RE-ENGINEERING A DIVERSE SERVICE SYSTEM: MULTI-CRITERIA RESOURCE ASSIGNMENT

By

Miguel Ángel Ruiz Hernández

A thesis submitted in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE IN
INDUSTRIAL ENGINEERING

University of Puerto Rico
Mayagüez Campus
2017

Approved by:

Betzabé Rodríguez Álamo, Ph.D. President, Graduate Committee	Date
Saylisse Dávila Padilla, Ph.D. Member, Graduate Committee	Date
Guillermo Ortiz Colón, Ph.D. Member, Graduate Committee	Date
Viviana I. Cesaní Vázquez, Ph.D. Member, Graduate Committee Department Head	Date
Edgardo Lorenzo González, Ph.D. Graduate School Representative	Date

Abstract

Balancing supply and demand in service organizations is a complex problem that should be approached using mathematical modeling. This work developed a resource assignment formulation that deals with the mismatch between personnel profile and tasks required skills, the effect of travel on resource's capacity, and the unknown relationship between tasks and workload. The formulation proved to be computationally unsolvable with regular computing capabilities; hence, an alternative formulation was developed. The University of Puerto Rico Agricultural Extension Service was used as test bed. Results showed that the alternative formulation performed very well, average *utility* results were less than 1% lower than the global optimum across 400 replicas, and *utility*/*size* for the full-size problem was in the 79th percentile across the 400-replica's optimal solution. This work introduces JAM, a contribution to clustering analysis, where cuts at different levels are used to extract clusters of different sizes and densities from Ward's dendrograms.

Resumen

El balanceo de la oferta y la demanda en las organizaciones de servicios es un problema complejo que se debe resolver usando modelos matemáticos. Este trabajo desarrolló una formulación de asignación de recursos que funciona bajo desajuste entre el perfil del personal y las habilidades requeridas para realizar las tareas, considerando el efecto del viaje en la capacidad de los recursos y el desconocimiento de la relación entre las tareas a realizar y la carga de trabajo. La formulación mostró ser demasiado pesada computacionalmente para unos recursos computacionales estándar. Debido a esto, se desarrolló una formulación alternativa. El banco de pruebas de este estudio fue el Servicio de Extensión Agrícola de la Universidad de Puerto Rico. Los resultados mostraron que la formulación alternativa desempeó muy bien, la diferencia en *utilidad* promedio a lo largo de 400 réplicas entre la formulación alternativa y el óptimo global estuvo por debajo del 1%, y la *utilidad*/*tamaño* para el problema de tamaño completo se situó en la percentila 79 de las soluciones óptimas de las 400 réplicas. Este trabajo presenta JAM, una contribución al análisis de clustering, un algoritmo capaz de mejorar la identificación de clusters a partir de dendrogramas generados con el método de Ward, cuando los grupos tienen tamaños y densidades diferentes.

Acknowledgments

This Thesis is dedicated to the ones that have gone. Juliana, Isidoro and Cristina. Juliana, your joy will always be my path. Isidoro, your impeccability will always inspire me. Cristina, you faced the biggest fear, thank you for your love and thank you for waiting.

When I think about the people to whom it is necessary to give thanks, two people instantly come to my mind, Betzabé Rodríguez and Andrés Calle. I learned a lot from both of you, fortunately, you already know this. Betzabé, it has been amazing being around you. Thank you for your honesty, for your good intentions, for keeping your word, for all the respect you always show, for your advice, for listening, for saying thank you, for understanding me, for your commitment, for your happiness... You are amazing; I can not imagine a better person to choose as my advisor. I also thank to my buddy, Andrés. We shared a great time together during this masters degree, studying, fixing cars, hanging out and learning together in this entrepreneurial dream we are living. Thanks for your help whenever I needed it.

I want to thank my Committee. Saylisse Dávila, your commitment is amazing, thanks for your time and your advice, I really appreciate it. Guillermo Ortíz, thank you for your commitment and all your help through this process. Viviana I. Cesaní, thank you for opening spaces in your tight schedule and for your punctual and valuable advice. I want to thank all my partners during these years in the industrial engineering department: Maximino, Dhania, Nicole, Joey, Elizabeth, ”los guarracos” (Heizel, Isis, Jeff, César Salazar),

César Dávila, Karen, Ismael, Yaritza, Samuel and all the others. Thank you for the laughs and the help.

I want to thank the undergraduate students that collaborated in this research: Jennifer González, Enery Lorenzo, Collin Stacy and Kohl Anderson. I also want to thank Lourdes Medina for giving me the chance in my first semester, i learned a lot from that experience.

Thanks to all my professors during this masters degree. David, Mayra, Betzabé, Sonia, Nazario, Saylisse, Mauricio and Juan. One of the best things during this masters degree was the overall quality of professors. Thanks to my yoga master Andrés. Andrés you showed me the two characteristics of a great teacher, being passionate about teaching and like the topic you are teaching. Teaching is the ultimate learning.

I decided to came back to university and study my masters degree because I wanted to learn. I am passionate about learning, I learned a lot from book and classes. It turns out that I also learned a lot from the people surrounding me, it has been amazing. I want to thank my brother Javier, my parents, my uncles (especially Jose Ángel and Rafa H.), my friends (Fer, Blanch, Eloy, Sambri, Eugenio, Alberto, Cremades, Willy, Juan Carlos) for always being there. And finally, thanks to my friends, Tito, you made this possible, Pablo, we initiated this amazing trip, and Marcos, sharing time and projects with you during the lasts 6 months has been a great experience.

Finally, I want to thank the University of Puerto Rico for this opportunity and the economic support during these years.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The assignment of resources to tasks or needs is a common challenge for any complex service system. The challenge soars in scenarios where a mismatch between resources capabilities and task requirements occurs. If a traditional assignment model is used in these scenarios, it will often result in a service organization assigning employees (or other resources) to tasks they will perform poorly. For example, a service organization may want to assign a resource only if the resource is going to perform accordingly to their standards, otherwise this assignment will harm the companys reputation. Additionally, the organization may want to complete a task only if completed according to those same standards. If a traditional assignment model is used on this scenario, the solution may suggest assigning a resource to a task that will be detrimental to the image of the organization. The assignment model proposed in this work considers the resource-task fit, and may result in a solution that suggests not fulfilling all tasks and/or not assigning a resource at all.

This research tailors the assignment model to account for common challenges faced by service organizations. The common challenges faced do not only include the detrimental benefit of fulfilling tasks with unfit resources, but also include unknown relation between demand (or tasks) and workload, myriad of tasks defined by their required soft and tech-

nical skills, and a diverse resource or personnel profile. To the best of our knowledge, no solution has been proposed in literature for assignment problems taking into account all of these challenges. Solutions for assignment problems with combinations of personnel profiles and tasks diversity has been addressed before by: Korkmaz et al. [2008], Peters and Zelewski [2007], Wongwien and Nanthavanij [2013]. The closest work to account for this is Caron et al. [1999] that shows that a particular formulation of the assignment problem can lead to both idle workers and unassigned tasks; however, no work was found where a real application of the assignment problem with a solution proposing the possibility of both idle workers and unassigned tasks simultaneously.

