PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Thiago Edmar de Oliveira

Solving a University Admission Exam Location Problem: an Application in Brazil

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Engenharia de Produção da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção.

Advisor   : Prof. Rafael Martinelli Pinto
Co-advisor: Prof. Bruno Milanez

Rio de Janeiro
March 2021

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Thiago Edmar de Oliveira

Solving a University Admission Exam Location Problem: an Application in Brazil

Dissertation presented to the Programa de Pós–graduação em Engenharia de Produção da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção. Approved by the Examination Committee.

Prof. Rafael Martinelli Pinto
Advisor
Departamento de Engenharia Industrial – PUC-Rio

Prof. Bruno Milanez
Co-advisor
Universidade Federal de Juiz de Fora – UFJF

Prof. Sanjay Dominik Jena
Université du Québec à Montréal – UQAM

Prof. Leonardo Goliatt da Fonseca
Universidade Federal de Juiz de Fora – UFJF

Rio de Janeiro, March 17, 2021

Thiago Edmar de Oliveira

Bachelor in Electrical Engineering (2017) at the Federal University of Juiz de Fora (UFJF) and specialization in Occupational Safety Engineering (2020) at the Faculdade Única de Ipatinga. The author works as a administrative technician at UFJF in the areas of logistics and computation.

Acknowledgments

I thank all my family, especially my mother, Rosimar, for always being by my side, helping and supporting me in the most challenging moments. Without her support and affection, this course's completion would not be possible.

My wife and best friend, Nádia, was fundamental in this journey for her true companionship and constant love. Her hand next to mine is the pillar that keeps me going, always.

I want to thank all the professors, especially my advisor Rafael Martinelli for the opportunity, lessons and support for the development of this study, and the co-advisor Bruno Milanez for his valuable guidance. PUC-Rio for the opportunity to offer this course and UFJF for collaborating effectively with the production of this study, providing access to the necessary data and promoting full support to the research.

I must thank also my friends Dalyara and Maicon, for the friendship of more than a decade, and my friend Renan, who, even in his absence, is present through his friendship. My colleagues in the course, especially the great companion Higor, friend and co-worker, for the partnership in the significant challenges we faced during this journey.

Abstract

Oliveira, Thiago Edmar de; Martinelli Pinto, Rafael (Advisor); Milanez, Bruno (Co-Advisor). **Solving a University Admission Exam Location Problem: an Application in Brazil**. Rio de Janeiro, 2021. 67p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

This work presents a methodology in order to reduce the displacement of candidates on test days in an admission exam of a university in Brazil. The exam has different types of tests according to the candidate's grade and each test site can offer only one type of exam. Based on the theory of facility location, a mathematical model of integer programming was developed to optimally allocate candidates considering the distance between them and the various existing exam locations. Several tests were carried out with the exam data already applied, which showed a reduction in total travel distance of more than 30%. Then, the methodology was applied directly in the most recent admission exam of the institution, which has more than 34 thousand candidates distributed over 70 exam places in 5 cities, with the proportion of candidates able to walk on foot being 4 times higher when compared to allocations used in previous years.

Keywords

Candidates Allocation; Exam Location; Binary Programming; Facility Location; Optimization

Resumo

Oliveira, Thiago Edmar de; Martinelli Pinto, Rafael; Milanez, Bruno. **Resolvendo um problema de localização de exame de admissão em universidade: uma aplicação no Brasil**. Rio de Janeiro, 2021. 67p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Este trabalho apresenta uma metodologia a fim de reduzir o deslocamento de candidatos em dias de prova em um exame de admissão de uma universidade no Brasil. O exame possui diferentes tipos de prova de acordo com a série do candidato e cada local de prova pode ofertar apenas um tipo de exame. A partir da teoria de localização de instalações, desenvolveu-se um modelo matemático de programação inteira para alocar candidatos de forma ótima considerando a distância entre eles e os vários locais de prova existentes. Foram realizados diversos testes com os dados de exames já aplicados, o que mostrou uma redução no deslocamento total superior a 30%. Em seguida, a metodologia foi aplicada diretamente no mais recente exame de admissão da instituição, que conta com mais de 34 mil candidatos distribuídos por 70 locais de prova em 5 cidades, com a proporção de candidatos aptos a se locomoverem a pé sendo 4 vezes maior quando comparada com alocações utilizadas em anos anteriores.

Palavras-chave

Alocação de Candidatos;   Localização de Exame;   Programação Binária;   Localização de Instalações;   Otimização

Table of contents

List of figures

List of tables

1
Introduction

Around the world, most higher education institutions promote some kind of written exam in their admission selection process to measure the applicants' knowledge. These exams can be organized and applied across a country or can be unique to each institution.

In the United States, nearly 2.2 million students in the high school class of 2020 took the Scholastic Aptitude Test - SAT (College Board 2020), a standardized test widely used for college admissions in the country. Other countries such as Turkey (Hafalir et al. 2018), Finland (Jokila et al. 2020), Chile, Japan, Mexico and others (Insider 2018) also apply nationwide tests for admission to higher education. China is the most prominent country in this area, applying the National College Entrance Examination - NCEE, commonly known as *gaokao* (Pires 2019), for more than 10.3 million students in 2020 (China Daily 2020), by far the largest existing standardized exam.

In Brazil, admission to a higher education institution to attend an undergraduate course is, by law, carried out through a selection process (Brasil 1996). This way, most universities apply a written test, some of them using the score of a national exam - called ENEM, organized and applied by the federal government - to rank their applicants. Despite this federal exam, there are still many universities that perform their own admission test, mainly the most prestigious universities in the country. These exams are called *vestibular* in Portuguese.

Among the institutions that elaborate and apply their own selection process is the Federal University of Juiz de Fora - UFJF (from *Universidade Federal de Juiz de Fora*, in Portuguese), one of the thirty largest federal universities in the country in the number of students (Instituto Nacional de Estudos e Pesquisas Educacionais 2018). UFJF destinates half of its vacancies in undergraduate courses to ENEM - which is organized and applied by the federal government - and another half to its own selection exam called Mixed Selective Admission Program - PISM (in Portuguese, *Programa de Ingresso Seletivo Misto*), which offered 2,303 vacancies in its last edition.

PISM was created in 1999 (Universidade Federal de Juiz de Fora 1999) and, unlike a traditional entrance exam, it represents a cumulative and serial

admission modality, in which the candidates take a test at the end of each high school grade. This selection process was designed with the objectives of improving the interrelationship between high school and higher education, allowing a gradual and systematic assessment of candidates, and valuing the work of high school teachers and the students' learning process (da Veiga 2020).

Since its creation, UFJF has promoted major changes in PISM. In the beginning, the university offered the exam only on its campus in the city of Juiz de Fora and today the test is applied in five cities. In its first edition, there were 8,986 candidates for PISM, two decades later the exam reached 40,126 candidates (Universidade Federal de Juiz de Fora 2020a). In 2016, the selection process was consolidated in its current format.

PISM is divided into three modules – one for each of the high school grades in Brazil –, each module having a different exam with the content of the specific series in which the candidate is. Thus, this exam is only for students attending regular secondary schools, and it has the goal of integration between higher education and secondary school since it assesses the student's knowledge year by year. At the end of the last module – which occurs when the candidate completes the third and last year of secondary school – the grades of the candidate in each of the previous modules are summed so that students can be classified according to the intended course.

As Juiz de Fora is the main city of Zona da Mata, a Minas Gerais state macro-region with more than 2 million inhabitants (Instituto Brasileiro de Geografia e Estatística 2010), PISM has a substantial impact in the city and its surroundings, especially in the days of application of the tests. Over the last years, UFJF has seen an increment of the candidates interested in the process and, as a consequence, a significant increase in the number of applicants in the PISM. For 2018's admission, for example, the number of candidates was about 30,000, with this amount increasing to more than 34,000 in 2019 and about 40,000 in 2020, with a total increase of 78% from 2015 to the 2020 edition.

Such growth makes the university's goal of providing a calm environment and offering a reliable and safe examination to candidates, economically and efficiently, much more challenging. In addition, it amplifies the burdens in the cities where the tests are applied: there are lots of records about traffic jams, a problem that is a consequence of the large number of people (candidates and, considering that they are underage, also parents) in transit on the test days and, especially, due to the fact that these people have to move to exam locations far away of their residences.

PISM nowadays requires the participation of more than 4,500 people directly involved in the application of the exams, with still hundreds of people

working with planning, logistics, and subsequent correction and processing of the results. Due to its large scale, the process officially begins in July with the publication of the public call. The exams are applied on the first weekend of December (in two days), with the grading being started already in the week following the exam. The results are published at the beginning of January for candidates of the third module of PISM who finished secondary school and, therefore, are competing for the available positions. Until March, all grades are published, including those that belong to candidates from the first and second modules who still do not compete for positions.

At the Federal University of Juiz de Fora, it was noted since previous exams that there was a demand from the candidates for their allocation in the exam places to start considering the location of the candidate's residence in order to facilitate their mobility on exam days, which also became then a demand of the administration of the institution. However, this project was not carried out due to technical challenges and resource availability.

This way, through the collaboration of UFJF, we performed a survey with candidates of the 2020 selection process edition to verify and reassure the candidates' demand for an optimization of their allocations process. This survey was available to them from November 15 to December 1 of 2019, the period in which the candidates needed to access the registration platform to consult the Proof of Registration - which informs them of the exam location where they were allocated.

We emphasize that participation in this survey was free to avoid conflict responses due to compulsory involvement. From a total of 40 thousand registrants in this selection process, there were 5,710 candidates who participated in the survey. It was meant to be short and simple, composed of four direct questions with pre-defined response options. These questions and their respective percentages of responses are presented in sequence.

From Figure 1.1 we perceive that the majority of candidates were allocated far or too far from their exam locations - a total of 64%. In Figure 1.2 we note that less than a half of candidates were satisfied with their allocation. In Figure 1.3 it is possible to note that at least 42% of candidates could be allocated in a closer exam location, which makes clear the existence of an opportunity to improve this process. At last, in Figure 1.4 we see that most candidates - 63%, divided between those who use their own vehicle and those who use taxi or app - use cars to reach their destination, which can cause a considerable negative impact on the traffic during exam days. Through the survey analysis, we can infer the candidates' demand for optimization in their allocation and corroborate the possible positive outcomes of this implementation.

How far away is your exam location from your residence?

Very close (I can arrive in 15min)
Nearby (I can arrive in 30 minutes)
Far (I can arrive in 45min)
Too far (I can arrive in more than 45 minutes)
I don't know
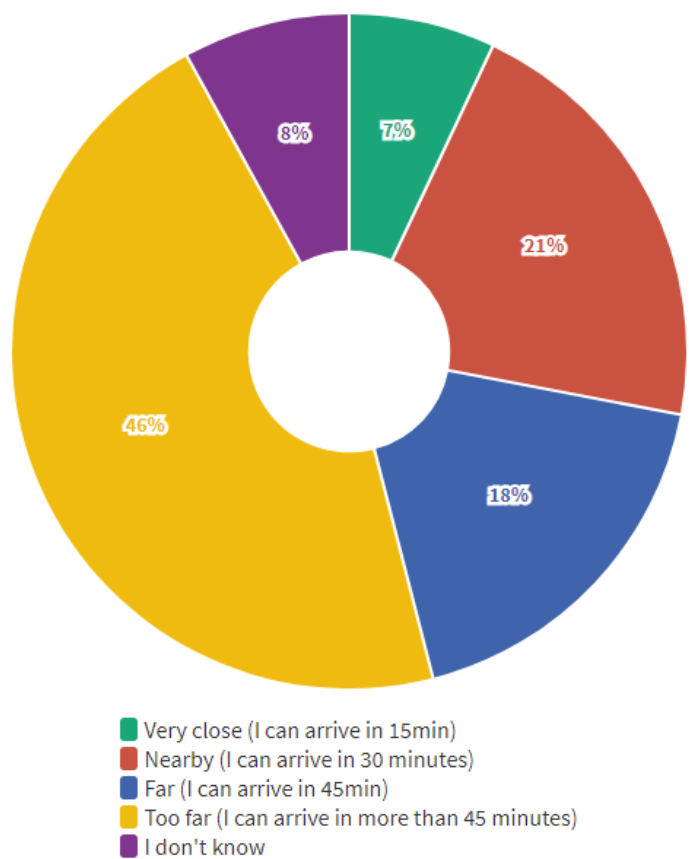
Figure 1.1: Question 1 and responses

How satisfied are you with the location you were allocated?
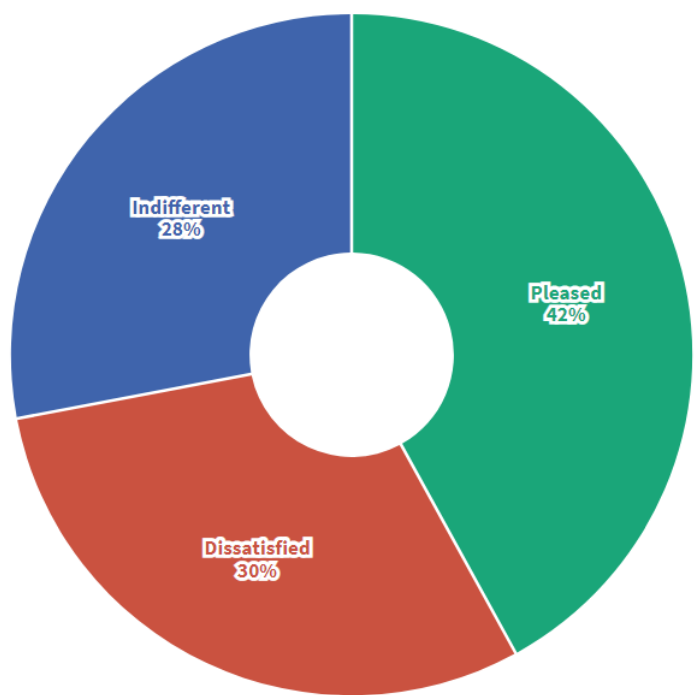
Figure 1.2: Question 2 and responses

Is there any location closer to your residence than the one you were allocated?

I don't know
32%

Yes
42%

No
26%
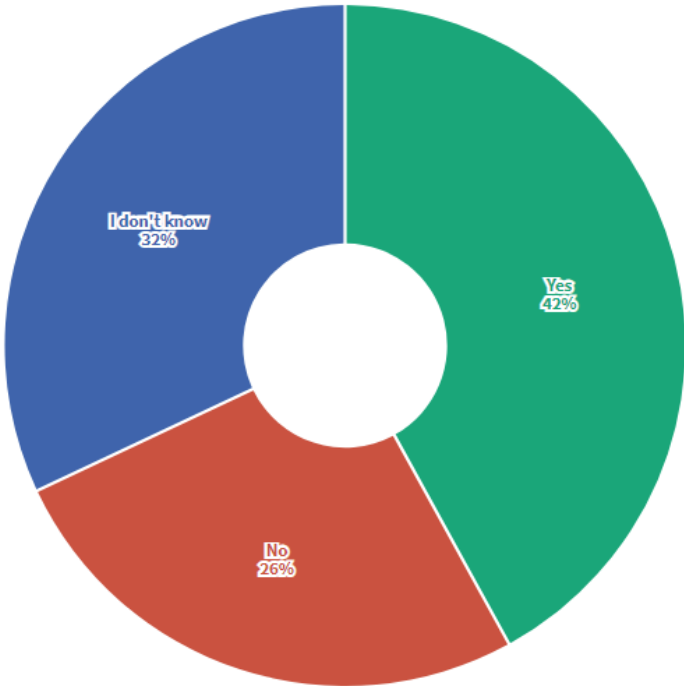
Figure 1.3: Question 3 and responses

How will you travel to the exam location?

