratic assignment problems by 'simulated annealing'. *IIE transactions*, 19(1):107–119.

Wolsey, L. A. and Nemhauser, G. L. (1999). *Integer and combinatorial optimization*, volume 55. John Wiley & Sons.

Xiao, N. and Murray, A. T. (2019). Spatial optimization for land acquisition problems: A review of models, solution methods, and gis support. *Transactions in GIS*, 23(4):645–671.

Xu, J., Murray, A., Wang, Z., and Church, R. (2020). Challenges in applying capacitated covering models. *Transactions in GIS*, 24(2):268–290.

Yin, P. and Mu, L. (2012). Modular capacitated maximal covering location problem for the optimal siting of emergency vehicles. *Applied Geography*, 34:247–254.

Zarandi, M. F., Davari, S., and Sisakht, S. H. (2011). The large scale maximal covering location problem. *Scientia Iranica*, 18(6):1564–1570.

Zhang, C., Liu, J., Wan, Y.-w., Murty, K. G., and Linn, R. J. (2003). Storage space allocation in container terminals. *Transportation Research Part B: Methodological*, 37(10):883–903.

Zhu, Z. and McKnew, M. A. (1993). A goal programming workload balancing optimization model for ambulance allocation: An application to shanghai, pr china. *Socio-Economic Planning Sciences*, 27(2):137–148.