The methodology and model in this study uses as a test bed the assignment of University of Puerto Rico Agricultural Extension Service's (AES) agents (i.e. resources) to tasks. AES is a service organization dedicated to disseminate scientific knowledge pertaining agriculture, and reaches out to farmers and disadvantaged communities in Puerto Rico (PR) to educate them in agricultural activities. The AES service paramount goal is to advance agricultural activity in Puerto Rico (PR), which is currently food insecure according to the report: Encuestas de la Oficina de Estadísticas Agrícolas, (Junta de Planificación de Puerto Rico, 2012), indicating that PR produces only 17.65% (in weight) of its consumed food products. AES contribution to agricultural activity in PR depends immensely in the quality of the assignment of agents to tasks. AES quality of service is grounded in resources capabilities on performing their assigned tasks; hence, it is imperative to find an assignment model that fits this scenario.

This work is divided into two phases primarily based on project execution and knowledge domains. The first phase includes the estimation of the tasks needed to be performed by the agents; this phase is analogous to a study of the demand. The methodology employed in this phase is primarily grounded in statistical analyses and techniques. Chapter

2 and Chapter 3 in this document pertain to this phase of the study. Although we acknowledge that there is contribution to the literature in this phase, the contribution we highlight in this document is in the second phase. The second phase includes the creation and test of the assignment model. The assignment model uses primarily optimization modeling techniques and it is described in Chapter 4. The second phase uses the results of previous chapters as inputs.

The remainder of the document is organized as follows. Chapter 2 presents a novel study. This study includes the development of JAM (Jump Analysis Method) a new analysis technique for outputs generated with Ward's hierarchical clustering method. Chapter 3 shows the methods and results used for the AES study of demand. This first phase of the study uses JAM (method described in Chapter 2) as part of the methodology. Chapter 4 shows the methods and results used for the assignment of AES agents to tasks. This second phase of the study has linear programming as main body of knowledge. Chapter 5 summarizes the concluding remarks of this work. The Appendix adds documentation relevant to support the work presented.

Chapter 2

Jump Analysis Method (JAM)

Data analysis allows us to learn about systems and make knowledge-based decisions. However, parameter evaluation for interpretation and data analysis often relies on multiple modeling decisions, making it a daunting task. Hierarchical clustering is an unsupervised learning technique that produces a graphical representation known as dendrogram that is highly subjective as its visual inspection can lead to multiple interpretations. In this chapter, Jump Analysis Method (**JAM**), a novel technique for analyzing dendrograms generated with Ward's hierarchical clustering method is presented. This technique requires simple parameter tuning and only basic statistical knowledge in order to be implemented. Additionally, this approach allows users to cut a dendrogram at different levels unlike traditional horizontal cuts (i.e. "stopping rules"). This methodology was tested with success for different scenarios: simulated by the research team, simulated by others and real data sets. Results show this method give some advantages compared to traditional dendrogram analysis and k-means. This will lead to better understanding of clustering outputs, especially where clusters vary in density size and shape, or few data is available.

2.1 Introduction

Hierarchical clustering is a popular technique in data mining. The applications of clustering technique range anywhere from grouping of customers and tasks [Ho et al., 2012, Shih and Liu, 2003, Elango et al., 2011] to storage allocation [Chuang et al., 2012], location routing problem [Barreto et al., 2007] and document retrieval [Willett, 1988].

Ward's hierarchical clustering method generates a hierarchical representation of the data set, known as *dendrogram*, but it does not divide the data into groups. Various methods for selecting groups given a dendrogram have been published; however, some of them suffer from subjectivity and others require determining numerous parameters. In this work, we present a method for group selection given a dendrogram generated with Ward's linkage [Jr., 1963] that requires a small number of parameters and only basic data mining knowledge in order to be implemented. Additionally, this method allows to cut the dendrogram at different levels simultaneously.

The most commonly used technique to find clusters from a dendrogram is by using the "horizontal cut method by visual inspection". Horizontal cutting method by visual inspection is a very simplistic method that will only require the user to choose the desired number of clusters, and, then, select the desired height for a horizontal cut. The number of clusters that fall under the horizontal cut are the desired clusters. This method, despite its simplicity, suffers from subjectivity and is highly dependent on the analyst's perception.The second most commonly used method for defining clusters is to evaluate different horizontal cuts and choose the best cut based on performance measures, which is the approach used by "stopping rules" [Salvador and Chan, 2004, Gong and Oard, 2009].

Horizontal cut is a good method for simple cases (spherical groups with similar size and density), but might fail when trying to select clusters out of a complex dendrogram

structure (groups vary in size or density). Given that most practical cases will fall in the "complex" spectrum, we propose a novel dendrogram analysis technique that provides adequate splitting in some complex cases where traditional dendrogram analysis methods might fail. The method proposed here is a Ward's dendrogram analysis technique. This method can be used in any clustering algorithm that uses Ward's hierarchical clustering. It has an intuitive parameter tuning and results show it can outperform current methods when dealing with groups with different density, number of observations and shape.

The rest of the chapter is organized as follows. In Section 2.2, we present a current state of the art in dendrogram analysis, and we identify the gaps in the literature that motivates our study. In Section 2.3, we discuss **JAM** (Jump Analysis Method), the methodology proposed for dendrogram analysis. In Section 2.4, we present results on simulated and real data sets and compare the performance of our method against K-means. In Section 2.4.4, we make the concluding remarks.

2.2 Literature review

There is rich literature on clustering techniques. The work by Berkhin [2006] presents a detailed survey on different clustering techniques, classifying each algorithm into hierarchical, partitioning density-based, grid-based, etc. The review also explains some of the principal advantages and disadvantages of each clustering algorithm.

For a performance comparison between different clustering techniques, we suggest to look into the studies of Mingoti and Lima [2006], Budayan et al. [2009]. The work by Mingoti and Lima [2006] studies the performance of different clustering methods under the presence of outliers, variable correlation and overlapped clusters. The study by Budayan et al. [2009] did a performance comparison for a case related to the Turkish construction industry and an interesting survey about performance comparisons between

different clustering techniques. In commonality, the survey papers Downs and Barnard [2002], Budayan et al. [2009] point out the poor performance of hierarchical clustering algorithms in situations where density, and number of observations varies between groups or groups are non-spherical. The works by Mingoti and Lima [2006], Berkhin [2006] concur that hierarchical clustering algorithms are sensitive to noise and outliers.

2.2.1 Hierarchical clustering

The result of hierarchical clustering is a dendrogram and, by itself, it cannot split the observations into groups. The clustering literature shows recent work on how to select groups after hierarchical clustering analysis [Sander et al., 2003, Almeida et al., 2007]. The works by Sander et al. [2003], Almeida et al. [2007] show methods for obtaining clusters given a hierarchical structure based on single linkage [Sibson, 1973]. The work by Sander et al. [2003] proposes to select the groups using the reachability plot and also shows how to convert a dendrogram into a reachability plot given a hierarchical structure based on single linkage. The work by Almeida et al. [2007] proposes a three-step algorithm, starting with a method for outlier removal, then forming the clusters, followed by selecting of the clusters that classify the removed observations. Our work resembles that of Sander et al. [2003], Almeida et al. [2007] in that we also create a method to create clusters derived from a dendrogram.

2.2.2 Stopping rules

There is a lot of research in a topic called “stopping rules,” still active with contemporary work by Salvador and Chan [2004], Gong and Oard [2009], Jung et al. [2003]. These stopping rules consist of establish an index or performance measure and stop the hierarchical clustering algorithm based on a stopping criteria. This limits all possible grouping solu-

tions to the ones given by horizontal cuts, making specially difficult to find good grouping solutions when groups vary in density and size. In contrast, our method allows to cut a dendrogram at different levels simultaneously, which is of interest in situations where densities and sizes vary between groups.