On foot
8%

Public transportation
29%

Taxi / App
12%

Own vehicle
51%
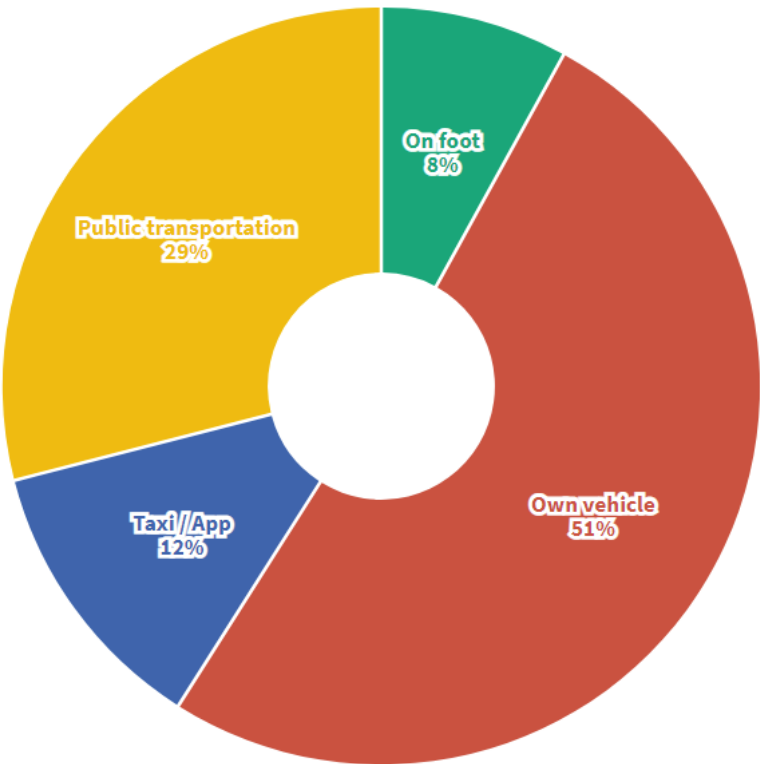
Figure 1.4: Question 4 and responses

1.1
Objectives

At UFJF, the past and current processes of allocating candidates at examination sites completely disregard the distances between one and the other, using just a personal identification number to distribute the candidates at random. Informal contact with other large companies that deal with the organization of selection process in the country revealed that, despite using a method to try to allocate candidates closer to their address of residence, such methods do not involve any kind of mathematical optimization or scientific approach, based only on the geographical knowledge of the employees without using a computational method.

This way, we aim to develop an original mathematical model to optimize the allocation of candidates in exam locations in order to reduce their travel distance on exam days, grounded on the scientific literature about facility location problems. With the collaboration of UFJF, we intend to apply our model in its entrance exam and test its effectiveness and acceptance among the applicants. To do so, we intend to use here only free tools to facilitate the adoption of this work by the university and other higher education institutions, especially public ones.

Despite its primary objective, it is expected that the present work will not only be able to reduce the total distance traveled by the candidates but, with this, significantly improve the allocation of candidates who live close to exam sites, allowing an increase in the number of candidates who are able to go to exam locations on foot and reduce traffic on exam days, which can decrease the environmental impact related to car use, in addition to improving the mobility of most candidates.

1.2
Organization

This work is structured in five chapters. The current chapter introduces the context in which the study was developed and its motivations, as well as the objectives that the author intends to achieve.

Chapter 2 presents the theoretical foundation and analyses the main works related to the topic addressed. The problem studied here is specified, the existing models are presented and the methods of available solutions are explored. At last, an original approach to the problem is presented.

In Chapter 3, the methodology developed and used in this work is explained. All its stages are demonstrated and the mathematical model used to perform the optimizations is elaborated and discussed.

The computational experiments are shown in Chapter 4, with all the data and results obtained through the various tests performed. The application of the methodology in real data is explained in detail, the algorithms and platforms used are described and then is presented the implementation of this work in a real case.

Finally, in Chapter 5, we discuss the study's conclusions and future research possibilities.

2
Literature Review

In this chapter, we review the literature associated with facility location problems. In Section 2.1, we discuss the most relevant academic works related to this theme. In Section 2.2, we address the main features present in the models. Section 2.3 focuses on the different kinds of models. Section 2.4 presents the solution methods and in Section 2.5 we introduce a specific type of problem addressed in this work.

2.1
Facility Location Problems

Facility Location Problem refers to the modeling, formulation and solution of a large class of problems whose objective is to allocate facilities in a given space. Its solution aims to optimize the fulfillment of a set of demands - or clients -, subject to a set of restrictions (Farahani and Hekmatfar 2009). It is a critical aspect of strategic planning for a broad spectrum of public and private firms (Owen and Daskin 1998).

The mathematical science related to facility location problems has experienced substantial growth in the last decades, as stated by Farahani and Hekmatfar (2009). To illustrate this, a quick search in the SCOPUS platform in November 2020 for "facility location problem" revealed more than 2.5 thousand documents, with its majority being published after the year 2000, as shown in Figure 2.1.

The first study found in the literature mentioning a location problem was released originally in 1909 by Alfred Weber with an English version in 1929 (Weber 1929). In his book of more than three hundred pages with an extensive analysis of the subject, the author addresses spatial factors to obtain the ideal location and minimize production facilities' costs. After this, few studies are found about this topic, with limited relevance. The subject then gains prominence with the publication of Hakimi (1964), who used optimum locations in a graph to perform allocations of a switching center in a communication network and police station in a highway system. After the 1990s, several consecrated books appeared, the works with more relevancy being Drezner (1995) and Daskin (1995).
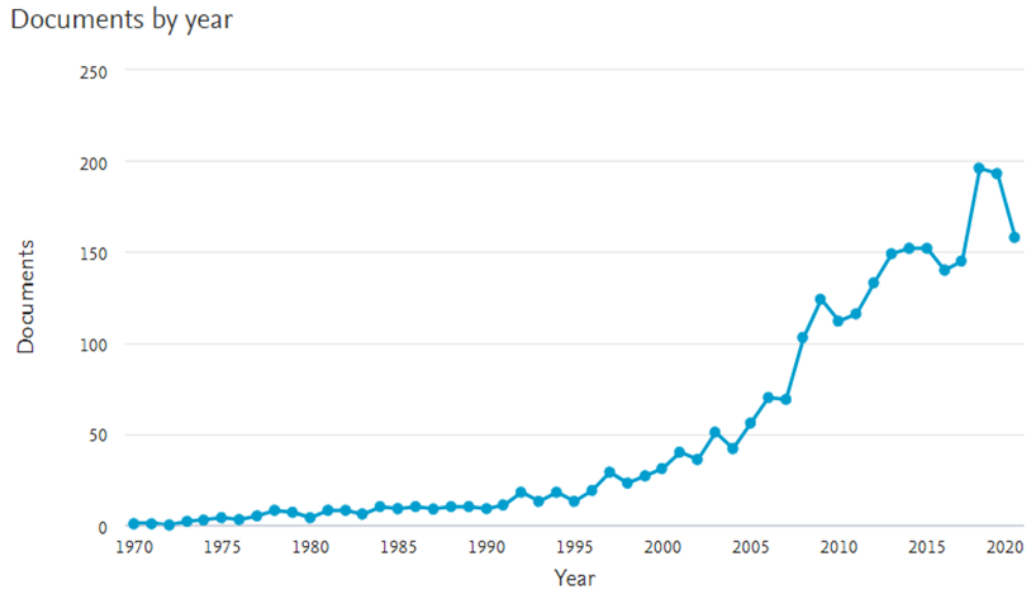
Figure 2.1: Results of SCOPUS for "facility location problem"

Source: SCOPUS platform at November 2020

There are four main components in this facility location problems: customers, commonly already allocated to certain points or routes; facilities, that must be allocated; space, where customers and facilities are located; and unit of measure, which indicates distance or time between customer and facilities. These, and more characteristics, will be discussed in the next section.

2.2
Model Features

Facility location problems have a set of features in which it is possible to classify and analyze them. The main features are: distance; capacity of facilities - capacitated or uncapacitated; number of facilities - single facility or multiple facilities; time - static (single period) or dynamic (multi-period); type of parameter - deterministic (inputs assumed to be known in certain) or probabilistic (probability distribution and scenarios); type of facility - different kind or identical ones; space - planar (facilities can be allocated anywhere in space), network (facilities are restricted to a network of nodes and edges) or discreet (facilities can be allocated at a pre-defined set of points). Each of these characteristics has a significant impact on the model's structure, study, and solution method.

Distance is a fundamental feature in facility location. There are several kinds of distance functions with a vast range of definitions - the kind of distance to be used depends on the problem's characteristics and its application. Euclidean Distance is the simplest and most common method to calculate

distance. It is obtained through the size of a line segment between two points on the Cartesian coordinates by applying the Pythagorean theorem, as shown in Figure 2.2.

$$\text{Euclidean distance (d)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
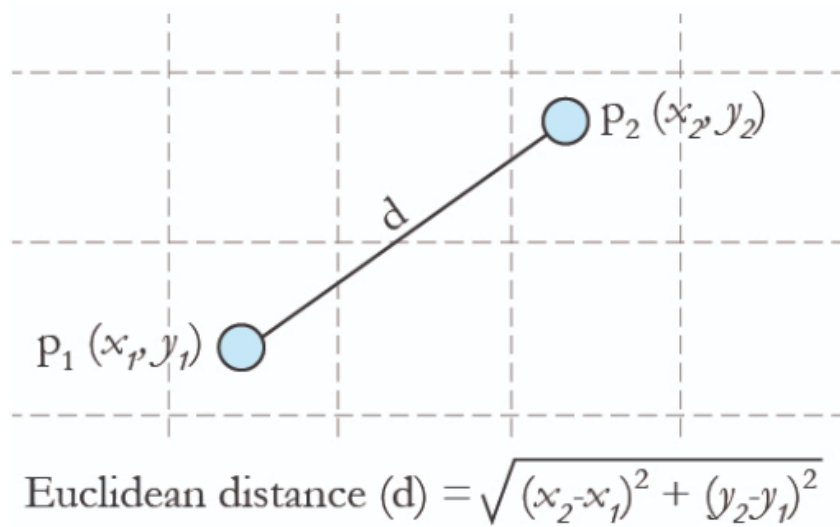
Figure 2.2: Euclidean distance formula

Source: Tutorial Example website

The Manhattan Distance is calculated from the sum of the absolute differences of its coordinates, it is commonly used for handling material inside a warehouse, for example. Geodesic Distance, Figure 2.3, is the distance between two points taking into account curvatures of the space and it is used to calculate the distance of objects very far from each other such as telecommunication antennas, in which disregarding the curvature of the earth can cause significant errors. The Euclidean and geodesic distances are considerably close for short distances, but they can differ over vast distances.

Real Distance considers the streets and directions of the roads. Because of that, it demands the use of a Geographic Information System - GIS, explained in detail later in Section 2.5. Real distances are widely used in Facility Location Problems, mainly those that intend to locate facilities in large geographic areas, such as neighborhoods or cities. As the calculation of real distance is not directly obtained through the locations, but instead, it requires an external service that provides such information, it is necessary to be very careful in the choice of which GIS provider to use. Figure 2.4 shows a comparison between Euclidean, Manhattan and real distance in a map.

There are still some less common kinds of distance with very specific applications. Hamming distance is the one between two strings of the same length, that is, the number of positions in which they differ. It is normally used in telecommunications, where it is better known as the signal distance. Some
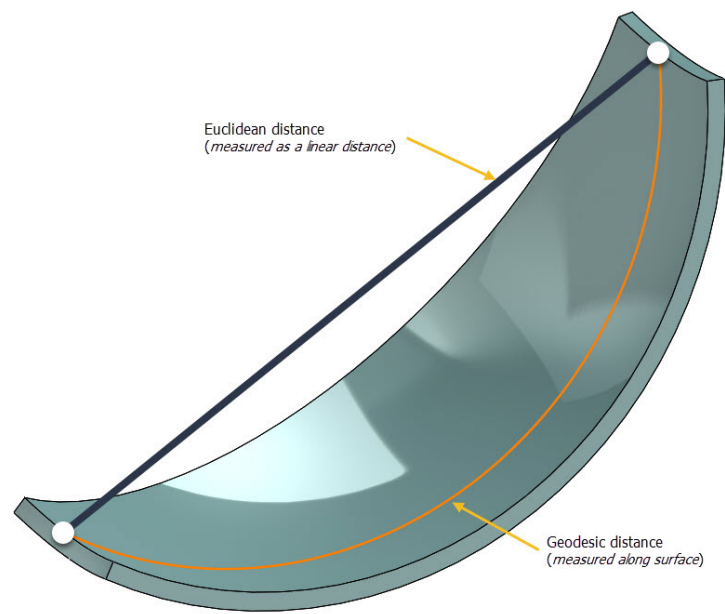
Figure 2.3: Geodesic and Euclidean distance comparison

Source: Using Geodesic Offsets on Complex Curves (https://www.javelin-tech.com/blog/)
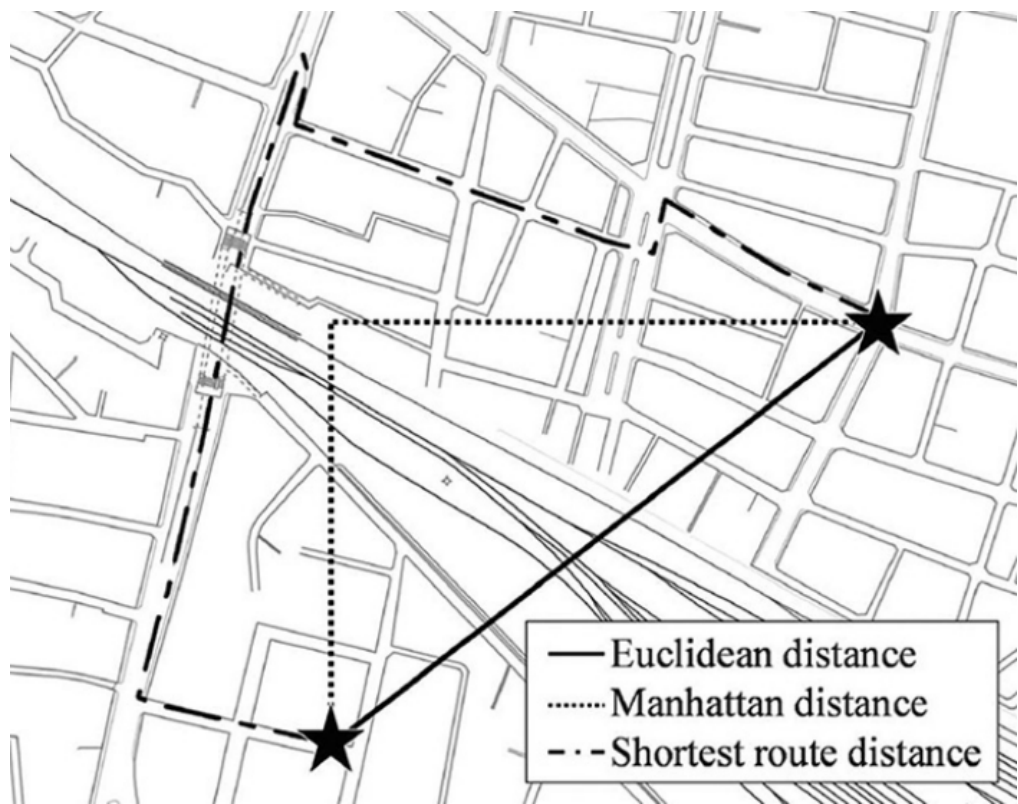
Figure 2.4: Euclidean, Manhattan and Real distance comparison

Source: Haginoya et al. (2020)

examples of its use in location problems are seen in Afrashteh et al. (2018) and Nguyen and Sepasian (2016). Levenshtein distance, also referred to as edit distance, is given by the minimum number of character operations (insertions, deletions or substitutions) necessary to transform a string into another. Finally, Hausdorff distance measures how far two subsets of space are from each other and it is mostly used in computer graphics - distances between 3D objects -, for further information the reader is referred to Mousavi et al. (2019) and Hu et al. (2006).

2.3
Location Models

As there is a vast range of different types of location problems, there are also many location models used to handle those problems. Researchers are continually developing these models and changes and updates are not uncommon in order to make them more suitable for the real world, which in turn increases mathematical complexity and computational challenges. There is extensive literature about the various types of existing models, making it essential to divide them into categories. As stated by Turkoglu and Genevois (2020), we will approach in this work a division of the location models in nine main categories as illustrated in Figure 2.5.
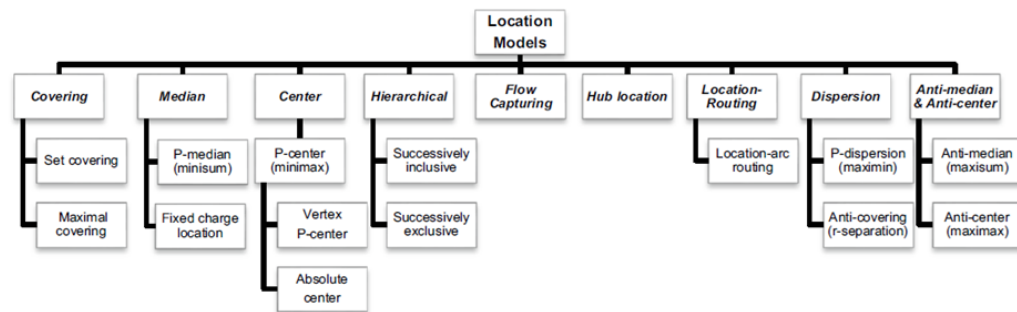
Figure 2.5: Categories of location models

Source: Turkoglu and Genevois (2020)

Covering location problems aim to cover a subset of customers, using a pre-defined number of facilities, depending on the distance between the customer and facilities. It is often categorized into two groups: Location Set Covering Problem - LSCP - and Maximal Covering Location Problem - MCLP (Farahani et al. 2012). Hierarchical facility systems consist of two or more levels of facilities, usually in terms of the types of services they provide. This way, a hierarchical system of facilities consists of various levels of service or facility and the demand nodes (Farahani and Hekmatfar 2009).