2.2.3 Hierarchical segmentation

There is a field of study called hierarchical segmentation related to image processing. For example, "Climbing" [Kiran et al., 2012] is an image processing technique that allows to cut a dendrogram at different levels. Due to nature of image data sets, where the data set is organized by equally distributed pixels, these techniques cannot be used in general data mining applications.

2.2.4 Iterative clustering algorithms

Extensive research has been conducted involving the creation of different iterative clustering algorithms, allowing to define clusters in complex data sets [Guha et al., 1998, Karypis et al., 1999, Zhang et al., 1996, Guha et al., 1999]. The work by Langfeldera et al. [2007] is an iterative process that first selects big groups out of a dendrogram and then creates a new dendrogram out of each of these groups. The method we propose, on the other hand, is a single step method. That is, we take a single Ward's dendrogram and select the best possible groups given a combination of parameters. Nevertheless, it is up to the user to carry out parameter tuning to find the most suitable combination of parameters.

The method we propose is not serial, in the sense that the results for a clustering output are not used to perform another clustering process. The purpose of this study is to simplify the process of extracting clusters from a dendrogram generated with Ward's linkage

method [Jr., 1963]. The proposed method uses multilevel cuts at a time and, to the best of our knowledge, there is no work in the literature that allows for multiple cuts at different levels from a Ward's dendrogram.

2.3 Methodology

(**JAM**) uses as input a dendrogram generated with Ward's hierarchical clustering [Jr., 1963] to select clusters out of a data set. First, the method selects merge points for cutting the dendrogram. The method aims to locate the merge points which separate two clearly different groups in the dendrogram. Once we select the merge points to cut the dendrogram, the groups or clusters will surface. Figure 2.1(a) shows a dendrogram with the merge points identified using capital letters and the observations represented using numbers. Each merge point fuses two groups and represents a step in the hierarchical clustering algorithm.

The work by Mingoti and Lima [2006] include a review of the most relevant hierarchical clustering applications using the open-source statistical software R [R Core Team, 2013]. For more details of the R software used refer to Appendix C. If we are using "Ward.D2" method we need to introduce `dist(data)` as distance, not `dist(data)`^{2} as in other methods.

In order to discriminate between groups using a dendrogram, it is desirable to search for merge points with large heights (based on Ward's minimum variance) relative to the last merge point in the same branch. The higher the height relative to the last merge, the more 'different' the groups are from each other. We use two different performance measures to categorize the relative heights in order to decide whether a merge point is selected for splitting or not.

First, we calculate the **absolute jump** of a merge point *i* relative to the previous merge

point in the same branch, k. Let J_{i} denote the jump of merge point i and define it as the difference between the *height* of that merge point, H_{i}, and the height of the previous merge point in the same branch, H_{k}, i.e. $J_{\mathrm{i}} = H_{\mathrm{i}} - H_{\mathrm{k}}$.

Merge points with higher jumps are candidates for splitting. In Figure 2.1(b), we provide a graphical example of the jump and the height of a merge point. Observe that the *height* of B is the y-axis (height) coordinate for merge point B and the **absolute jump** of B is the difference between the y-axis coordinate of B minus the y-axis coordinate of the closest merge point in the same branch, that is, the y-axis coordinate of merge point D.

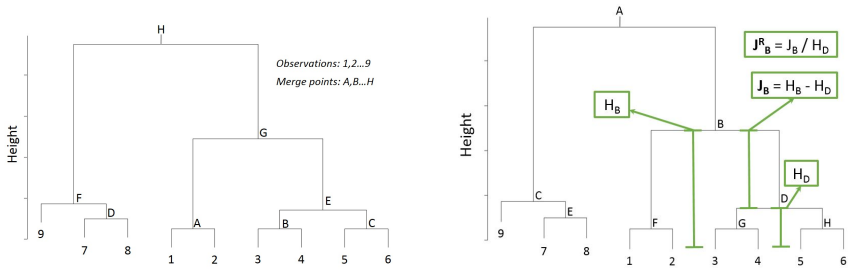
In a Ward's dendrogram, it is expected to have higher **absolute jumps** for the merge points at the top of the dendrogram (reducibility property 2.1. For more detail see the work by Olson [1995]). Therefore, the exclusive use of **absolute jumps** to determine the cutting points would result in a misrepresentation of the different groups.

$$d(A \cup B) \geq min(d(A, C), d(B, C)) \tag{2.1}$$

In order to mitigate the effect of the reducibility in Ward's dendrograms, we look at the **relative jump** of a merge point i, $J_{\mathrm{i}}^{\mathrm{R}}$ as a new metric to consider when cutting a merge point. The relative jump is defined as the ratio of the jump, J_{i}, and the height of the precedent (i.e. previous merge point in the same branch) merge point, H_{j}, i.e. $J_{\mathrm{i}}^{\mathrm{R}} = J_{\mathrm{i}}/H_{\mathrm{j}}$. The merge points with higher **relative jump** values are also candidates for cutting. Next we show a graphical example to clarify these concepts

In Figure 2.1(a), we show a dendrogam with merge points identified by capital letters A to H and observations identified by numbers 1 to 9. In Figure 2.1(b), we provide an example of the **relative jump**, **absolute jump** performance measures relative to merge point B as example. H_{B} Identifies the height of merge point B. Then, J_{B} identifies the **absolute jump** for merge point B and is calculated as $J_{\mathrm{i}} = H_{\mathrm{B}} - H_{\mathrm{D}}$. Finally, $J_{\mathrm{B}}^{\mathrm{R}}$ identifies

the **relative jump** for the merge point B and is calculated as $J_{\mathrm{B}}^{\mathrm{R}} = J_{\mathrm{B}}/H_{\mathrm{D}}$



(a) Dendrogram with merge points identified (b) Relative jump of merge point B, $J_{\mathrm{B}}^{\mathrm{R}}$

Figure 2.1: Dendrogram description and jumps

Combining both performance measures **absolute jump** and **relative jump**, we select the merge points we are cutting which results in the definition of the clusters. This method uses as input a data set and two parameters, defining the percentage of merge points we are selecting from each metric. The thresholds work as follows: If we establish an 80% threshold for the first metric, this will mean only the 20% merge points with higher **absolute jump** are candidates for splitting. The threshold for the second metric works the same way on the marginal distribution of **relative jumps**. The merge points that satisfy both performance measures are the points we are finally splitting. Next, we explain how to select the groups in a dendrogram given the merge points selected by the cutting method.

Each combination of thresholds can result in a different selection of merge points for splitting, i.e. will result on different groupings. We suggest changing the parameters using the scheme depicted in Figure 2.2 for the parameter tuning stage or when the desired number of clusters has not been found. We recommend to start the search with the largest value for both thresholds (it produces the least computationally demanding solutions). We found that grouping solutions with **relative jump** threshold higher than **absolute jump** threshold are not usually good solutions, therefore, we suggest to start to reducing the **relative jump** threshold, see Figure 2.2. The user defines the amount and value of thresholds

at tuning stage. We normally suggest values greater than 80% for both thresholds.

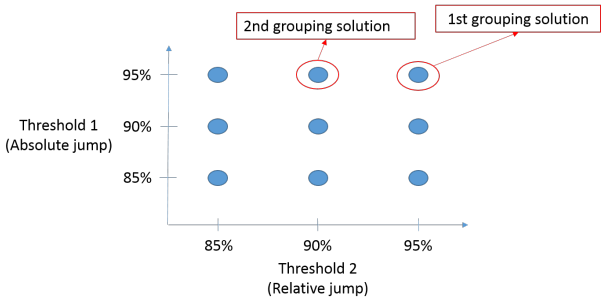
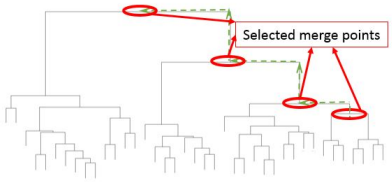


Figure 2.2: Thresholds combinations and grouping solutions.

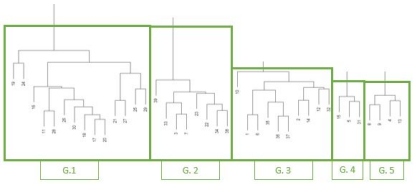
In Figure 2.3, we present a sequence of graphical representations for the group selection. First, we select the merge points that go over both thresholds. Next, we draw the path from each selected merge point to the top of the dendrogram 2.3(a). Later, we include as selected merge points each merge point in contact with the drawn paths, see Figure 2.3(b). Then, we delete from the dendrogram all the horizontal lines contacting with the selected merge points, see Figure 2.3(c). Finally, the formed groups are the ones containing the observations still in contact with dendrogram lines, see Figure 2.3(d).

(a) selected merge point & drawn Path



(b) Merge points in contact with the path included as selected merge points

(c) Lines in contact with path deleted

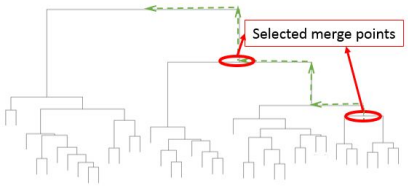


(d) Groups defined

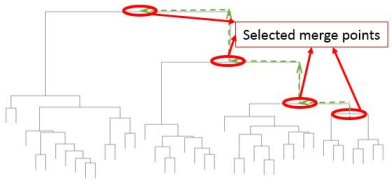
Figure 2.3: Sequence for identification of groups.

In order to limit the possibility of having groups containing only one individual. If a merge point is merging two individual observations, that point will not be selected for cutting.

Figure 2.4 has different set of parameters than Figure 2.3(a), leading to one additional selected merge point in Figure 2.4(a). Finally, both cases lead to the same grouping. In conclusion, different set of parameters can lead to different combinations of selected merge points for cutting a dendrogram and still produce same grouping results.

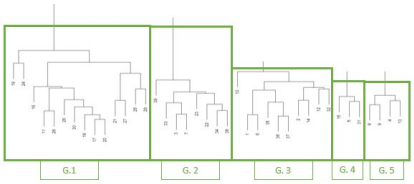


(a) selected merge point & drawn Path



(b) Merge points in contact with the path included as selected merge points

(c) Lines in contact with path deleted



(d) Groups defined

Figure 2.4: Example with grouping result equal to Figure 2.3.

We have defined a methodology that allows remove the subjectivity in the process of defining groups out of Ward's dendrograms. Next, we show the Algorithm 1 describing the presented methodology.

2.4 Results and Discussion

Previous work have stated that K-means and Ward's method tend to fail when separating groups of different sizes, densities, or shapes [Downs and Barnard, 2002, Budayan et al.,

Input : A Ward's dendrogram D /*based on a dataset of n observations.*/
Output : Clustered observations
Require: $t_A \leftarrow$ a /*User defined absolute jump threshold.*/
$t_R \leftarrow$ b /*User defined relative jump threshold.*/
M_i:i=1,2,...,n-1 /*M is the set of all merge points in D*/;
for *each merge point* $M_i \in M$ **do**
  $J_i \leftarrow CalculateAbsoluteJump(M_i)$;
  $R_i \leftarrow CalculateRelativeJump(M_i)$;
end
$p_A \leftarrow CalculatePercentile(J,t_A)$ /*Calculate the a^{th} percentile of the set of absolute jumps J*/;
$p_R \leftarrow CalculatePercentile(J,t_R)$ /*Calculate the b^{th} percentile of the set of relative jumps R*/;
$S \leftarrow \emptyset$;
for *each merge point* $M_i \in M$ **do**
  if *(*$J_i \geq p_A$*) and (*$R_i \geq p_R$*)* **then**
    $S \leftarrow S \cup M_i$;
    /*S is the Set of all merge points that satisfy both thresholds*/;
  end
end
for *each merge point* $M_i \in M$ **do**
  if *(*$M_i^l \in S$*) or (*$M_i^r \in S$*)* **then**
    $S \leftarrow S \cup M_i$;
    /*M_i^l the adjacent child merge point to the left of M_i*/;
    /*M_i^r the adjacent child merge point to the right of M_i*/;
  end
end
ClustResult $\leftarrow \emptyset$;
for *each merge point* $M_i \in S$ **do**
  $C_i^l \leftarrow LeftChildCluster\ (M_i)$;
  $C_i^r \leftarrow RightChildCluster\ (M_i)$;
  $ClusterLeft_i \leftarrow \overline{ClustResult} \cap C_i^l$;
  $ClusterRight_i \leftarrow \overline{ClustResult} \cap C_i^r$;
  if *ClusterLeft*$_i \neq \emptyset$ **then**
    ClustResult $\leftarrow$ ClustResult $\cup$ ClusterLeft$_i$;
  end
  if *ClusterRight*$_i \neq \emptyset$ **then**
    ClustResult $\leftarrow$ ClustResult $\cup$ ClusterRight$_i$;
  end
end
Return ClustResult;

Algorithm 1: JAM; Ward's dendrogram analysis technique.

2009]. Ward's method can also fail with non-convex shapes [Berkhin, 2006]. The work by Kaur and Kaur [2013] further suggests K-means performs better than hierarchical clustering in large data sets, whereas hierarchical clustering takes the lead in smaller data sets. K-means outperforms Ward's method with horizontal cuts, mainly because these cuts fail to address the group diversity in size and density commonly found in large data sets.

Next, we compare the performance of the proposed method on eight scenarios: (1-2) two simulated data sets. (3-5) three publicly available simulated data sets, and (6-8) three real data sets provided by collaborators.

We use the logic explained in Algorithm 2 to compute the time consumed and to find the grouping solutions. The search stops when it finds the desired number of clusters for each data set.

Depending on how restrictive is the threshold criteria, it is possible that our method cannot find the desired number of clusters. If the method does not find the desired number of clusters, it could be that we did not examine the correct thresholds, or that our method cannot find the desired number of clusters for the given dendrogram structure. We use purity as a performance measure. We define purity as the proportion of observations located (by the clustering method) in the group they belong. We also compute the standard error for the purity proportion and the average system time consumed. Next, we show the

grouping results for the tested scenarios.

Data: Data set

T1seq; *threshold* 1 *sequence*

T2seq; *threshold* 2 *sequence*

#clust; *the desired number of clusters to find in the dataset*

Input : data set d_S /**based on a dataset of n observations.*/

Output : ClustResult

Require: k ← 5 /**Known number of groups.*/

A ← 0.99,0.98,...,0.80 /**Sequence of absolute jump thresholds.*/

R ← 0.99,0.98,...,0.80 /**Sequence of relative jump thresholds.*/

D ← hclust(d=dist(d_S),method = "ward.D2") /**hclust defined as the function in R software.*/;

ClustResult ← ∅;

for *each* $A_i \in A$ **do**

  for *each* $R_j \in R$ **do**

    $Clustering_{ij}$ ← JAM(A_i, R_j,D);

    if *NumberClusters(*$ClustResult_{ij}$*) == k* **then**

      ClustResult ← $Clustering_{ij}$;

      break outerloop;

    end

  end

end

Return ClustResult;

Algorithm 2: Algorithm applied to data sets of the different scenarios.