Ebery et al. (2000) explains that hub location problems have their origin in occasions where cargo (products or passengers) or information (data transmission) must be transported between an origin-destination pair of nodes, but where it is expensive or impractical to dedicate exclusive transport links to each origin-destination pair. Thus, a node is used as a hub - it is connected to other nodes to service all origin-destination node pairs. It is largely used in airline transportation, road systems and computer networks.

Farahani and Hekmatfar (2009) defines the Location-Routing Problem (LRP) as a combination of facility location (strategic policy) and vehicle routing (operational decisions). Being a composition of two difficult problems, LRP belongs to the class of NP-hard problems and heuristic approaches become the only practicable solution method to large problems. It is applicable to food and drink distribution, waste collection, newspaper delivery and others (Yu et al. 2010).

The Median and Center problems are the main location models among facility location problems, since most classical location studies focus on some aspects of these two approaches, with the choice of which model to use depending on a fundamental consideration: spatial efficiency or geographical equity (Ogryczak 1997). Center models aim to minimize the maximum distance between customers and facilities from a fixed number of facilities, hence the reason it is also known as minimax. It is also introduced under the title of p-center problem, where the objective is to find locations of p facilities so that all demands are covered and the maximum distance between a demand node and the nearest facility (coverage distance) is minimized, that is, the concern is about the worst case and we want to make it as good as possible (Farahani and Hekmatfar 2009). The minimax approach primarily considers geographical equity issues, which makes it used in emergency services, civil defense, and accident rescue (Ogryczak 1997).

The Median models aim to minimize the total distance between customers and facilities from a pre-determined number of facilities. Known as p-median problem, it aims to find the location of p facilities on a network so that the total cost is minimized (Daskin 1995). They are also called minisum and are mainly concerned with spatial efficiency. These problems are intended to find the median points among the candidate points so that the sum of costs can be minimized through the target function (Farahani and Hekmatfar 2009). Batista (2010) presents a diagram that helps to identify some models, as depicted in Figure 2.6. The problem addressed in this work and presented in the following chapters is characterized as an extension to a median location problem, since its main objective is to minimize the total distance traveled by the candidates.
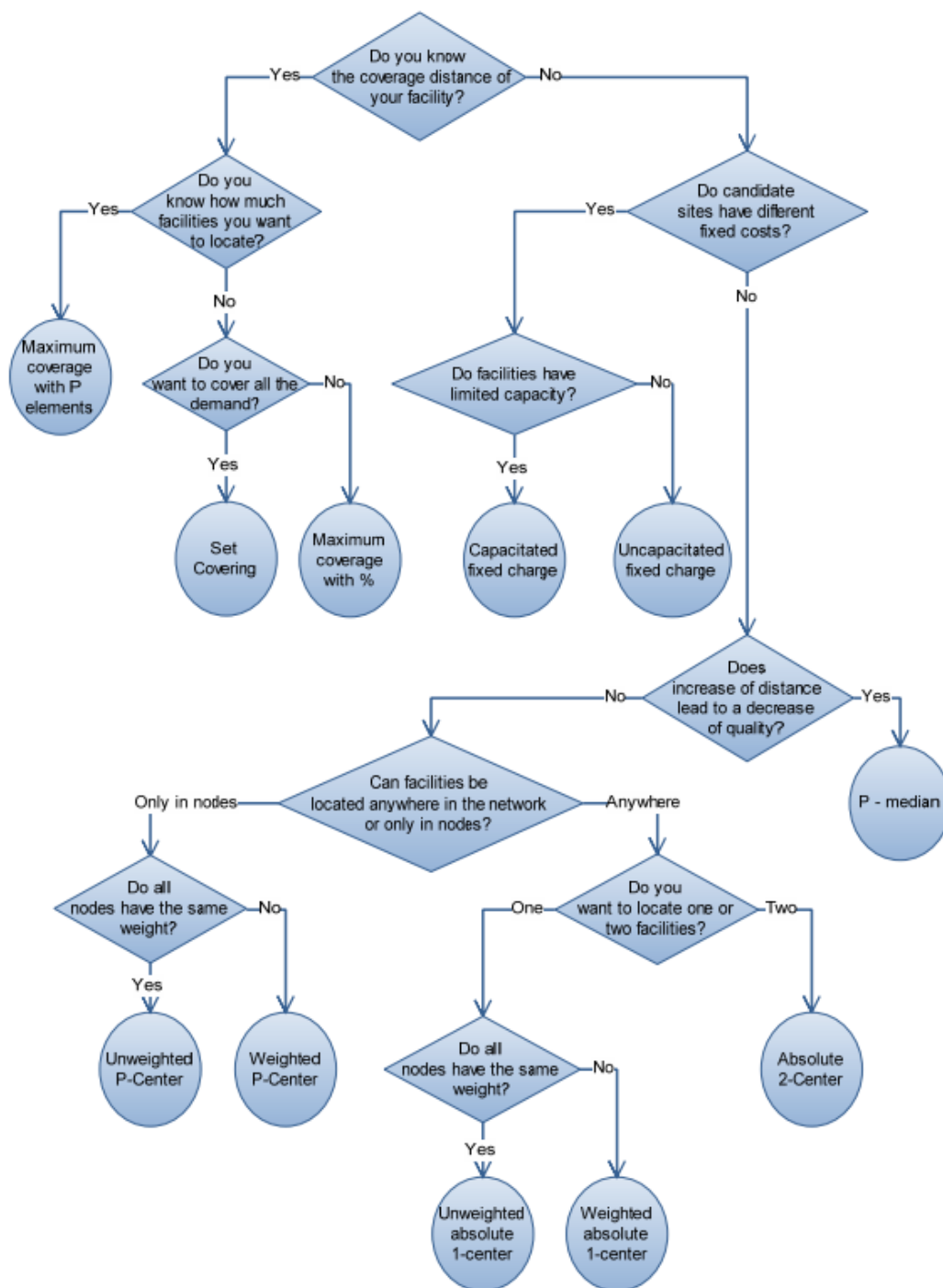
Figure 2.6: Question diagram for model identification

Source: Batista (2010)

2.4
Solution Methods

Facility location problems are, in most cases, classified as Mixed Integer Linear Programming - MILP, since there is in these models a binary variable about the installation decision of a facility: this variable assumes the value one if the facility is installed or zero otherwise.

From this, we can suppose a problem with N demand and P facilities and note that the set of possible solutions is a combination of P elements in N:

$$\binom{N}{P} = \frac{N!}{P!(N-P)!}$$

Simulating a small example (compared to problems with a large amount of data, commonly found in the literature and the real world) of N = 100 and P varying from 10 to 50, we can see in Table 2.1 exponential growth in the number of enumerated solutions.

Table 2.1: Combinatorial simulation

N	P	$\binom{N}{P}$
100	10	1.73×10^{13}
100	15	2.53×10^{17}
100	50	1.01×10^{29}

This simple example illustrates, as confirmed in the works of Owen and Daskin (1998) and Daskin (1995), that general MILP problems have NP-Hard complexity, so the same goes to the facility location problems and, specifically, the *p*-median problems, as stated by Kariv and Hakim (1979). Due to its complexity, there are significant studies on which methods are best suited to solve each type of problem and its efficiency in doing so. These methods are generally separated into two groups: exact methods and heuristics.

Exact methods are those in which the optimal global solution is found, when feasible; heuristics methods, instead, aim to find a satisfactory solution, whether it is the best solution possible or not. For this reason, they are considered to be, respectively, precise and approximate methods. Generally, exact methods are most used in limited complexity problems, while heuristics are most applied in high mathematical complexity problems. Both methods might be applied in facility location, depending on the problem's characteristics and the method to be used. A combination of the two techniques can be applied, such as, for example, solutions that start with exact methods and are concluded

with heuristic approaches to obtain desirable optimal solutions (Adeleke and Olukanni 2020).

Among the exact methods, the branch-and-bound is the most widely used tool for solving large-scale NP-hard optimization problems (Clausen 1999). Its algorithm seeks the best solution in an entire space of possibilities, but as total enumeration is impracticable due to the exponentially increasing number of potential solutions, it builds a decision tree with lower and upper bounds in order to avoid that bad solutions are explored. It is named after the strategy used to create restrictions through ramifications in a tree - branching - which represents subsets of the solution set, while the optimal solution of each branch is compared to lower and upper limits - bounds. Variations of branch-and-bound, as branch-and-cut and branch-and-price, are used to solve problems with specific characteristics. More information about these methods can be found in Ralphs et al. (2003) and in Dasgupta et al. (2006) where is presented a complete course about algorithms and solution techniques. The solver used in this work and described in Section 4.4 applies the branch-and-bound method.

2.5
Geographic Information System

A Geographic Information System is a structured system that stores, process and distribute spacial data. It was first introduced by Tomlinson (1967) in service for the Canadian government and evolved expressively, mainly due to the advent of the Global Positioning System - GPS -, to a collaborative tool that acquires geographic information through voluntary participation from its users.

A GIS can show many different kinds of data on a map, such as streets, road directions and buildings. It expresses information about locations that can be presented as latitude and longitude, address, or postal code (National Geographic 2017). They are important platforms for facility location since they can handle a large amount of data and posses a graphic interface that allows easy identification and visualization of a map and its points of interest. Relevant examples of studies that use GIS in facility location problems are seen in Church (2002) and Panichelli and Gnansounou (2008).

2.6
The University Admission Exam Location Problem

Across the literature, it is possible to find works related to the planning of exam times, such as the University Exam Timetabling Problem (de Werra 1985), which, besides the exam times, also schedules rooms, people to watch

the tests, among others. Another type of university-related topic is the College Admission Problem (Gale and Shapley 1962), which asks a matching between schools and candidates, based on their preferences, part of what is commonly called the College-Choice Process (Cabrera and La Nasa 2000). Some works apply facility location models in the subject of Education (Antunes and Peeters 2000, Bruno et al. 2016), but with different objectives from the one discussed in this study.

To the best of our knowledge, no work deals directly with the University Admission Exam Location Problem - UAELP -, an original approach of the facility location problem addressed in this work in which the objective is to minimize the total travel distance of candidates of an entrance exam for a university. Some studies with similar content can be found, such as de Andrade and Silva (2011), which conducted research that addresses our study's subject, but with some limitations. Although it has the same objective - to reduce the candidates' travel distance - the approach is significantly different. To achieve the objective, they propose choosing the cities with a higher concentration of candidates in the vicinity, not involving the distances performed within those cities. Still, their approach does not assign the candidates to a location or city and, most importantly, it does not use the current exam data, but uses the data from the last selection process to plan the next one. This decision can present a significant limitation to the model, since there could be notable changes from one exam to another, mainly in the number of candidates, making the optimization ineffective. This difference can be seen in the article itself, in which only one in five simulations has shown significant improvement; in two of these simulations, there were even negative results or no gains.

The study developed by Kripka et al. (2011) shows strong similarities with this work, even presenting a resembling mathematical model. However, both works have entirely different purposes, with their approach aimed at reducing the travel distance performed by students within the area of a university and also not assigning individuals to a location, assigning subjects to classrooms instead.

Despite the lack of direct correlation, as the UAELP is an extension of the Single Source Capacitated Facility Location Problem (Sridharan 1995), a vast literature can be found for its base problem. There is a large production of works related to the facility location subject with different approaches, presenting a wide range of mathematical models, heuristic algorithms, and theoretical analyses (Klose and Drexl 2005, Owen and Daskin 1998). It is a very relevant topic when it comes to supply chain management, as shown by the in-depth analysis of Melo et al. (2009), but also especially in areas such as health care

(Burkey et al. 2012, Huang et al. 2010, Vieira et al. 2019) and humanitarian logistics (Cotes and Cantillo 2019, Lin et al. 2012, Murali et al. 2012). Moreover, like the approach presented in this work, it is widespread in facility location problems the use of geographic information systems, as in Matisziw and Murray (2009), Shen (2005). Finally, the reader is referred to Ulukan and Demircioglu (2015) for a survey of discrete facility location problems.

3
Methodology

Here, we propose a methodology to allocate candidates from a selection process to the respective exam locations available in the process. This methodology is divided into five consecutive and interdependent stages, as shown in Figure 3.1. The first stage is data collection, then the distances acquisition (distance between each candidate for each of the exam locations) and then the third step, which is the treatment of the thousands of distance values obtained. With that, we can apply the proposed mathematical model and perform the analysis of the results for its implementation in sequence.
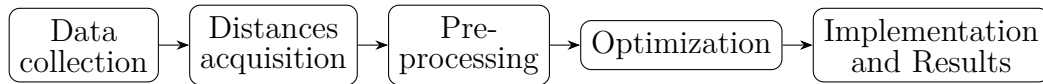
Figure 3.1: Stages of the Proposed Methodology

3.1
Data Collection and Distances Acquisition

These stages require the most time, they are quite important and should be done very carefully since any small error will spread to the next steps and amplify the failures throughout the process. Data collection must be performed with the institution responsible for carrying out the selection process and, therefore, it owns the necessary candidates' data. These data are: the addresses of the candidates, the type of test that each candidate will take (in case of different tests in the same selection process) and the addresses of the exam locations available for application.

With these data in hand, we can proceed to obtain the distances between the candidates and the exam locations. This is probably one of the main steps of the presented method and also the one that requires careful analysis due to the large amount of data involved. This stage is executed through the use of a GIS, a service that, between other things, provides distances between two addresses.

There is a significant range of tools that can provide this kind of service, with some of them charging for their use and others being completely free. Some examples of them are Google Maps, TomTom, Bing Maps, Open Street

Maps, QGIS and many others. It is important to highlight the need to write an automated code to perform the task of consulting these distances on the chosen platform, since a large number of distance queries will be obtained. A Geographic Information System platform usually offers an Application Programming Interface (API) to allow the access of a code to their service.

Commonly, this stage is divided into two steps with the need to use two different codes - allowing the query of the following steps to be performed in different platforms. The first step is the Geocoding - where physical addresses (description of street addresses) are transformed into geographic coordinates (latitude and longitude) -, and the second is the Distance Matrix - the calculation of distances between two points represented by the obtained coordinates. It is important to note that this calculation of distances made by the GIS is the real distance between the points considering the existing public roads between them, and not a simple Euclidean or geodesic distance.

3.2
Pre-processing

Once we have all the distance values between candidates and exam locations, it is necessary to analyze and process the acquired data. Pre-processing is an essential stage in order to determine what information will be used in the mathematical model by treating data inconsistencies and missing values. The first question that must be analyzed is what to do about the candidates for whom the GIS tool was not able to find some, or even all, of their distances to the exam locations. In this case, we decided to set to zero all distances from these candidates to the exam locations, but they will remain in the allocation since they still must be allocated. The amount of disregarded candidates (in terms of their distances) must be observed, because if the proportion of these candidates concerning the total of candidates is high, the allocation performed by the model may be impaired.

There are also lots of candidates who live too far from the exam locations, actually living in cities other than the exam cities or even in different states. These candidates will not travel from their homes to the exam locations on exam days due to the long distance required to do so; instead, they are going to travel days before to the exam cities and stay in a hotel or a friend's house to go to the exam location in the test days.

Given that it is not possible to obtain the candidates' information about where they will be staying and that their informed addresses will not be used for trips to the exam places on exam days, it is necessary to establish a cutoff parameter for the distance values. In this way, those candidates who live more

distant than the cutoff parameter will have their distance values set to zero, so they will be allocated in a test location, but the mathematical model will be indifferent to them, as their distances will not impact the objective function.

This cutoff parameter's value is open to discussion and computational tests or even sociological studies may be necessary to define the best value to fit the proposed solution. An initial idea is to consider a distance that involves the entire exam city, since the candidates who live beyond the city limit are more likely to travel days before.

In this matter, another concern is what to do with candidates whose distances to some exam places are greater than the cutoff parameter. The simplest thing to do would be just reset these distances, but in this case the same candidate would have some valid distances (the ones that are smaller than the cutoff parameter) and reset distances (the ones that are greater than the cutoff parameter), what would cause the mathematical model to tend to allocate that candidate to the exam locations that had the distance values set to zero but are, in fact, more distant than the locations that had the distant unaltered. This goes against the whole idea of using the cutoff parameter to make large distance values set to zero and, this way, make the model ignore the respective candidate. Thus, to address this issue, we propose that a candidate have all their distance values to exam location reset to zero only if each one of their distances is greater than the cutoff parameter. Thereby, we prevent that aforementioned distortion after the applying of the cutoff parameter.