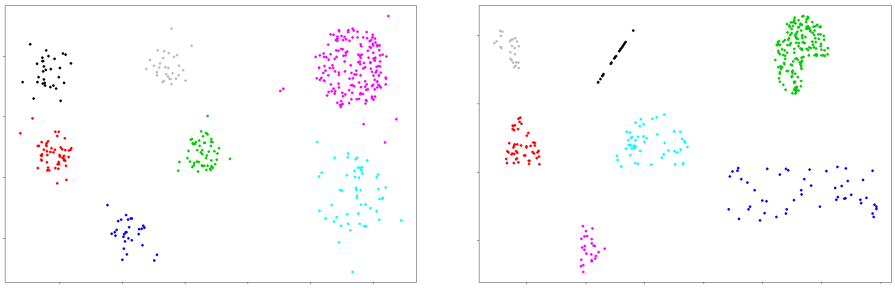
2.4.1 Simulated data sets - Scenarios 1 and 2.

We tested the proposed method on two different simulated data sets. These scenarios were designed to prove our method can overcome complexities where traditional clustering methods fail (e.g, groups of different sizes, densities, or shapes). Both scenarios have groups varying in density, area, and number of observations. Figure 2.5 has a graphical representation of the sets. Observe that Scenario 1 has a small sample of outliers (5 percent) and spherical groups. In Scenario 2, there are no outliers and groups have different shapes. Each scenario has seven groups and around 400 observations.

We generated 30 replicas of each scenario. Purity is the performance measure used in order to evaluate the clustering algorithms. We also show the standard deviation of the purity measure and the average time consumed for the 30 replicates in each scenario. We then compare the grouping results of our method with K-means, using purity as performance measure. For K-means we search for seven clusters, using random generated initial centroids in one case and initial centroids generated with Ward's hierarchical clustering and a horizontal cut to form seven groups in the other. Our method searches for seven clusters using the procedure described in Algorithm 2.

Next, in Figure 2.5, we show the two simulated scenarios and the clustering results supporting what we propose.

Results in Table 2.1 show our method always outperforms K-means for the proposed scenarios. The solutions proposed by our cutting method could have never been found with a horizontal dendrogram cut (i.e. stopping rules), since these solutions require cutting the dendrogram at different levels simultaneously. Time consumed is 22% greater in average for our dendrogram cutting method. Next, we show the results for the three publicly available data sets simulated by others.



(a) Scenario 1 (b) Scenario 2

Figure 2.5: Plot of tested scenarios: (a) first scenario (b) second scenario.

Table 2.1: Description of each scenario's results $n = 30$ replicates using as input parameters Threshold1 sequence = (0.99, 0.98, 0.97 ... 0.82, 0.81, 0.80), Threshold2 sequence = (0.99, 0.98, 0.97 ... 0.82, 0.81, 0.80) and Number of Clusters = 7.

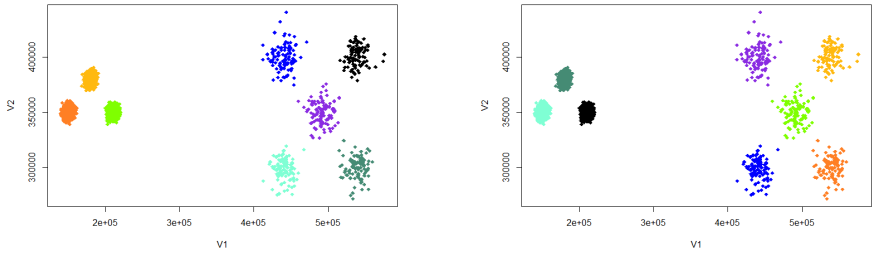
		Purity		System time(sec.)
		Average	Standard error ($x10^{-4}$)	
JAM	Scenario 1	0.997	1.06	0.022
	Scenario 2	1.000	0	0.022
K-means 1	Scenario 1	0.843	7.98	0.017
	Scenario 2	0.852	7.57	0.018
K-means 2	Scenario 1	0.932	5.61	0.019
	Scenario 2	0.971	2.64	0.018

2.4.2 Publicly available simulated data sets - Scenarios 3, 4 and 5.

We tested the method on 3 different data sets available at Speech and Unit [2016] The data sets we used are: "Unbalance", this data set has 8 groups and 6,500 observations, "S1" this data set has 15 groups and 15,000 observations and "S2" this data has 15 groups and 15,000 observations. . These data sets do not have a variable indicating the group for each observations; Hence we cannot calculate purity as performance measure after grouping for these data sets. Therefore we make a qualitative comparison. We plot the results for each method with color codes to see if the grouping was successful.

For data set "Unbalance", we compare the grouping results of our method with K-

means searching for 8 groups, our method stops the search with "threshold1" = 0.99 and "threshold2" = 0.94. Figure 2.6 shows the grouping results for the "Unbalance" data set. Both methods divided data set "Unbalance" in groups with similar performance. (K-means with initial centroids at random was not used because we compare the results graphically, so we cannot make an average of different K-means repetitions).



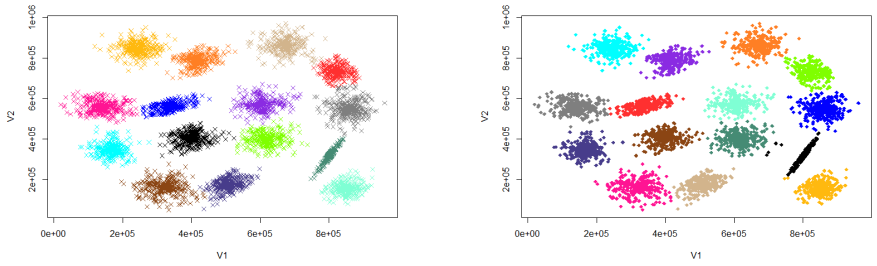
(a) JAM, Threshold sequences as in 2.4.1

(b) K-means using initial centroids generated with Ward's hierarchical clustering

Figure 2.6: Results for the "Unbalance" data set: (a) with JAM and (b) with K-means

For data set "S1", we proceed as with the set "Unbalance", searching for 15 clusters, using initial centroids generated with Ward's hierarchical clustering. Our method looks for 15 clusters using the procedure described in Algorithm 2, our method stops the search with "threshold1" = 0.99 and "threshold2" = 0.97. Figure 2.7 shows the grouping results for the "S1" data set. Both methods successfully group data set "S1".

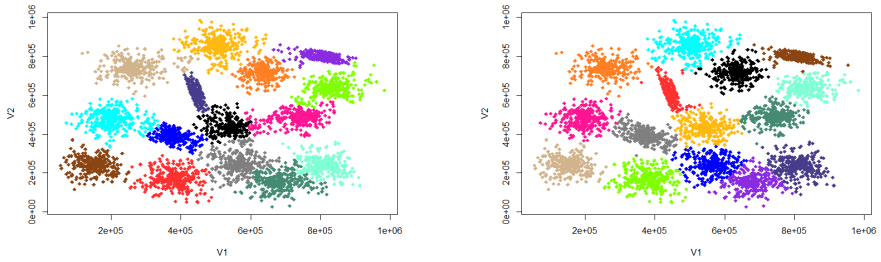
The data set "S2" has 15 groups and 15,000 observations. We proceed as with the set "S1", our method stops the search with "threshold1" = 0.99 and "threshold2" = 0.93. Figure 2.8 shows the grouping results for the "S2" data set. K-means showed slightly better performance than **JAM** clustering this data set. **JAM** had some problems separating the dark blue and light blue group 2.8(a), and the pink and black groups. K-means had problems separating the red group



(a) JAM, Threshold sequences as in 2.4.1

(b) K-means using initial centroids generated with Ward's hierarchical clustering

Figure 2.7: Results for the "S1" data set: (a) with JAM and (b) with K-means



(a) JAM, Threshold sequences as in 2.4.1

(b) K-means using initial centroids generated with Ward's hierarchical clustering

Figure 2.8: Results for the "S2" data set: (a) with JAM and (b) with K-means

2.4.3 Real data sets

We tested the method on three different real data sets. The first data set is the "Wine data set" available at Lichman [2013]. The second data set is known as PAM50 related to breast cancer and gene expression, for more detail please see Parker et al. [2009]. We used the gene expression variables to group the data. The cancer subtype variable is used to evaluate purity of the formed groups. The third data set is known as RNA Expression used in Network et al. [2012], we used this data set, normalized at level 3 format, and only the 50 PAM50 genes for grouping observations, then the cancer subtype variable is used

to evaluate purity of the formed groups again.

Next we show the results for these 3 data sets in Table 2.2. We use purity as performance measure as in 2.4.1. For both cancer datasets, Scenarios 7 and 8, our method didn't find 5 groups, so we decided to select the solution with 4 clusters for both cases. In scenario 7 the purity results were better for our method even with 4 groups.

Table 2.2: Description of data set's results using as input parameters Threshold1 sequence = (0.99, 0.98, 0.97 ... 0.82, 0.81, 0.80), Threshold2 sequence = (0.99, 0.98, 0.97 ... 0.82, 0.81, 0.80) and Number of Clusters = 3(for wine data set), 5 (for breast cancer data sets). 4

	Purity			System time(sec.)		
	JAM	K-m 1	K-m 2	JAM	K-m 1	K-m 2
Scenario 6	1	0.843	0.932	0.05	0.001	0.02
Scenario 7	0.906	0.846	0.892	0.03	0.004	0.01
Scenario 8	0.76	0.806	0.859	0.02	0.001	0.01

Our method performs very well for the first two data sets. The third data showed low purity results. Examining Scenario 8, we found that the root cause for the low purity grouping result was on the Ward's dendrogram construction 2.9. Observations of different cancer subtypes are mixed in the same branch all along the dendrogram. The observations in colors green, yellow and blue are hard to separate in groups given this dendrogram.

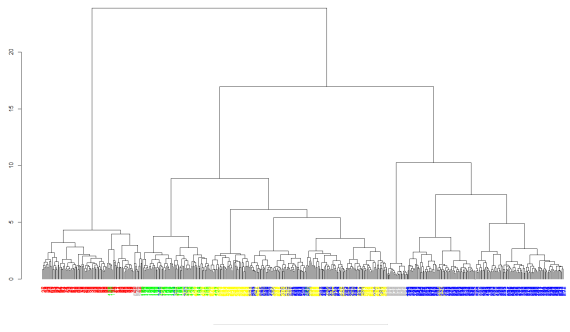


Figure 2.9: Dendrogram for the third data set.

2.4.4 Conclusion

There is a lot of research involving the creation of different clustering algorithms that work in complex situations. Still, classic algorithms such as K-means and hierarchical clustering algorithms with the different linkage methods are widely used in scientific articles.The main reason for that is the wide variety of algorithms available and their complexity. In this study we present **JAM**, (Jump Analysis Method) a novel technique for identifying groups in Ward's dendrograms.

Overall, results show that the method is capable of: automating the process of extracting clusters from a Ward's dendrogram, and outperforming k-means in complex scenarios where clusters may vary in density, size, and shape.

It has some limitations as it is a method based on Ward's hierarchical clustering. Big data sets can cause problems to this analysis technique. Non convex shapes can also cause problems. Still, **JAM** may overcome some of the traditional problems for hierarchical clustering, such as diversity in size, shape and density between groups. Additionally, **JAM** is an easy to use and understand analysis tool. Finally, this method shows a new way to look at dendrograms, encouraging researchers to develop methods for cutting dendrograms at different levels simultaneously.

Chapter 3

Farmer's Needs in Puerto Rico

This chapter describes the methodology used to collect, analyze, and transform information to estimate regional agricultural needs for the test bed. The methodology includes a study that characterizes farmer's needs given his/her profile. The study uses a survey of needs (questionnaire) and clustering techniques described in the previous chapter. Furthermore, this chapter most important contribution is to add a model that uses farmer-level needs characterization combined with municipal profiles to estimate municipality-level needs. Municipality needs will be used as input for the resource assignment model.

3.1 Introduction

On a service-system it is extremely important to know customers' needs. Management should plan system's operation and resources in order to satisfy those needs. For some systems, specially in service systems, knowledge of services needed by the clients is not readily available. Tax et al. [1998] shows how important and difficult is to keep and develop relationships with current customers in service organizations, this is the case for the test bed in this thesis. The resource assignment model is tested using as test bed the

University of Puerto Puerto Rico Agricultural Extension Service (AES). Specifically, we target the instructional service Agricultural Agents (AA) provide to farmers. Hence, information about the farmer's needs is used as ground to develop a prediction model on service needs by region (i.e. municipality). Regional needs will be used as input for the assignment model. This Chapter describes the methodology used and the results obtained for the analysis of farmer's needs for advice in each Puerto Rican Municipality.

The methodology designed to estimate Puerto Rican farmers' needs at a municipality-level, has as starting ground information obtained from farmers employing a questionnaire. The questionnaire collected information about farmer's profile and his/hers particular need for assessment. This questionnaire was administered to 101 farmers in PR where different sectors of farmer's community were well represented. Clustering analysis was used to identify or link farmer's needs to his/her profile.

Classification techniques were later used to link farmer's profile to municipalities profile. The term "municipality profile" in this chapter refers to the agricultural status and composition of that municipality. In order to construct a municipality profile, the 2012 Puerto Rico' Agriculture Census 3.1 was used. This profile includes as input, number of farmers in each municipality, number of farmer's per product categories in each product category and education level of farmers in each municipality.

This chapter adds to the literature, as the first study predicting farmer's needs by municipality in Puerto Rico. Moreover, no similar study was found that uses the methodology provided in this chapter were individual's profiles were used, combined with Census data, to establish "regional" profiles.

This methodology includes a sampling strategy to ensure population representativeness, see Section 3.3.1. Details of Data Collection Process in Section 3.3.1. Data Transformation and Integration in Sections 3.3.1 and 3.3.1. Clustering analysis using Ward's

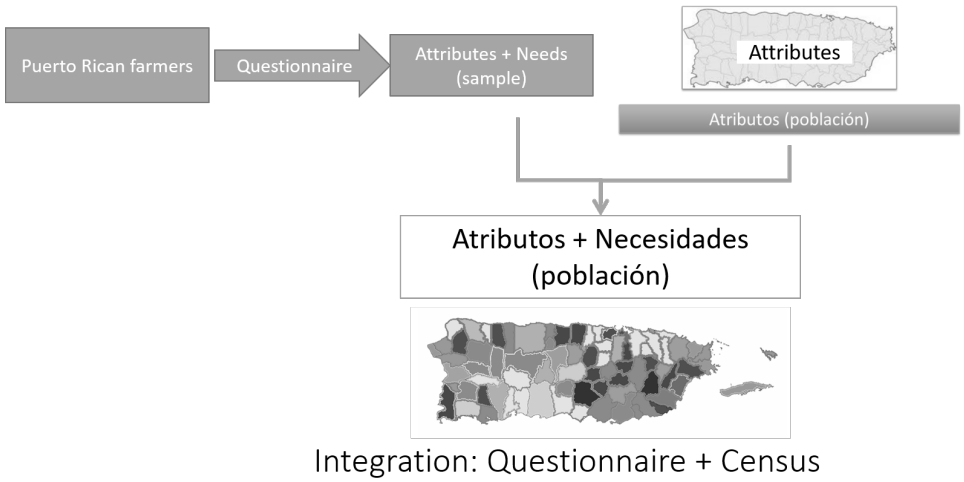


Figure 3.1: Farmer's questionnaire and Census integration strategy.

hierarchical clustering [Jr., 1963] and JAM (Chapter 1). A classification model to predict the cluster of a municipality. Finally an estimation of farmers' needs in Puerto Rican municipalities. We show and discuss the results in Section 3.4 and make some concluding remarks in Section 3.4.1.