3.3
Optimization

The core of the proposed mathematical formulation is an extension of the original formulation of the Single Source Capacitated Facility Location Problem (Sridharan 1995). Let C, L, and T be the set of candidates, locations, and exams types, respectively, obtained in the previous stages. The distance between the candidates and the exam locations is given by d_{ij}, where $i \in C$ and $j \in L$. The information about which type of test each candidate has to perform is given by the binary parameter cpr_{it}, where $t \in T$. The cost of using a location to offer any exam is c_j, and, if it is opened, it provides a capacity of Q_j candidates. We then have two binary variables, x_{ij} and y_j^t, representing whether candidate $i \in C$ is assigned to location $j \in L$, and whether location $j \in L$ offers the exam of type $t \in T$, respectively. Note that variables y_j^t are also responsible for controlling whether a location $j \in L$ is opened at all. The original integer programming formulation is presented as follows.

$$\min \sum_{i \in C} \sum_{j \in L} d_{ij} x_{ij} + \sum_{j \in L} \sum_{t \in T} c_j y_j^t \qquad (3\text{-}1)$$

subject to

$$\sum_{j \in L} x_{ij} = 1 \qquad\qquad \forall i \in C \qquad (3\text{-}2)$$

$$\sum_{t \in T} y_j^t \leq 1 \qquad\qquad \forall j \in L \qquad (3\text{-}3)$$

$$\sum_{i \in C} cpr_{it} x_{ij} \leq Q_j y_j^t \qquad\qquad \forall j \in L, t \in T \qquad (3\text{-}4)$$

$$x_{ij} \in \{0, 1\} \qquad\qquad \forall i \in C, j \in L \qquad (3\text{-}5)$$

$$y_j^t \in \{0, 1\} \qquad\qquad \forall j \in L, t \in T \qquad (3\text{-}6)$$

The objective function (3-1) minimizes the total distance between the candidates and the available exam locations. Constraints (3-2) force all candidates to be associated with exactly one location. Constraints (3-3) imposes a given place to offer only one type of exam, in order to reduce logistical complexity and avoid problems with candidates taking a different type of test than the candidate entered. Constraints (3-4) bind both variables and also limits the number of candidates in a given place. Variables domains are given by (3-5) and (3-6).

Some improvements were developed from the presented model with the objective to make its solution more efficient, obtaining quality results in significantly less time. To do so, two actions were taken: introduction of a valid inequality and the suppressing of generating the variables x_{ij} of the disregarded candidates. Since these candidates might be allocated at any location, there is no need for these variables to be generated, which reduces the total number of variables significantly and, as a consequence, reduces the running time. As explained by Avella et al. (2020), the valid inequality is redundant, but induces MIP solvers to derive knapsack cuts, which can effectively tighten the formulation. This way, the final model of the integer programming formulation developed - the one used in the optimization -, is presented.

$$\min \sum_{i \in C} \sum_{j \in L} d_{ij} x_{ij} + \sum_{j \in L} \sum_{t \in T} c_j y_j^t \qquad (3\text{-}7)$$

subject to

$$\sum_{j \in L} x_{ij} = 1 \qquad\qquad \forall i \in C \qquad\qquad (3\text{-}8)$$

$$\sum_{t \in T} y_j^t \leq 1 \qquad\qquad \forall j \in L \qquad\qquad (3\text{-}9)$$

$$\sum_{i \in C} cpr_{it} x_{ij} \leq Q_j y_j^t \qquad\qquad \forall j \in L, t \in T \qquad\qquad (3\text{-}10)$$

$$\sum_{j \in L} \sum_{t \in T} Q_j y_j^t \geq |C| \qquad\qquad\qquad (3\text{-}11)$$

$$\sum_{j \in L} Q_j y_j^t \geq Ct_t \qquad\qquad \forall t \in T \qquad\qquad (3\text{-}12)$$

$$x_{ij} \in \{0, 1\} \qquad\qquad \forall i \in C, j \in L \qquad\qquad (3\text{-}13)$$

$$y_j^t \in \{0, 1\} \qquad\qquad \forall j \in L, t \in T \qquad\qquad (3\text{-}14)$$

From the original model, it is added a new parameter Ct_t - number of candidates for type of exam - and two new constraints: (3-11) the valid inequality and (3-12) that imposes a lower limit related to the amount of places that offer the same type of exam. These improvements in the model made it much more efficient, with an expressive reduction in solving time around 40%.

4
Computational Experiments

4.1
The Real Problem

In the following sections, we show the proposed methodology applied to the data of the 2019 and 2020 editions and, from that, we compare the obtained results with the real data retrieved from the process in order to measure the efficiency and applicability of the proposed study. Then, with the approval of the university administration, we implemented the methodology in the real allocation used in the 2021 edition, significantly changing the historical allocation applied by UFJF. To verify the acceptance of the candidates about this new method, a survey was conducted, similar to the survey performed in the 2020 edition (presented in Section 1). As the methodology applied to the data of both editions of the selection process is the same, we describe below its application together.

In the 2019 edition of the selection process (PISM 2019), the exam was applied in four cities of two states for a total around of 34,300 candidates, allocated in 70 exam locations; in the following edition (PISM 2020) there were exams also applied in two states, but in five cities - the city of Petrópolis was included as an exam city for the first time -, with about 40,000 candidates distributed in 78 test locations.

In the current exam - PISM 2021 - there are 34,098 registered candidates, a reduction probably caused by the impact of the coronavirus pandemic and the closure of the schools since March. The exam cities, shown in Figure 4.1, are the same five exam cities of the previous edition and there are candidates from all regions of Brazil, with 77% of all candidates living in cities other than Juiz de Fora (Universidade Federal de Juiz de Fora 2020b).

It is important to emphasize that in this work, we aim to use only free tools in order to allow the implementation of this method by the university without additional cost, as well as to facilitate its adoption by other institutions. Still, using only free tools allows us to perform a wide range of tests at no expense, improving this study significantly.
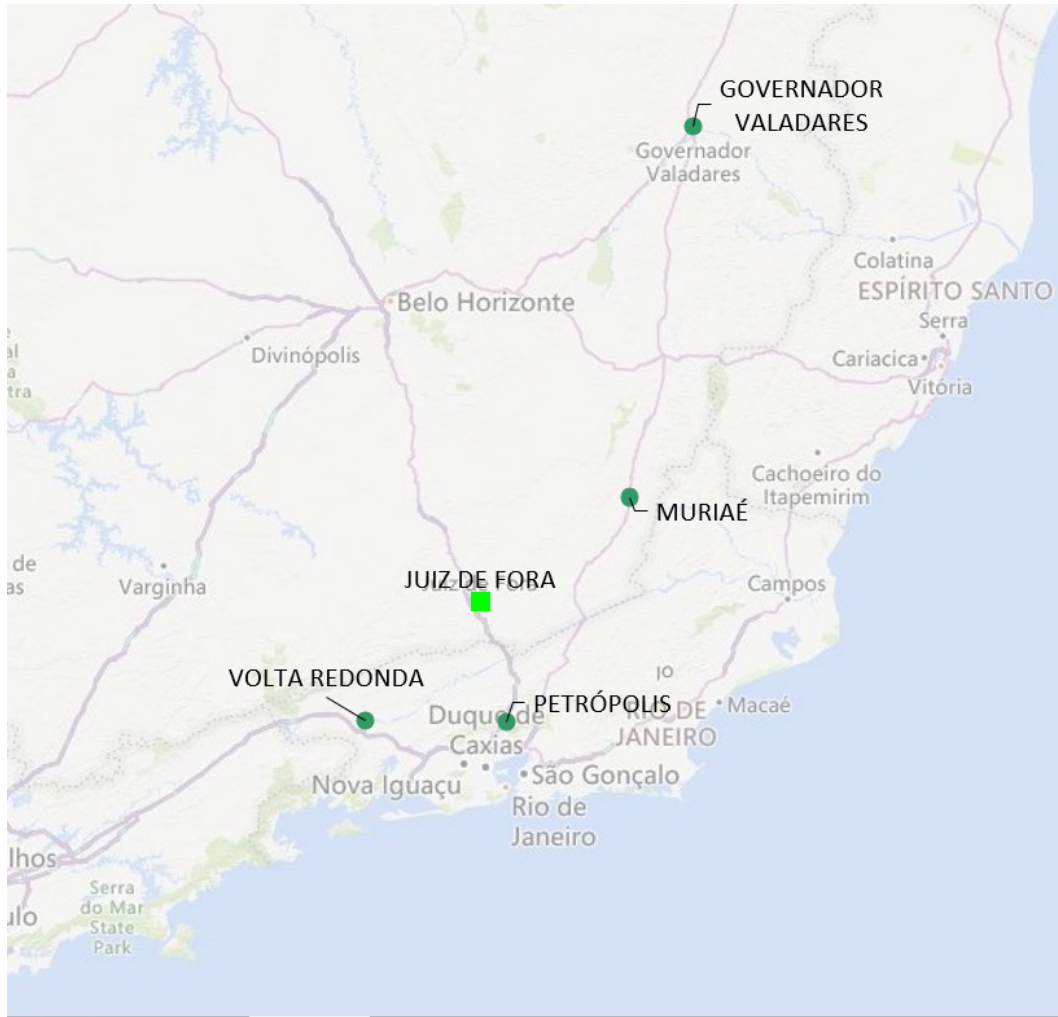
Figure 4.1: Exam cities of PISM

4.2
Data Collection and Distances Acquisition

It is necessary to draw attention to one of the main concerns of this work: strict respect for the personal data of candidates, taking into account the *Lei Geral de Proteção de Dados* - the Brazilian national law of data protection -, and best practices in the treatment and safeguarding of private information. Thus, we emphasize that there was no access to personal data, but to the addresses of the candidates, and this way there is no possibility of acquiring or identifying any other data of the applicants in the selection processes covered here.

The data collection was performed directly with UFJF and, it is worth highlighting, belong to 2019, 2020 and 2021 editions. The institution provided access to the main necessary information: candidates' addresses, original allocation of candidates, exam locations addresses and capacities, and type of test performed by each candidate. This last type of data is important due to

the fact that a relevant constraint of the model is the limitation that an exam location can offer only one kind of exam of the three existing types.

It is not any kind of place that can be used as an exam location. This site must have at least five classrooms, each of which can accommodate at least fifteen candidates. To meet these requirements, schools or colleges are often used as examination sites.

Not all candidates and exam locations were included in our model, since there are special cases in which no optimization regarding the distances are applicable. These cases consist of two types of candidates: those who for religious reasons cannot perform tests on Saturdays - which are a regular day of exam -, and those who have special needs due to mental or physical disability. These candidates are allocated in specific locations, in each city, destined exclusively to attend them. They represent around 1% of all candidates, which makes their exclusion from the model not significant for the optimization as a whole.

From the possession of data provided by the university, we must start the stage of distance acquisition. To do so, as explained in the previous chapter, we need to use a Geographic Information System - GIS. The definition of which system to use is very important since the quality of the acquired data will impact all the stages ahead and influence the efficiency of the method.

In this stage, with the objective of using only free tools into account, we analyze various GIS platforms available. Most of the GIS offers an initial amount of queries at no cost and then start to charge the queries that exceed this initial amount. Among those are the most prominent and used GIS, like Google Maps and Bing Maps.

The stage of distance acquisition is divided into two steps: Geocoding and Distance Matrix. First, it is necessary to define the best-suited GIS platform to perform the Geocoding - the step of obtaining the geographical coordinates from physical addresses. After an objective search for an efficient and simple platform we came to two main options: Google Maps and Bing Maps. Then, tests were carried out with them in order to measure their effectiveness: first we chose 500 random addresses and, using just the postal code and country name, performed queries through both platforms. In this test, we noted an amount of 11,6% invalid queries - empty or notoriously incongruous values of geographic coordinates obtained - for Bing Maps and just 3% for Google Maps. Next, we did a test using the same 500 random addresses in a full format: street name, house number, neighborhood, city and country name. This test proved to be much more efficient, especially for Bing Maps that presented only 1% of invalid queries against 2,1% for Google Maps. This way, we decided to use the full description of addresses instead of just the postal code, since besides the better

results in the tests, using the postal code would be impracticable because in some of the cities of exam, mainly the smaller ones, a large variety of streets have the same postal code.

From that, as Bing Maps showed higher efficiency in the experiments, there was just one thing left in order to define it as our geocoding platform in further tests: the limitation of free queries. As appointed early, we intend to use only free tools in this research to facilitate its replication with a minimal cost or at no cost at all, so through a search in the Bing Maps Licensing documentation (Microsoft 2020), we can see that for an educational or non-profit purpose is offered a basic key capable of performing 50,000 transactions within a 24-hour period, which is sufficient for our needs since the larger amount of data to perform the geocoding corresponds to candidates' addresses and does not surpass 41,000 transactions, what would allow us to obtain the geographical coordinates of an entire PISM edition in one single day. This was another advantage compared to Google Maps, that offers a free amount of initial queries but when this limit is exceeded it starts to charge the user (Google 2020).

With a defined tool to perform the geocoding the next step is to choose the platform to perform the distance matrix. This step is the one that requires a larger number of queries, since it consists of the multiplication of each exam location to each candidate. This way, we obtain a distance matrix between all candidates and exam locations to be used as an input parameter in the mathematical model. The amount of distance values to be acquired in this step for both editions is summarized in Table 4.1.

Most GIS platforms offer both services: geocoding and distance matrix. Nevertheless, the amount of queries needed varies a lot for each step: for the city of Juiz de Fora, for example, as 24,000 queries are necessary for geocoding, more than 1,000,000 are required for the distance matrix. This way, it was not possible to use the same platform to perform both steps once the limitation of free queries in Bing Maps meets the needs for the geocoding, it is quite low when compared to the millions of queries needed for the distance matrix.

Thereby, as Bing Maps and Google Maps were not available options, we expanded our search to find a suitable platform to execute the distance matrix. So, after extensive work, we came across Open Route Service. It is a platform maintained by Heidelberg Institute for Geoinformation Technology (HeiGIT) group, supported by the Klaus Tschira Foundation Heidelberg (KTS), and it is part of the GIScience Research Group at the Department of Geography, within the University of Heidelberg (Open Route Service 2020a), having available a high-quality group of researchers to develop and improve their own algorithm. Founded in 1386, the University of Heidelberg is the oldest

Table 4.1: Amount of distances

PISM 2019	Total Locations	Total Candidates	Distance Values
Volta Redonda	6	2,388	14,328
Muriaé	9	3,617	32,553
Governador Valadares	6	3,764	22,584
Juiz de Fora	50	24,587	1,229,350
Sum	71	34,356	1,298,815
PISM 2020	Total Locations	Total Candidates	Distance Values
Volta Redonda	7	2,607	18,249
Petrópolis	7	3,681	25,767
Muriaé	7	4,433	31,031
Governador Valadares	8	5,091	40,728
Juiz de Fora	45	23,544	1,059,930
Sum	74	39,366	1,175,705
Total	145	73,722	2,474,520

university in Germany and one of the strongest research universities of the continent according to Times Higher Education Ranking, figuring third place in the country, 12th place in Europe and 42nd in the world (Heidelberg University 2021).

Open Route Service has the objective to facilitate the transfer of knowledge and technology from academia to practical applications and professional services, with a focus on the analysis, refinement, enrichment and usage of spatial data for innovative applications and services. Its API consumes user-generated and collaboratively collected free geographic data, directly from Open Street Map, and offers a global coverage (Open Route Service 2020a), but, mainly, it has a large limitation of queries to its distance matrix of 500 requests per day, with each request consisting of a matrix of 50 x 50, which reaches over 1 million queries in just one day.

Despite being an efficient platform, Open Route Service could not be used for both steps of distance acquisition, because its geocoding API offers only 1,000 requests per day, making its use impracticable for this work. Thus, after a long search and multiple tests, we finally defined the best suitable GIS platforms to perform the stage of distance acquisition: Bing Maps for geocoding and Open Route Service for distance matrix.

From the candidates' addresses and exam locations we created an algorithm using Bing Maps (2020) documentation to execute the geocoding.

Such algorithm was written in Python language with the objective to allow a very fast acquisition of the geographic coordinates through a Microsoft API, using a Bing Maps Basic Key of non-profit type. This way, in a few hours for each city, the code was able to collect thousands of the demanded coordinates. For Juiz de Fora, which has almost 60% of all addresses of the process, the code took almost half a day to perform all the queries, which - considering the enormous quantity of data involved - shows its efficiency and simplicity. The time required to perform the geocoding step for each city and selection process edition are shown in Table 4.2, as well as the quantity of addresses queried - the number of candidates plus the number of exam locations for each city.