3.2 Related Literature

Ideally, for the assignment model, the input of demand should be on the form of predictions on types and quantity of services needed for each PR municipality. However, no study or document exists where such information is readily available; neither a study of instructional needs of farmers at an individual level. To be able to test the resource assignment model, it was imperative to estimate and define farmers' needs.

Farmer's needs are diverse. The general assumption of many studies (made before this one) is that farmers' need specific technical support related to their agricultural product. For example, legume farmers need advice in technical topics related to legumes and cattle farmers need advice in technical topics related to cattle. However, we could argue that

most of the farmers are experts producing their products and they need advice in other topics, for example renewable energy or even marketing.

This chapter challenges the assumption of farmers needing advice in topics related to their specific agricultural product and explore other possibilities of needs. In literature there exists several studies of farmers' needs, most of them focus on studying technical issues, often in specific topics such as pest management [Hashemi et al., 2009]. So far, the only study of farmers' needs in a general way, was one from a reform of Chinese Agricultural Extension Service that began in 2005 (Inclusive Public Agricultural Extension Service) [Hu et al., 2012]. The main focus of this reform and its corresponding study was the farmers' access to counseling Extension Service as shown in Hu et al. [2012], a questionnaire was conducted to assess the technical needs of farmers. However, the study did not include any business need for farmers. Therefore, the studies about farmers needs we have found, only cover technical needs. Next, we present the methodology for the study of farmer's needs in Puerto Rico.

3.3 Methodology

The methodology employed in this chapter is divided into three stages. The first stage will study and link farmer's profiles to needs. The second stage profiles municipalities using its agricultural activities reflected in the 2012 Agricultural Census. The third stage will merge the characterization of farmer's needs to municipal agricultural activities in order to obtain a prediction of the AES services needed of a municipality.

3.3.1 Characterization Farmer's Needs

In this study, we are going to group farmers using Ward's hierarchical clustering [Jr., 1963] combined with JAM (Chapter 2), as a complete clustering technique. Grouping farmers will make the questionnaire output easy to understand. Previous studies like Govind et al. [2008] and Shih and Liu [2003] have proposed customer grouping. These cases serve as examples to the usefulness of grouping clients because easily understandable information is obtained.

First, we created and delivered a questionnaire to a sample of Puerto Rican farmers, see section 3.3.1. We designed an integration strategy 3.3.1, so we can use the same models with Census and questionnaire data sets. Then we group customers using Ward's hierarchical clustering [Jr., 1963] and JAM (Chapter 1). This clustering analysis 3.3.2 allowed us to characterize farmers in four different profiles considering their demographic variables (attributes) and their needs. The output from clustering analysis was the input for a classification model. The purpose of the classification model was to predict the cluster of farmers given their demographics 3.3.3. We trained the model with the questionnaire data set. Then, we use the classification model to predict the cluster of each Puerto Rican municipality. Assuming it is reasonable to use the same model for both data sets thanks to the data transformation and integration strategy. Finally, we predicted the needs for each Puerto Rican municipality calculating a weighted average of the k nearest neighbours inside the predicted cluster. Next, we present the details of this methodology.

Farmer's Needs Questionnaire

A Farmer's Needs Questionnaire was created as a first step in order to obtain data on farmer's attributes and their needs. A sample questionnaire is included in Appendix A. This questionnaire was created using the following different domains of information: (a) feedback from Dr. Guillermo Ortiz (Professor at UPRM College of Agricultural Sciences),

(b) informal interviews with different farmers, obtaining first-hand information about possible farmer's needs, (c) specialized articles in agricultural sciences [4-H, 2015, Hu et al., 2012, Chase et al., 2006, Hashemi et al., 2008, Manjala et al., 2009, Ruifa et al., 2009]. The questionnaire administration protocol was approved by the institutions' "Comité para la Protección de los Seres Humanos en la Investigación" (CPSHI/IRB) under protocol number 20141110 B. Approval letter is included in Appendix B.

Questionnaire data was structured using the key provided in Table 3.1. Observe from the table that there are six main categories, i.e. municipality, relevance of agriculture as economic activity, education level, importance of each product, needs, and preferred advice channel. The key indicates the possible answer alternatives for each question.

It is important to disclose that the last category in the key (preferred advice channel) was added to the questionnaire per request from AES management. Management were interested in learning the results from this question, hence it was added to the questionnaire effort. Results from this question are reported in this work but it is not used for the assignment model.

Data Collection

Recruitment of farmer's to participate in the study was challenging. The sampling plan required that the venues and contact with farmer's were not coordinated by the AES. The research team was interested in obtaining unbiased answers to the questionnaire to reduce possible pressures from farmer's part. This in turn was challenging given that the team had limited number of venues to capture farmer's interest. Moreover, each questionnaire took 15 minutes (approximately) to answer and for some farmers' venues that implied 15 minutes they were not attending customers. To this end, it is important to acknowledge that the personnel from Puerto Rico's Department of Agriculture helped in farmer's recruitment in a great way. Given the challenges provided above, considerable time was invested in data

Table 3.1: Information output from questionnaire

Item	Key
Municipality	Municipality where farm is located
Agriculture: main economic activity	0 - No
	1 - Yes
Maximum education level	0 - Intermediate School
	1 - High School
	2 - Certificate
	3 - Associate grade
	4 - Bachelor
	5 - Master
	6 - Phd
Importance of product 1 (Plantains)	No production or insignificant
	1 - Little importance
	2 - Important
	3 - Very important
	4 - Extremely important
...	...
Importance of product N (Swine)	No production or insignificant
	1 - Little importance
	2 - Important
	3 - Very important
	4 - Extremely important
Need 1	0 - N/A
	1 - Low
	2 - Medium
	3 - High
...	...
Need 20	0 - N/A
	1 - Low
	2 - Medium
	3 - High
Preferred advice channel	0 - Visit to farms
	1 - Phone advice
	2 - Article publication
	3 - Method demonstration at farm
	4 - Method demonstration at Experimental Station
	5 - Meetings to talk about a specific topic

collection.

Sampling Strategy

The sampling methodology objective is to ensure representativeness of the population. The sample has a product category stratum. The sampling strategy is described next.

> First Alternative: The first alternative we considered was to establish two categories; crops and livestock. This alternative will be very easy to implement although has a drawback that may lead to unrepresented categories. For example, if we require the sample to collect data from owners/managers of 40 livestock farms, it could result in a sample with some of the animal production activities (swine, cattle, etc.) having little or no representativeness.
>
> Second Alternative: The second alternative was to establish one category for each product family. This alternative has an advantage in terms of tasks assignment because this will allow us to see if there is a relationship between product categories and needs. For example, it would be interesting to learn if fruit farmers have different needs compared to others farmers, and what are these needs.
>
> Third Alternative: The third alternative would be to establish one category for each agricultural product. This would imply the need of collecting data from 2 lemon farmers, 2 coconut farmers and 4 plantain farmers instead of 8 fruit farmers. This detailed level would make data collection a very difficult task.