Table 4.2: Geocoding information

	PISM 2019		PISM 2020	
	Addresses	Time	Addresses	Time
Volta Redonda	2,393	2h0min15s	2,613	1h20min13s
Petrópolis	-	-	3,687	1h52min34s
Muriaé	3,626	2h35min18s	4,439	2h13min49s
Gov. Valadares	3,770	2h34min08s	5,098	2h35min21s
Juiz de Fora	24,637	15h59min16s	23,586	15h07min13s

It is worth noting that not every request returns a valid coordinate, with very few of them returning an empty value, this might occur due to imprecision of the platform used or due candidates' mistake in the information provided during the registration application. Despite this, the number of empty coordinates after geocoding is quite low, not surpassing 0.1%. A substantial matter in this stage is to verify the integrity of the exam locations' coordinates acquired, since in sequence every single distance between thousands of candidates and dozens of exam locations will be calculated. This way, we examined all exam locations' coordinates and ensured that the data retrieved related to them was valid and correct.

Possessing every geographic coordinate needed - for candidates and exam locations - we can move further to the distance matrix step. As in the previous step, we also created an algorithm to execute this task in an automated and fast way through an Open Route Service API key, again the algorithm was developed using Python language from Open Route Service (2020b) documentation. This code is much more complex than the previous one, since it requires the partition of input data into blocks of 50 lines due to Open Route Service restrictions and the subsequent assembly of the acquired information into one large matrix

containing the distance values from each candidate for each one of the possible exam locations.

The amount of data involved in this step is way larger than in the geocoding, whereas - for each city - the number of distance values is composed of the number of candidates times the number of exam locations and, due to the cutting edge technology embedded in Open Route Service, it took just a few minutes to complete. The information about the distance matrix step is in Table 4.3, where we can see the efficiency of the platform through the quite low percentage of distances it was not able to obtain.

Table 4.3: Distance Matrix information

PISM 2019	Total Distances	Distances not found	% not found
Volta Redonda	14,328	384	2,68%
Muriaé	32,553	342	1,05%
Governador Valadares	22,584	210	0,93%
Juiz de Fora	1,229,350	10,600	0,86%
PISM 2020	Total Distances	Distances not found	% not found
Volta Redonda	18,249	70	0,38%
Petrópolis	25,767	35	0,14%
Muriaé	31,031	119	0,38%
Governador Valadares	40,728	128	0,31%
Juiz de Fora	1,059,930	2,655	0,25%

The impressive low number of distances not found by Open Route Service is due to its advanced technology, but to achieve these numbers was also necessary a specific treatment of the data. ORS has a limit in its API of 6,000 Km, if during a request - that is divided into blocks of 50 x 50 as already mentioned - one of the distances calculated exceeds that limit an error is attributed to all distances of such block of addresses, which can cause a significant increase in the number of distances not found.

In order to solve this issue we established boundary conditions for every pair of coordinates: latitudes must be between -35° and 0° and longitude must be between -70° and -35°. These coordinates ranges define a geographic quadrant where Brazil is contained, which guarantees that no calculated distance will exceed the API limit of 6,000 Km, since the extreme points within Brazil's limits do not reach 4,500 Km. Hence, the pair that posses at least one coordinate that does not meet the boundary conditions has its both coordinates changed

to -50°, which represents a spot in the middle of the Atlantic Ocean and will make ORS disregard that pair and turn its output in an empty value.

As an example, for the city of Juiz de Fora in the 2020 edition, the case with most data, we had 94 coordinates pairs that did not meet the boundary conditions and then they were changed to -50°, which generated to these pairs an empty value of distance in the distance matrix. This is a small number compared to the total number of input data but shows the importance of this treatment, because each of these changed pairs, if not treated, could generate a worst-case scenario of 235,000 empty values in the distance matrix (94 inputs times a block of 50 x 50), which would represent an alteration from 0.25% of distance values not found in the experiment to 22.17%, an unsustainable increase in errors and disregarded candidates that would compromise the entire methodology.

4.3
Pre-processing

The pre-processing stage is an essential one in order to treat the vast number of data acquired in the previous stages, as well as eventual errors produced by the GIS platforms used. As now, we have the distance matrix for each exam city with thousands, and in some case millions, of distance values we once again developed an algorithm in Python language to eliminate any inconsistent and unnecessary data.

The primary objective of this code is to discard candidates that live too far from exam locations, that is, candidates that due to the large distance they will need to travel, it is impossible for them to go to test sites and return home on the same day, being necessary to stay overnight in the exam city, whether in a friend's house or a hotel. This way, these candidates will be part of the mathematical model but will not have any impact on the objective function, since they are not going to travel the distance assigned to them.

An important question related to this objective is the definition of the distance cutoff parameter, which above it the candidate' distance values of the will be disregarded: such distance values will be set to zero and, this way, they will not be taken into account in the mathematical model. Some proposals were made and after studying the exam cities and the distance matrix obtained, we noted that all candidates who live in the exam cities - resident candidates - have their distance values smaller than 50 kilometers, as shown in Table 4.4. There are significant variations depending on the area of the city, but 50 km remains a plausible distance to be traveled within a day: Volta Redonda, for example, has just 182 km^2 of area, which is equivalent to near 13.5 km from

end to end, while Governador Valadares has an area of 2,342 km^2, which is close to 49 km from end to end.

Table 4.4: Candidates of PISM 2020 within the cutoff parameter

PISM 2020	Resident Candidates	Candidates with distances < 50 km
Volta Redonda	503	1,248
Petrópolis	1,631	1,734
Muriaé	738	923
Governador Valadares	1,620	1,687
Juiz de Fora	9,605	10,978

As a better allocation will impact mainly the candidates who live some way close to the exam locations - primarily candidates who live in the exam city - we decided, therefore, to establish the value of 50 kilometers as the cutoff parameter. After applying the pre-processing algorithm, as explained in Section 3.1, candidates who possess all their distance values greater than the cutoff parameter had their values set to zero and will, therefore, be disregarded in the optimization - that is, those candidates will have no impact in the objective function of the optimization, since their all their distances are zero, but will be allocated in any of the exam locations.

In Table 4.5, we can see that the number of disregarded candidates varies significantly for each exam city, as we have cities that have most candidates coming from surrounding towns and other cities that have a greater number of candidates living within their limits. So, by the end of this stage, we were able to verify the amount of valid candidates for each exam city, that is, the amount of candidates who will have indeed their distance values considered in the mathematical model, such data is also present in Table 4.5. It is important to note that the empty values retrieved from the distance matrix also were set to zero in order to prevent any kind of computational error in the following stages and, as a consequence, the respective candidates are also disregarded.

Table 4.5: Disregarded and Valid Candidates

PISM 2019	Disregarded Candidates	Valid Candidates	% Valid Candidates
Volta Redonda	1,296	1,092	45.73%
Muriaé	2,924	693	19.16%
Governador Valadares	2,627	1,137	30.21%
Juiz de Fora	13,932	10,655	43.34%
PISM 2020	Disregarded Candidates	Valid Candidates	% Valid Candidates
Volta Redonda	1,359	1,248	47.87%
Petrópolis	1,947	1,734	47.11%
Muriaé	3,510	923	20.82%
Governador Valadares	3,404	1,687	33.14%
Juiz de Fora	12,576	10,978	46.61%

4.4
Optimization and Results

In the last stage of the proposed methodology, we execute the mathematical optimization. To perform such a task, we implemented the model presented in Section 3.3 using Julia language through JuMP library (Kwon 2019). JuMP has at its disposal a great range of solvers, with more than two dozens of available choices between paid and free tools (JuMP Developer 2020). Of all solvers that offer a free license, only two supports MILP - Mixed-Integer Linear Programming, the kind of problem we intend to approach. These solvers are Cbc and GLPK.

Cbc - Coin-or Branch and Cut - is an open-source mixed integer programming solver written in C++ by COIN-OR Foundation, Inc., a non-profit educational and scientific foundation with the mission to create and disseminate knowledge related to all aspects of computational operations research (COIN-OR Foundation 2020). GLPK - GNU Linear Programming Kit - is intended for solving large-scale linear programming, mixed integer programming, and other related problems. It is a set of routines written in ANSI C and organized in the form of a callable library, being part of the GNU Project (GNU Project 2020). For mixed integer problems, GLPK uses a branch-and-bound algorithm together with cutting-plane method.

In order to define the best suitable solver for our purpose, we performed tests with both of them in a smaller instance of the mathematical model. Cbc proved to be very inefficient in our tests, not being able to obtain any result or

even run properly. In contrast, GLPK was able to reach reliable results in all tests performed in a reduced time window.

The structure of the optimization code is then defined: the solver is called, as well as any other library needed; the input parameters to be read are the matrix distance table, the table that discriminates which kind of exam each candidate has to perform, the capacity of each exam location and its opening cost. It is important to highlight how relevant is the choice of the opening cost values is to the model: values too low can make the model open exam locations that were not necessary - which is an undesired outcome due to logistical costs -, while medium or very high values can make the model take more time to find a feasible solution. In our experiments, we used the same opening cost to all exam locations - since there are no preferable places - and the value was 1,000,000,000, which proved to be more efficient in the various tests performed. Next, we declare the variables, set the objective function and add the constraints. Finally, there is the command to execute the optimization with the output being two tables: one containing the exam locations used and the other containing the candidates assigned to each exam location.

All the following experiments were performed using JuMP and GLPK solver in a machine with an Intel Core i5 6400 2.7GHz with 16GB of RAM. First, we ran tests for the selection process already applied, PISM 2019 and 2020 editions. This way, we are able to compare its results with the information about the previous exams and precisely measure the impact of our methodology, as well as its efficiency and applicability.

For each exam city an independent test was executed in both exam editions. The running time for four of the cities, the smallest ones, were quite low, from a few seconds for Volta Redonda to 12 minutes and 9 seconds for Governador Valadares. For Juiz de Fora, the largest city of the exams with the majority of candidates, for both PISM editions, the model did not reach an optimal solution, so a time limit of 5 hours was established. This way, for PISM 2019 the optimization model obtained a sub-optimal value with a gap of 1.3% and for PISM 2020 the model presented a gap of 3.5%.

Subtracting the cost to open an exam location from the objective function we obtain the total displacement performed by the candidates of each city, as well as the average displacement. We then are able to compare the optimized displacements with the real displacement performed in the previous exams and analyze if there is any reduction in the total travel distance of the candidates and how significantly such reduction might be. All these information are summarized in Table 4.6.

As we can note from the obtained results, there was a significant reduction

Table 4.6: Travel distances and Reductions

PISM 2019	Total Disp.(m)		Average Disp.(m)		Reduction
	Original	Optimized	Original	Optimized	
Volta Redonda	21,492,954	20,228,923	19,682	18,525	5.88%
Muriaé	9,835,128	8,561,819	14,192	12,355	12.95%
Gov. Valadares	7,142,587	6,536,421	6,282	5,749	8.49%
Juiz de Fora	136,111,307	80,277,775	12,774	7,534	41.02%
Total	174,581,976	115,604,938	12,858	8,514	33.78%
PISM 2020	Total Disp.(m)		Average Disp.(m)		Reduction
	Original	Optimized	Original	Optimized	
Volta Redonda	24,174,958	22,113,438	19,371	17,719	8.53%
Petrópolis	21,863,103	19,331,781	12,608	11,149	11.58%
Muriaé	11,451,049	10,438,205	12,406	11,309	8.84%
Gov. Valadares	9,934,014	7,511,815	5,889	4,453	24.38%
Juiz de Fora	148,741,088	99,164,921	13,549	9,033	33.33%
Total	216,164,212	158,560,160	13,046	9,569	26.65%

in the total distance travel in the experiments when compared to the total travel distance originally performed by the candidates in each exam edition. This reduction goes from 5% for Volta Redonda to 41% for Juiz de Fora. These data show that the proposed optimization proved to be highly effective, expressively improving the candidates' allocation, allowing them to be assigned to an exam location much closer to their residence than in the original allocation. If we consider the data from the two editions of the exams together, it is possible to state that the methodology presented an overall reduction in the distance covered by almost 30%.

Yet, it is possible to discriminate the results and perform a detailed analysis for the valid candidates. Thereby, we compare the optimized allocation of valid candidates with their original allocation, then we separate them into three groups for both exam editions: candidates who have their allocation improved and candidates who have their allocation unchanged - both in Table 4.7; and candidates who have their allocation penalized - those who were allocated more distant in the optimized model than in the original allocation - in Table 4.8. In those tables, there is also the proportion of each type of candidate and the average gain and loss in travel distance.

With this analysis, we are able to measure more precisely the effectiveness of the optimization and observe its specific impact on the candidates of each city. Through the data of the candidates who have their allocation maintained,

Table 4.7: Valid candidates with improved and mantained allocation

PISM 2019	Improved Allocation			Maintained Allocation	
	Cand.	Average Disp. (m)	Average Gain (m)	Cand.	Average Disp. (m)
Volta Redonda	576 *52.75%*	20,034	2,228 *11.12%*	475 *43.50%*	17,882
Muriaé	583 *84.13%*	12,447	2,408 *19.35%*	38 *5.48%*	1,553
Governador Valadares	500 *43.98%*	5,632	1,402 *24.89%*	546 *48.02%*	5,869
Juiz de Fora	9,778 *91.77%*	7,541	5,782 *76.67%*	520 *4.88%*	5,307
PISM 2020	Improved Allocation			Maintained Allocation	
	Cand.	Average Disp. (m)	Average Gain (m)	Cand.	Average Disp. (m)
Volta Redonda	999 *80.05%*	17,367	2,265 *13.04%*	77 *6.17%*	13,211
Petrópolis	1,243 *71.68%*	10,896	2,198 *20.17%*	204 *11.76%*	6,353
Muriaé	597 *64.68%*	9,969	1,787 *17.93%*	278 *30.12%*	14,635
Governador Valadares	1,222 *72.44%*	4,318	2,302 *53.31%*	171 *10.14%*	2,565
Juiz de Fora	10,134 *93.95%*	9,078	4,836 *53.27%*	464 *4.23%*	7,727

we note that just a small amount of the candidates were originally allocated in the closest exam location possible - this amount does not surpass 10% for most cities - which makes clear that in fact there was considerable room for improvement in the original allocation used so far.

It is also possible to observe the high number of candidates that had their allocation improved - almost 79% on average for 2020 edition - and the surprisingly positive numbers of the average gain in displacement for these candidates, which reaches, for example, 53.27% for Juiz de Fora, a reduction of almost 5 km in the travel distance for each candidate on average. Another data that is worth highlighting is the minor amount of candidates who had their allocation penalized in order to achieve the optimal total travel distance for each city.

To further enhance the analysis we present Tables 4.9 and 4.10, where we separate candidates who were originally assigned to an exam location within grouped distance ranges of 12.5 km. For each range, there is the average original and optimized distances in meters, and the total distance reduction, for both exam editions. This way, we are able to examine the displacement of candidates

Table 4.8: Valid candidates with penalized allocation

PISM 2019	Penalized Allocation		
	Candidates	Average Disp. (m)	Average Loss (m)
Volta Redonda	41	4,772	472
	3.75%		*9.89%*
Muriaé	72	17,305	1,813
	10.39%		*10.48%*
Governador Valadares	91	5,668	1,040
	8.00%		*18.35%*
Juiz de Fora	357	10,591	1,971
	3.35%		*18.61%*
PISM 2020	Penalized Allocation		
	Candidates	Average Disp. (m)	Average Loss (m)
Volta Redonda	172	21,784	1,170
	13.78%		*5.73%*
Petrópolis	287	15,654	698
	16.55%		*4.46%*
Muriaé	48	8,174	1,124
	5.20%		*12.90%*
Governador Valadares	294	6,113	1,330
	17.43%		*21.76%*
Juiz de Fora	200	9,724	1,512
	1.82%		*15.55%*

for each range carefully and note that the most benefited for the optimization are the candidates who live close to exam locations, which is an expected outcome, since those candidates are the ones who are logically more impacted for any alteration in the allocation due to the fact that it might lead to a significant change, for example, between a location near their neighborhood and a location across the city. In contrast, as the distance range increases, the impact of the reduction decreases, once a few kilometers of improvement in the allocation has a limited benefit for candidates who live far from any exam locations.

In order to allow the analysis of average travel distances considering the distance ranges of all candidates of previous PISM editions, Figures 4.2 and 4.3 are presented, where the average travel distances - original and optimized - of all candidates for an entire selection process are visualized in bars, separated for

Table 4.9: Average travel distances considering distance ranges of PISM 2019

PISM 2019		Average travel distance		
	Range	Original (m)	Optimized (m)	Reduction
	[0.0, 12.5]	5,959	5,111	14.23%
Volta	[12.5, 25.0]	15,806	13,618	13.84%
Redonda	[25.0, 37.5]	35,199	34,187	2.88%
	[37.5, 50.0]	44,454	43,178	2.87%
	[0.0, 12.5]	3,526	1,900	46.11%
	[12.5, 25.0]	22,051	20,242	8.20%
Muriaé	[25.0, 37.5]	32,623	30,554	6.34%
	[37.5, 50.0]	42,925	40,633	5.34%
	[0.0, 12.5]	4,451	3,938	11.53%
Governador	[12.5, 25.0]	19,584	18,812	3.94%
Valadares	[25.0, 37.5]	30,427	29,716	2.34%
	[37.5, 50.0]	44,732	43,868	1.93%
	[0.0, 12.5]	6,065	2,680	55.81%
	[12.5, 25.0]	16,639	8,506	49.78%
Juiz de Fora	[25.0, 37.5]	30,667	14,629	52.30%
	[37.5, 50.0]	44,614	38,386	13.96%

distance ranges of 12.5 km each. In addition, the percentage reduction between original and optimized average distance is also shown. Through these graphs, the notable reduction in the displacement of candidates is confirmed, especially for those who were originally allocated up to 25 km from their addresses, with a reduction in the travel distance from 35% to almost 50%. As the originally allocated distance grows, we see that the percentage reduction in displacement decreases, but still quite expressive, being around 9% to 32%.

Using another metric, through a meticulous analysis, it is possible to verify the original amount of candidates who were able to go to their exam locations on foot and its respective optimized amount, then we can compare the number of these candidates throughout all the selection process editions. First, we must define the plausible distance limit that a person is able, and willing, to travel on foot. Studies calculated that the ideal distance between residents and transport stations - a routine walk - varies from 0.8 km (El-Geneidy et al. 2013) to 1 km (Instituto de Políticas de Transporte e Desenvolvimento 2015), so we defined the double of that value - 2 km, which is equivalent to nearly a 25-minute walk -, as the distance limit to go to exam location on foot. This value was defined considering that taking a test is not an everyday event and,
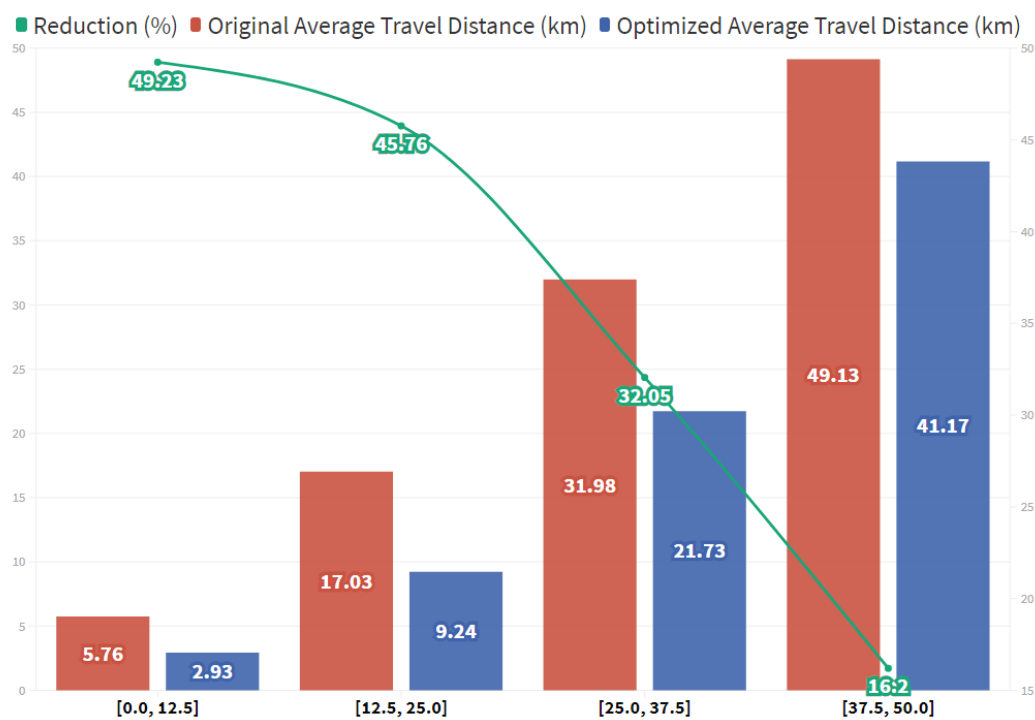
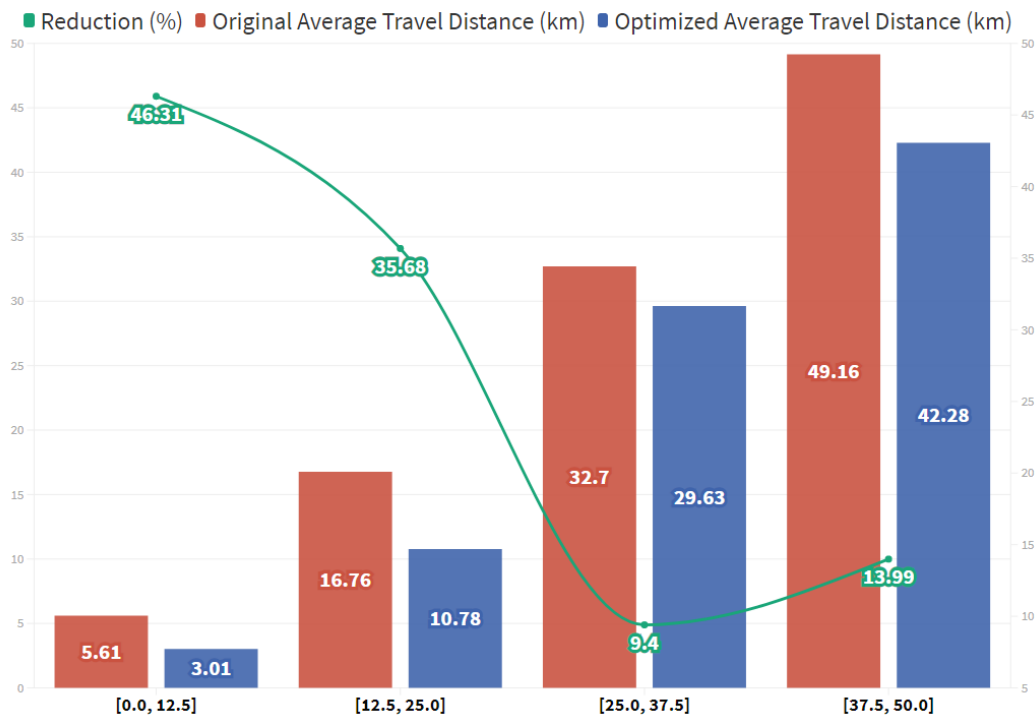Figure 4.2: Average distances of all candidates for PISM 2019

Figure 4.3: Average distances of all candidates for PISM 2020

Table 4.10: Average travel distances considering distance ranges of PISM 2020

PISM 2020		Average travel distance		
	Range	Original (m)	Optimized (m)	Reduction
Volta Redonda	[0.0, 12.5]	6,223	4,875	21.66%
	[12.5, 25.0]	15,034	12,713	15.44%
	[25.0, 37.5]	35,529	35,206	0.91%
	[37.5, 50.0]	43,680	41,461	5.08%
Petrópolis	[0.0, 12.5]	5,257	4,065	22.67%
	[12.5, 25.0]	17,540	16,354	6.76%
	[25.0, 37.5]	30,052	27,242	9.35%
	[37.5, 50.0]	44,505	43,215	2.90%
Muriaé	[0.0, 12.5]	3,216	2,113	34.30%
	[12.5, 25.0]	22,354	20,881	6.59%
	[25.0, 37.5]	32,758	31,873	2.70%
	[37.5, 50.0]	42,859	41,430	3.33%
Governador Valadares	[0.0, 12.5]	4,168	2,739	34.29%
	[12.5, 25.0]	18,571	17,772	4.30%
	[25.0, 37.5]	36,766	36,137	1.71%
	[37.5, 50.0]	41,364	39,601	4.26%
Juiz de Fora	[0.0, 12.5]	6,101	2,817	53.83%
	[12.5, 25.0]	16,696	9,494	43.14%
	[25.0, 37.5]	32,210	26,042	19.15%
	[37.5, 50.0]	44,951	36,065	13.09%

therefore, the candidates would be more willing to perform a walk greater than that an ideal one, even more in order to avoid complications in traffic.

That said, is presented in Table 4.11, the number of candidates that are considered to be able to go to the exam locations on foot - those within the distance of 2 km. We can observe that there was a very expressive increase - more than four times - in the total amount of candidates who, after the optimization, could potentially go to their exam locations on foot. If we analyze the data separately for each city, the improvement is even more positive, with the city of Juiz de Fora for both editions presenting a surprising increase greater than 700% in the number of candidates who, after optimization, were allocated within the distance of 2 km.

Table 4.11: Candidates able to travel on foot

	PISM 2019			PISM 2020		
	Original	Optimized	Increase	Original	Optimized	Increase
Volta Redonda	57	89	56.14%	57	121	112.28%
Petrópolis	-	-	-	141	299	112.06%
Muriaé	109	285	161.47%	176	332	88.64%
Gov. Valadares	194	325	67.53%	307	661	115.31%
Juiz de Fora	456	3,666	703.95%	401	3,421	753.12%
Sum	816	4,365	434.93%	1,082	4,834	346.77%

4.5
Implementation in the most recent selection process

As the experiments of applying the developed methodology in the 2019 and 2020 editions of PISM presented successful and consistent results, we received the permission of UFJF's administration to implement our work in its next selection process - the PISM 2021 edition - to be applied in February (tests of Modules 2 and 3) and March (tests of Module 1) of 2021. It is important to remind that the exam is divided into three modules with different content, one for each of the years in high-school.

Due to the Covid-19 pandemic, this edition will, exceptionally, be applied on two weekends - when the habitual is the application in just one weekend, usually in December of the previous year. It was decided to assure the attendance of biosafety protocols adopted by UFJF in order to avoid the spread of the Sars-Cov-2 virus, which demands the utilization of half of the location's capacity to guarantee social distancing between candidates. Thus, the application of the 2021 edition was split in two: in the first weekend it will be offered tests for candidates registered in the second and third modules, and two weeks later there will be tests for candidates registered in the first module. Such division was elaborated because the amount of candidates of the first module is close to the sum of candidates of the second and third modules, so the same available locations in one application weekend could be used in the next one.

From the data collected from UFJF - again with complete respect to private data and no access to candidate's personal information - we performed the proposed methodology as we did for the previous editions. The main difference in this case is that, as the application is divided into two weekends for different candidates, we must execute independent models for each weekend of application.

As was done in previous editions, the step of geocoding was applied

through an algorithm in Python using Bing Maps. The time of execution of the algorithm varies from 49 minutes and 15 seconds to candidates of Module 1 from the city of Volta Redonda to 8 hours and 29 minutes to candidates of Modules 2 and 3 from the city of Juiz de Fora. With all the geographic coordinates we then pass to the distance matrix step and acquire all the necessary distances using Open Route Service. The results of this initial stage are summarized in Table 4.12, where we can note that the available locations for each city are the same, regardless of the modules. Comparing the data of distances acquisition for 2021 edition with previous editions in Table 4.3, it is possible to perceive a higher percentage of values not found for cities of Volta Redonda and Petrópolis in the current exam edition, but still within an acceptable range.

Table 4.12: Distance Matrix information of PISM 2021

Modules 2 and 3	Available Locations	Total Candidates	Distance Values	Values Not Found
Volta Redonda	6	1,159	6,954	2.76%
Petrópolis	9	1,781	16,029	4.27%
Muriaé	8	2,301	18,408	0.91%
Governador Valadares	8	2,579	20,632	0.66%
Juiz de Fora	49	11,743	575,407	0.63%
Sum	80	19,563	637,430	0.75%
Module 1	Available Locations	Total Candidates	Total Distances	Distances Not Found
Volta Redonda	6	997	5,982	2.11%
Petrópolis	9	1,731	15,579	2.60%
Muriaé	8	1,559	12,472	1.09%
Governador Valadares	8	2,022	16,176	0.59%
Juiz de Fora	49	7,719	378,231	0.95%
Sum	80	14,028	428,440	1.01%

In the experiments regarding PISM 2019 and PISM 2020, the number of available locations was exactly the same of the used locations, once the data were about exams already applied. For the PISM 2021, as we started our implementation in the planning phase of the exam, there were plenty of locations available, in a number considerably higher than the number of locations needed. Thus, the definition of which exam locations will be used will be up to the optimization model executed for each weekend of the exam. Figure 4.4 illustrates the distribution of available locations in the city of Juiz de Fora of the current PISM edition, the dots are the exam sites and the square represents the UFJF campus and its 21 locations.
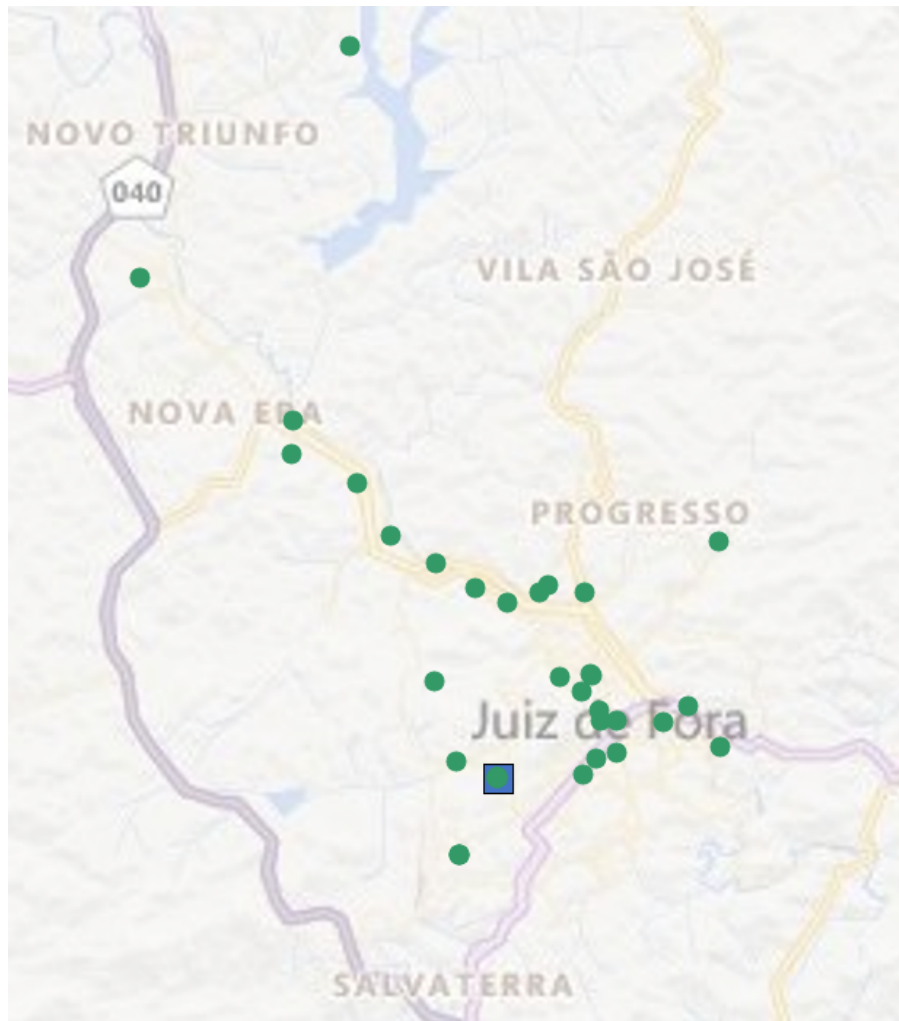
Figure 4.4: Distribution of available exam locations of PISM 2021 for Juiz de Fora

In sequence, we executed the pre-processing using our Python algorithm - which allows this stage to be performed very quickly and directly, despite the large amount of data involved. In Table 4.13, the numbers of valid and disregarded candidates are presented - those who reside more than the cutoff parameter of 50 km. We can observe that the ratio of valid candidates for each city is close to the ratio of the 2019 and 2020 editions, which lead us to assume that for the current exam the cutoff parameter also includes most of the resident candidates of the cities, even for Muriaé, which has a much smaller proportion of valid candidates, either the exam edition, because this city actually has fewer resident candidates when compared to others - in PISM 2020, for example, only 16.6% of the candidates of Muriaé lived in the city.

We then executed the optimization again with JuMP and GLPK solver in the same machine with an Intel Core i5 6400 2.7GHz with 16GB of RAM. For modules 2 and 3, that posses more candidates, the running times were larger than to the respective exam city of module 1. The results of Volta Redonda

Table 4.13: Disregarded and Valid Candidates of PISM 2021

Modules 2 and 3	Disregarded Candidates	Valid Candidates	% Valid Candidates
Volta Redonda	591	568	49.01%
Petrópolis	979	802	45.03%
Muriaé	1,901	400	17.38%
Governador Valadares	1,797	782	30.32%
Juiz de Fora	6,205	5,538	47.16%
Sum	11,473	8,090	41.35%
Module 1	Disregarded Candidates	Valid Candidates	% Valid Candidates
Volta Redonda	473	524	52.56%
Petrópolis	1,105	626	36.16%
Muriaé	1,252	307	19.69%
Governador Valadares	1,455	567	28.04%
Juiz de Fora	4,594	3,125	40.48%
Sum	8,879	5,149	36.71%

were obtained in 13 seconds for all modules; for Petrópolis the times were 15 seconds for each execution; Muriaé presented a solution time of 16 and 86 seconds; Governador Valadares ran in 20 seconds and 1 minute and a half; and for Juiz de Fora the solutions again were performed with a running time limited to 5 hours, obtaining sub-optimal solutions with a gap of 4.2% for tests of module 1 and a gap of 7.9% for tests of modules 2 and 3. The total and average optimized travel distance follows in Table 4.14.

Table 4.14: Optimized travel distances of PISM 2021

	Total Displacement(m)		Average Displacement(m)	
	Modules 2&3	Module 1	Modules 2&3	Module 1
Volta Redonda	10,283,515	8,919,017	18,105	17,021
Petrópolis	9,158,050	8,187,346	11,419	13,079
Muriaé	5,187,476	4,469,737	12,969	14,559
Governador Valadares	3,500,626	2,423,485	4,477	4,274
Juiz de Fora	44,854,799	22,150,650	8,099	7,088

In the standard tests, Juiz de Fora had 49 available locations, with 21 of these being places located on its campus. For the experiments of modules 2 and 3 were used 36 locations, with 12 of them belonging to the campus; and for

module 1 it was necessary only 18 locations, and just 5 were located on campus. This is a great paradigm shift, once historically the application of PISM always used every available location on the UFJF campus. Thus, as our objective is to provide efficient support to decision making and knowing the impacts of such change, we performed further tests for Juiz de Fora adding in the model the obligation of using all the locations in UFJF's campus, this was simply made resetting the opening costs of UFJF's locations, what makes the optimization to fill these locations preferably.

Comparing the two options for Juiz de Fora - the standard optimization and the one with the obligation to use all campus locations -, it is notorious that the first option brings better results. The experiments with locations restrictions present an increase in the number of used locations and average travel distance: for modules 2 and 3 are used four more exam locations and for module 1 eight more locations are used, which increases the logistic complexity of the process and its costs; the travel distances of candidates for modules 2 and 3 are 9.23% larger and for module 1 the increase is 15.09%, what proves that these restrictions make the model less efficient in what concerns the main objective of reducing the candidates travel distance. These results were presented to the university's administration and, despite the historical modification, they chose to apply the standard test that proved to be more effective. Its data are the ones presented in this Section.

As for PISM 2021, the allocation used was the one proposed here, we have no benchmark to compare the results in order to measure its effectiveness. Thereby, as seen in Table 4.15, we analyzed the average travel distance for each city for the current edition and the other two editions that we have data. Despite the significant alteration of data from one exam edition to another, since the candidates are different, as well as the exam sites, the average travel distance provides us a reasonable panorama of the optimization benefits.

We can see that the average travel distance for the 2021 edition of PISM is very similar to the average travel distance of previous editions, which assures us about the benefits of the developed methodology. In fact, for the main exam city - Juiz de Fora, which posses about 60% of all candidates -, there was a considerable reduction in the average travel distance of 14.37% when comparing with PISM 2020. It is important to note that the average travel distance for PISM 2021 was calculated considering the global travel distance divided by the total of valid candidates for each city, since this exam was applied in two phases.

Another possible analysis on PISM 2021 is to compare the ratio between candidates within the distance of 2 km - as explained earlier, those who may

Table 4.15: Average optimized travel distance per exam edition

	PISM 2019	PISM 2020	PISM 2021
Volta Redonda	18,525	17,719	17,585
Petrópolis	-	11,149	12,147
Muriaé	12,355	11,309	13,659
Governador Valadares	5,749	4,453	4,391
Juiz de Fora	7,534	9,033	7,735

go to exam locations on foot - and the number of valid and total of candidates from all editions. Thus, we have one more parallel observation for the current edition and the selection process previously applied. As seen in Table 4.16, PISM 2021 presented a proportion of valid candidates within 2 km of distance very similar to previous editions, with small variations. Even when we analyze the values of each city, we observe that the percentage of the total of candidates who fit our distance parameter is close, showing that our model was as efficient as in the previous exam editions that we tested.

As a whole, we can see that the optimization proved to be very effective, allowing more than a quarter of the valid candidates to be able to go to the exam locations on foot, this figure being around 40% for some of the cities. If we analyze the proportion of the total candidates shown in Table 4.17, this number is higher than 10% of all candidates.

Table 4.16: Ratio of valid candidates able to travel on foot

	PISM 2019		PISM 2020		PISM 2021	
	Cand.	% Valid	Cand.	% Valid	Cand.	% Valid
Volta Redonda	89	8.15%	121	9.70%	146	13.37%
Petrópolis	-	-	299	17.24%	230	16.11%
Muriaé	285	41.13%	332	35.97%	282	39.89%
Gov. Valadares	325	28.58%	661	39.18%	587	43.51%
Juiz de Fora	3,666	34.41%	3,421	31.16%	2,456	28.36%
Sum	4,365	32.15%	4,834	29.17%	3,701	27.96%

To better illustrate the results obtained and improve their analysis, we present Figure 4.5, which shows the proportion of the total candidates able to travel on foot through each selection process - and the original data in relation to the optimized ones -, separated by city, with also the totalization for each exam. Thus, it is easy to see the positive impact of the optimization with a significant increase in the proportion of candidates able to travel on foot for

Table 4.17: Ratio of total candidates able to travel on foot

	PISM 2019		PISM 2020		PISM 2021	
	Cand.	% Total	Cand.	% Total	Cand.	% Total
Volta Redonda	89	3.73%	121	4.64%	146	6.77%
Petrópolis	-	-	299	8.12%	230	6.55%
Muriaé	285	7.88%	332	7.49%	282	7.31%
Gov. Valadares	325	8.63%	661	12.98%	587	12.76%
Juiz de Fora	3,666	14.91%	3,421	14.52%	2,456	12.62%
Sum	4,365	12.71%	4,834	12.28%	3,701	11.02%

every exam city, mainly in Juiz de Fora, which had the best results and has the most candidates.

Comparing the improvements achieved from the use of the mathematical model, there were in original data from PISM 2020 a total of 1,082 candidates allocated within 2 km of distance, which means that 2.75% of all candidates were able to go to exam locations on foot, while in PISM 2021 - where the optimization was truly implemented in a real case - there were 3,701 candidates with that characteristic, which corresponds to 11.02%. This way, we can affirm that with the utilization of the model developed here the proportion of candidates able to go to exam locations on foot in the current exam is four times greater than that proportion in PISM 2020.

Figure 4.6 presents, for each edition of the selection process, a radar chart of occupation of the exam locations: the inner areas refer to the percentage of the exams location with the lowest occupation for each city and the outer areas refer to the percentage of average occupation for an entire city. One can notice that there is a significant discrepancy between the minimum and average percentage occupation throughout the editions of PISM and cities, such as edition 2019 in which the difference between the minimum and average occupation is very low, unlike the edition 2021 where these differences are more accentuated, mainly in Petrópolis and Muriaé. However, the analysis of the outer areas in the figure reveals that the average percentage of occupation is very high, regardless of city or selection process, not being lower than 92%.

These results show the overall efficiency of the model considering the aim of allowing the use of the least amount of exam locations possible - and, consequently, their use as close as possible to full capacity - in order to reduce the logistical complexity and general costs. This way, we are able to verify that this objective has been fulfilled: for all editions of PISM, most of the exam locations used are completely occupied and those with vacancies present great
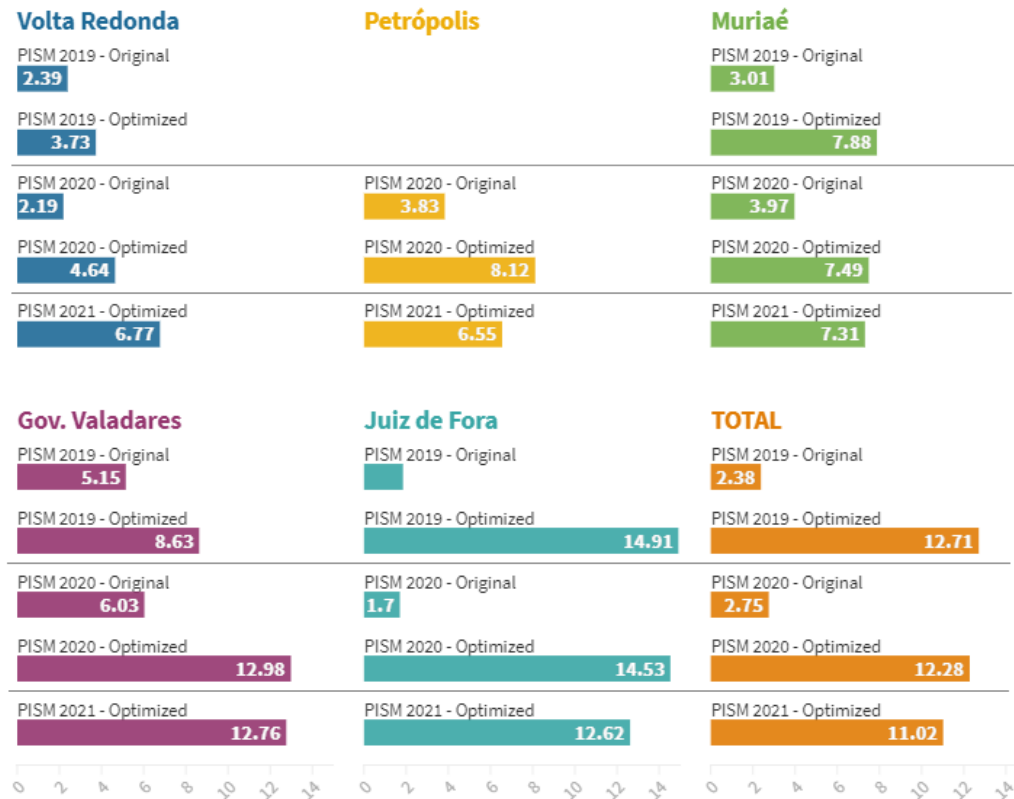
Figure 4.5: Percentage of candidates able to travel on foot by exam city

values of occupations, which makes the average occupation very high, even if there are places in a city with a significant number of vacancies.

For PISM 2021, we again performed a survey with the candidates in order to measure their acceptance of the optimized allocation. We highlight that UFJF published a public call at the beginning of February, splitting the applications of the exam again due to the coronavirus pandemic. Now, tests of module three - which would be applied together with module two - will be applied alone in the last weekend of February, module two will be applied in the last weekend of March and module one will be applied in the last weekend of April. This way, the Proof of Registration - where the survey is available - was destined only for candidates of module three, with the rest of the candidates being able to access the survey just after the presentation of this thesis. From a total of 8,014 candidates of this module, 1,636 candidates participated, which represents 20.4% of the total.

The comparison between the responses of this survey and the survey applied for the candidates of PISM 2020 is then presented. From Figure 4.7, we can note a relevant increase in the number of candidates allocated very close or nearby of their addresses: from 28% in 2020 to 45% in 2021. We see in Figure 4.8 that 8% more candidates considered to be pleased with their allocation

Figure 4.6: Percentage of minimum and average occupation of exam locations

in the current exam. In Figure 4.9 it is possible to note a reduction in the perception of the candidates about the existence of exam locations closer to their residence different from the one where they were allocated. Finally, in Figure 4.10, we can see a significant increase in the percentage of candidates who will travel on foot - as indicated by the numeric results of the optimization - and, as a consequence, a similar reduction in the number of candidates who will use public transportation, which can help to avoid traffic jams on exam days.

How far away is your exam location from your residence?

PISM 2020 PISM 2021

● Very close (I can arrive in 15min) ● Nearby (I can arrive in 30 minutes) ● Far (I can arrive in 45min) ● Too far (I can arrive in more than 45 minutes) ● I don't know
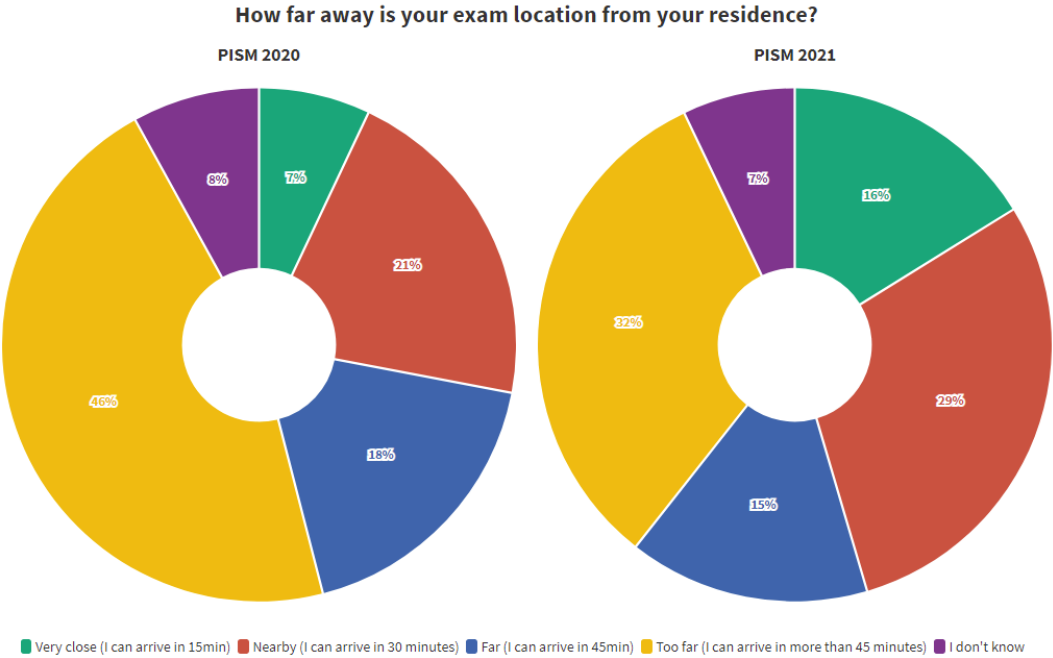
Figure 4.7: Question 1 and responses - comparison between PISM 2020 and PISM 2021

How satisfied are you with the location you were allocated?
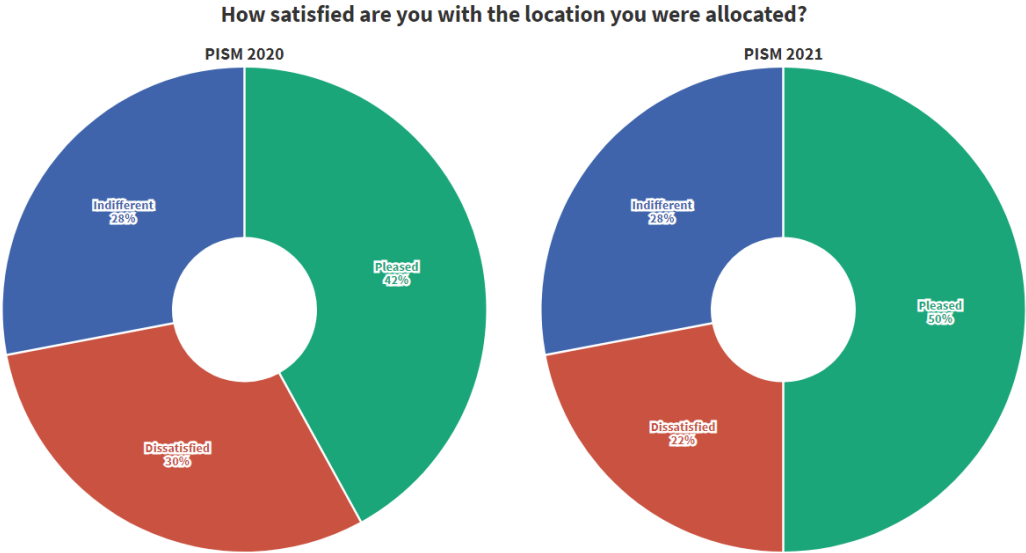
PISM 2020 PISM 2021

Figure 4.8: Question 2 and responses - comparison between PISM 2020 and PISM 2021

Is there any location closer to your residence than the one you were allocated?

PISM 2020   PISM 2021

I don't know 32%
Yes 42%
No 26%

I don't know 42%
Yes 36%
No 22%

Figure 4.9: Question 3 and responses - comparison between PISM 2020 and PISM 2021

How will you travel to the exam location?

PISM 2020   PISM 2021

On foot 8%
Public transportation 29%
Taxi / App 12%
Own vehicle 51%

On foot 15%
Public transportation 23%
Taxi / App 13%
Own vehicle 49%

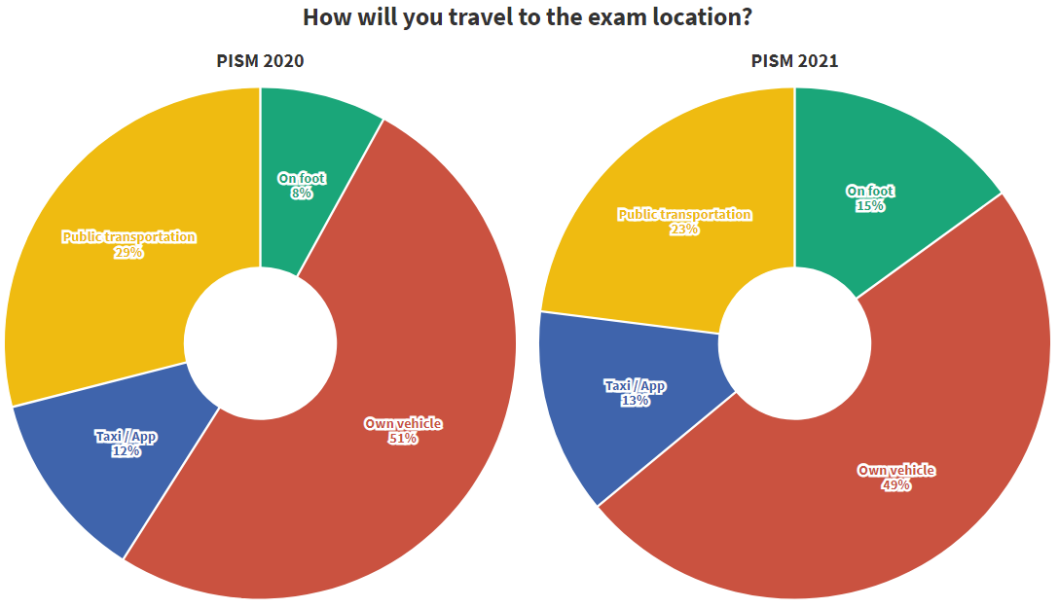Figure 4.10: Question 4 and responses - comparison between PISM 2020 and PISM 2021

5
Conclusions and Future Work

This work was elaborated with the objective of developing, testing and verifying the applicability and effectiveness of a model to improve the allocation of candidates in exam locations of university admission exams, reducing the distance traveled by the candidates. Throughout the study it was presented all the data and results gathered across its progress: once an original mathematical model based on the literature review was established, it was tested on the real data of the previous exam editions of a Brazilian University (Federal University of Juiz de Fora - UFJF) in order to measure its efficiency and, from the positive outcomes obtained through extensive experimental analyses, our work was applied in most recent exam edition.

As the results show, in all performed experiments the main objective of reducing the total travel distance of candidates was accomplished, with reduction for a city reaching significantly values as high as 33% and 40%, and more than 50% when we analyze specific range distances. The number of candidates able to go to exam locations on foot more than doubled for most cities and had an expressive increase of seven times for Juiz de Fora, with that amount meaning more than 11% of all candidates and more than 27% considering only valid candidates. It is quite important to highlight that all these successful outcomes were achieved using only free or open-source tools available for everyone, producing no significant cost to its application related to equipment or service subscription.

Another positive property of this study was the possibility to perform the experiments through different editions of the selection process and in a range of exam cities with varied traits: locations in different states and realities; a medium town like Muriaé with about 100,000 inhabitants and a large city with almost 600,000 inhabitants like Juiz de Fora; locations with nearly 24,000 candidates and others that did not reach 3,000; cities with almost half of its candidates who live in its limits and others where no more than 1/6 of the candidates are residents. Such characteristics allowed us to exhaustively test our model in problems with very different features. Yet, it provided significant reductions in travel distances, making the mathematical model and its results even more meaningful.

Thus, it is possible to conclude the significant potential of the work developed here in the improvement of university admission exams. The reduction of the distance traveled by the candidates facilitates their travel to the exam locations and reduces costs for them - potentially even decreasing the number of absences - and presents a substantial reduction in traffic on exam days, which also has a positive impact on the environment.

An opportunity to develop future work is seen in the exam cities' definition. These cities, in addition to those with UFJF campuses - Juiz de Fora and Governador Valadares -, are chosen by the university at the request of its resident candidates. However, they are not necessarily the cities with the most residents. In fact, we identified that Muriaé varies from the sixth to the eighth position in the number of resident candidates, while Volta Redonda does not even reach the top ten places. Thus, the analysis of other factors - such as the concentration of candidates through the regions involved - combined with the consideration of cities with more resident candidates, can improve the definition of exam cities, increasing the number of valid candidates and, as a consequence, the number of positively impacted candidates.

In this work, it was established a cutoff parameter of 50 km, which above that the candidates have their distances reset and are disregarded in the mathematical model. Therefore, a relevant possibility to future research is to carry out experiments with different values of the cutoff parameter and analyze the impact of this change in the number of valid candidates and the model's efficiency as a whole.

Finally, as there is room for improvement in the definition of exam cities, it is also possible to move forward in the process of choosing exam locations within a city. Future studies may use a clustering method to suggest the most suitable exam sites to the administration of the selection process in order to allow the utilization of an optimal distribution of those places for each city, which can significantly increase the number of candidates able to go to their test on foot and reduces the impact of the exam application in traffic.

Bibliography

Adeleke, O. J. and Olukanni, D. O. (2020). Facility location problems: Models, techniques, and applications in waste management. *Recycling*, 5.

Afrashteh, E., Alizadeh, B., and Baroughi, F. (2018). Optimal algorithms for integer inverse undesirable p-median location problems on weighted extended star networks. *Journal of the Operations Research Society of China*.

Antunes, A. and Peeters, D. (2000). A dynamic optimization model for school network planning. *Socio-Economic Planning Sciences*, 34:101–120.

Avella, P., Boccia, M., Mattia, S., and Rossi, F. (2020). Weak flow cover inequalities for the capacitated facility location problem. *European Journal of Operational Research*.

Batista, I. C. B. (2010). A module for solving point location problems in a gis environment. Working Paper from Universidade Tecnica de Lisboa.

Bing Maps (2020). *Bing Maps Documentation*. Accessed July 09, 2020.

Brasil (1996). *Lei 9.394/1996 - Lei de Diretrizes e Bases da Educação Nacional*. Accessed October 28, 2020, in Portuguese.

Bruno, G., Esposito, E., Genovese, A., and Piccolo, C. (2016). Institutions and facility mergers in the italian education system: Models and case studies. *Socio-Economic Planning Sciences*, 53:23–32.

Burkey, M., Bhadury, J., and Eiselt, H. (2012). A location-based comparison of health care services in four u.s. states with efficiency and equity. *Socio-Economic Planning Sciences*, 46:157–163.

Cabrera, A. F. and La Nasa, S. M. (2000). Understanding the college-choice process. *New directions for institutional research*, 2000(107):5–22.

China Daily (2020). *10.71 million students to sit gaokao*. Accessed October 21, 2020.

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

Clausen, J. (1999). Branch and bound algorithms - principles and examples. Department of Computer Science, University of Copenhagen.

COIN-OR Foundation (2020). *COIN-OR Foundation - purpose and structure*. Accessed September 17, 2020.

College Board (2020). *Nearly 2.2 Million Students in the Class of 2020 Took The SAT At Least Once*. Accessed November 13, 2020.

Cotes, N. and Cantillo, V. (2019). Including deprivation costs in facility location models for humanitarian relief logistics. *Socio-Economic Planning Sciences*, 65:89–100.

da Veiga, C. L. (2020). O atendimento especial do programa de ingresso seletivo misto (pism) da universidade federal de juiz de fora (ufjf): histórico, desafios e possibilidades. Master's thesis, Universidade Federal de Juiz de Fora.

Dasgupta, S., Papadimitriou, C. H., and Vazirani, U. V. (2006). *Algorithms.* New York: McGraw-Hill Higher Education.

Daskin, M. S. (1995). *Network and discrete location: models, algorithms, and applications.* John Wiley and Sons, Inc.

de Andrade, R. Q. and Silva, A. L. (2011). Metodologia baseada no algoritmo de branch-and-bound para localização de sedes de vestibulares da UFOP. In *Simpósio Brasileiro de Pesquisa Operacional*, pages 2876–2885. in Portuguese.

de Werra, D. (1985). An introduction to timetabling. *European journal of operational research*, 19(2):151–162.

Drezner, Z. (1995). *Facility Location: A Survey of Applications and Methods.* Springer-Verlag New York.

Ebery, J., Krishnamoorthy, M., Ernst, A., and Boland, N. (2000). The capacitated multiple allocation hub location problem: Formulations and algorithms. *European Journal of Operational Research*, 120:614–631.

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., and Surprenant-Legault, J. (2013). New evidence on walking distances to transit stops: identifying redundancies and gaps using variable service areas. *Transportation*, 41:193–210.

Farahani, R. Z., Asgari, N., Heidari, N., Hosseininia, M., and Goh, M. (2012). Covering problems in facility location: A review. *Computers and Industrial Engineering*, 62:368–407.

Farahani, R. Z. and Hekmatfar, M. (2009). *Facility location: concepts, models, algorithms and case studies.* Contributions to Management Science. Physica-Verlag.

Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15.

GNU Project (2020). *GLPK documentation.* Accessed September 1, 2020.

Google (2020). *Google Maps Usage and Billing.* Accessed July 20, 2020.

Hafalir, I. E., Hakimov, R., Kübler, D., and Kurino, M. (2018). College admissions with entrance exams: Centralized versus decentralized. *Journal of Economic Theory*, 176:886–934.

Haginoya, S., Hanayama, A., and Koike, T. (2020). Linkage analysis using geographical proximity: a test of the efficacy of distance measures. *Journal of Criminological Research, Policy and Practice*.

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

Heidelberg University (2021). *Heidelberg University in international comparison.* Accessed January 27, 2021.

Hu, Y., Wu, H., Wang, R., and St-Onge, B. (2006). Locating targets under perspective with genetic algorithm and tabu search. *American Society for Photogrammetry and Remote Sensing*, 1:563–574.

Huang, R., Kim, S., and Menezes, M. B. (2010). Facility location for large-scale emergencies. *Annals of Operations Research*, 181:271–286.

Insider (2018). *What standardized tests look like in 10 places around the world.* Accessed January 18, 2021.

Instituto Brasileiro de Geografia e Estatística (2010). *Censo Demográfico.* Accessed February 26, 2020, in Portuguese.

Instituto de Políticas de Transporte e Desenvolvimento (2015). *ITDP lança o indicador PNT (People Near Transit).* Accessed June 09, 2020, in Portuguese.

Instituto Nacional de Estudos e Pesquisas Educacionais (2018). *Censo da Educação Superior.* Accessed February 21, 2020, in Portuguese.

Jokila, S., Haltia, N., and Kosunen, S. (2020). Market-making practices of private tutoring in finland: Commercialization of exam preparation for admission to higher education. *ECNU Review of Education*, pages 1–25.

JuMP Developer (2020). *JuMP Installation Guide.* Accessed March 17, 2020.

Kariv, O. and Hakim, S. L. (1979). An algorithmic approach to network location problems. ii: The p-medians. *SIAM Journal on Applied Mathematics*, 37:513–538.

Klose, A. and Drexl, A. (2005). Facility location models for distribution system design. *European Journal of Operational Research*, 162(1):4–29.

Kripka, R. M. L., Kripka, M., and da Silva, M. C. (2011). Formulação para o problema de alocação de salas de aula com minimização de deslocamentos. In *Simpósio Brasileiro de Pesquisa Operacional*, pages 1941–1951. in Portuguese.

Kwon, C. (2019). *Julia Programming for Operations Research.*

Lin, Y.-H., Batta, R., Rogerson, P. A., Blatt, A., and Flanigan, M. (2012). Location of temporary depots to facilitate relief operations after an earthquake. *Socio-Economic Planning Sciences*, 46:112–123.

Matisziw, T. C. and Murray, A. T. (2009). Siting a facility in continuous space to maximize coverage of a region. *Socio-Economic Planning Sciences*, 43:131–139.

Melo, M., Nickel, S., and da Gama, F. S. (2009). Facility location and supply chain management – a review. *European Journal of Operational Research*, 196:401–412.

Microsoft (2020). *Bing Maps Licensing.* Accessed July 23, 2020.

Mousavi, S. M., Gitinavard, H., Vahdani, B., and Foroozesh, N. (2019). Hierarchical group compromise ranking methodology based on euclidean–hausdorff distance measure under uncertainty: An application to facility location selection problem. *Journal of Optimization in Industrial Engineering*, 12:93–105.

Murali, P., Ordóñez, F., and Dessouky, M. M. (2012). Facility location under demand uncertainty: Response to a large-scale bio-terror attack. *Socio-Economic Planning Sciences*, 46:78–87.

National Geographic (2017). *GIS (Geographic Information System)*. Accessed July 13, 2020.

Nguyen, K. T. and Sepasian, A. R. (2016). The inverse 1-center problem on trees with variable edge lengths under chebyshev norm and hamming distance. *Journal of Combinatorial Optimization*, 32:872–884.

Ogryczak, W. (1997). On the lexicographic minimax approach to location problems. *European Journal of Operational Research*, 100:566–585.

Open Route Service (2020a). *Open Route Service Data*. Accessed July 17, 2020.

Open Route Service (2020b). *Open Route Service Documentation*. Accessed August 03, 2020.

Owen, S. H. and Daskin, M. S. (1998). Strategic facility location: A review. *European Journal of Operational Research*, 111:423–447.

Panichelli, L. and Gnansounou, E. (2008). Gis-based approach for defining bioenergy facilities location: A case study in northern spain based on marginal delivery costs and resources competition between facilities. *Biomass and Bioenergy*, 32:289–300.

Pires, M. D. J. (2019). Gaokao: far more than an exam. *Diadorim*, 21:168–185.

Ralphs, T., Ladányi, L., and Saltzman, M. (2003). Parallel branch, cut, and price for large-scale discrete optimization. *Mathematical Programming*, 98:253–280.

Shen, G. (2005). Location of manufactured housing and its accessibility to community services: a gis-assisted spatial analysis. *Socio-Economic Planning Sciences*, 39:25–41.

Sridharan, R. (1995). The capacitated plant location problem. *European Journal of Operational Research*, 87(2):203–213.

Tomlinson, R. F. (1967). An introduction to the geo-information system of the canada land inventory. *Department of Forestry and Rural Development*, pages 1–25.

Turkoglu, D. C. and Genevois, M. E. (2020). A comparative survey of service facility location problems. *Annals of Operations Research*, 292:399–468.

Ulukan, Z. and Demircioglu, E. (2015). A survey of discrete facility location problems. *Int. J. Soc. Behav. Educ. Econ. Bus. Ind. Eng*, 9(7):2487–2492.

Universidade Federal de Juiz de Fora (1999). *RESOLUÇÃO Nº 18/99 - Institui o Programa de Ingresso Seletivo Misto da Universidade Federal de Juiz de Fora.* Accessed January 23, 2021, in Portuguese.

Universidade Federal de Juiz de Fora (2020a). *Anos 2000: como surgiram Pism e cotas e como acabou o vestibular tradicional.* Accessed January 27, 2021, in Portuguese.

Universidade Federal de Juiz de Fora (2020b). *Pism 2021 conta com candidatos de todas as regiões do país e um de fora do Brasil.* Accessed November 28, 2020, in Portuguese.

Vieira, I. F. G., de Barros, M. F., and Cormack, A. (2019). Hierarchical facility location model for allocating cancer treatment units in interior of rio de janeiro. *Production*, 29.

Weber, A. (1929). *Theory of the location of industries.* University of Chicago Press.

Yu, V. F., Lin, S.-W., Lee, W., and Ting, C.-J. (2010). A simulated annealing heuristic for the capacitated location routing problem. *Computers and Industrial Engineering*, 58:288–299.