The alternative chosen for the product category stratum was the Second Alternative given that it provides an advantage for the study without over-complicating the data collection process.

Two different sources of information were used in order to establish the product family categories as described before in the Second Alternative. The first source of information

used was the Puerto Rico Agricultural Census 2012. This Agricultural Census shows a classification of livestock farms in line with the needs of this study. However, crop farms classification is not in line with the chosen stratum alternative, which posses a challenge.

The Agricultural Census defines a category named "main crops". This category includes coffee, plantains and pineapples. The reason to group these crops together is because they are the most common in Puerto Rican farms. This category groups crops that are not necessarily in the same family; this could mean these farmers have different needs in terms of the technical assessment. Therefore, the expertise needed by farmers in this group is not expected to be the same. This is why we chose the crop category division used by the United Nations in the World Programme for the Census of Agriculture 2012. Please see Table 3.2 and Table 3.3 for details about the categories used.

Table 3.2: Puerto Rico Agricultural Census's (2012) livestock farms categories

Livestock Farms Categories
Cattle and Calves
Swine
Poultry
Aquaculture
Other Animals

Table 3.3: World Programme for Census of Agriculture's crop farms categories

Crop Farms Categories
Fruits
Vegetables and melons
Legumes
Tubers
Spices and beverage crops
Sugar crops
Oilseed crops
Cereals
Other crops

These categories are important for this study, since we want the sample for each category to be proportional to the total number of farms in Puerto Rico (using the Puerto Rico Census of Agriculture 2012) belonging to each class. This ensures representativeness of all farmers' activities.

Data Collection Efforts and Venues from Questionnaire

This study's data was collected by administering the Farmers Needs Questionnaire. We started to collect data with questionnaires in paper assisting the farmer during the completion of the questionnaire. This procedure is approved by CPSHI 20141110 protocol. Due to the effort required to obtain data this way we later decided to administer the questionnaire online. The data collection included the following venues and contacting mechanisms:

a. Mayagüez family market
b. Las Marías family market
c. Rincón farmers' market
d. San Sebastián farmers' market
e. Market at an agricultural fair in Mayagüez
f. UPRM Agricultural Market
g. Exhibition of goats and sheeps at an agricultural fair in Mayagüez
h. AES stakeholders meeting of vegetables and grains in Juana Díaz
i. AES stakeholders meeting of cattle in Hatillo
j. Conference on goats and sheep in Mayagüez
k. Workshop on soil conservation incentives in Utuado
l. Puerto Rico Department of Agriculture
m. Online

Data Transformation

The questionnaire has one variable indicating the importance of each agricultural product (e.g. tomatoes, milk, etc.) for each farmer. However, it is of interest to aggregate this variable at the agricultural product category level (e.g. fruits, cows) so we can identify similar farmers for the study of the data. It is desired an aggregation meeting the following requirements.

1. The aggregated value cannot have a value greater than the highest individual value, 4 in this case.

   Example: If having a farmer growing only mangoes with the highest importance level 4, in this case the farmer should have a 4 in the importance level for fruits. A farmer growing lots of fruits cannot have a greater value; If allowed a greater value, this would mean that fruits would be more important for the second farmer and both farmers only grow fruits.

2. The aggregated value should be greater or equal than the highest value of an individual product in the category.

3. Each value greater than zero should contribute to the aggregated value for the corresponding category.

   This only applies if we have not reached the maximum value, 4 in this case.

With 4 as the highest value, the aggregation method is described next. Let V_i^j be the importance-value of product $i = \{1, ..., N\}$ within category j, where N is the number of products in category j. Now consider V^j as the aggregated importance-value of category j for the farmer. This value, i.e. V^j, for any j will be calculated as follows:

Initialization: $i = 1$, $V^j \leftarrow V_i^j$

Recursive Steps: Next i, $V^j \leftarrow V^j + (4 - V^j) + (4 - V^j) * \left(V_i^j/5\right)$

Final Step: $i = N$, $V^j \leftarrow V^j + (4 - V^j) + (4 - V^j) * \left(V_i^j/5\right)$

The process described above fulfills all the initial requirements established.

Questionnaire and Census Data Integration

The objective is to estimate farmer's needs in Puerto Rico. Two data sets are used for this purpose: (1) The data set from farmer's survey, and (2) the data set from the Census. The data set from farmer's survey has demographic variables and farmer's needs variables. The data set from Census has only demographic needs. The questionnaire will work as a training set for the prediction model. Then and only then, the integration of the questionnaire and Census data set can be performed. The integration process is described next in Table 3.4.

Table 3.4: Census data transformation

Variable	Raw Census data	Transformed Census data
Education	0 - No Education	0 - (No Education + Elementary + Secondary)
	1 - Elementary School	1 - High School
	2 - Secondary School	2 - Certificate or Associate Degree (Some university)
	3 - High School	3 - Bachelor Degree
	4 - Some University	4 - Master or PhD Degree
	5 - Bachelor Degree	
	6 - Master or PhD	
Product category importance (i.e. fruits)	Total Number of farms per product	Importance of product [Product Category, Municipality] = Number of farms [Product, Municipality] / Total Number of farms [Product, Puerto Rico] X Sales [Product, Puerto Rico]
	Total Number of farms	
	Sales per product in P.R.	

After this transformation, one can use the same prediction model for the Census and

the questionnaire data sets. It is important to clarify that the variable "Agriculture: Main economic activity" is not used for the characterization model because it is binary (Yes/No), and one cannot create an equivalent with the Census Data. The variable "Municipality" is also ignored in the characterization model given that it was collected only to ensure the sample representativeness.

Data Preprocessing

To prepare for the clustering techniques, is important to process data beforehand. The clustering algorithm computes distances between observations considering all the variables. Hence, one must pay attention to the variables introduced into the algorithm so the distance is proportionate to the difference in responses (i.e. has an appropriate meaning). Table 3.5 lists and provide explanations of the variables.

Table 3.5: Variable transformation: Explained

Variable	Explanation
Maximum education level	It is a variable with different levels in gradient form, meaning for example that there is more difference between 2 and 4 than between 3 and 4. This variable is kept as-is.
Importance of product category	Once aggregated as explained in Section 3.3.1, the variable is ready to be used in the algorithm
Need	It is a variable with different levels in gradient form; hence, this variable is kept as-is.
Preferred advice channel	This variable should be transformed. The variable is defined in gradient form. So if we do not transform this variable, the clustering algorithm would assume the difference between Telephone Advice (value of 1) and Article Publication (value of 2) is lower than the difference between Telephone Advice (value of 1) and Meetings to Talk about a Specific Topic (value of 5). For solving that, we decided to transform this variable in a vector with 6 variables containing all zeros but a 1 for the selected advice channel.

Another consideration that is needed for this technique is that clustering algorithm

computes distances across observations taking into account all the variables, so we need to be sure each variable has the same importance level. In order to do that, all variable were scaled between 0 and 1 for each variable.

3.3.2 Clustering Analysis

After careful exploration of different clustering techniques alternatives the clustering method selected was hierarchical clustering with Ward's linkage. This clustering method shows good results for small data sets [Kaur and Kaur, 2013], which is case of the test bed. This clustering method groups data based on its intrinsic characteristics, that is, it does not use any prior knowledge about the number of groups to be formed [Berkhin, 2006]. We use as variables both the farmer's attributes and the farmer's needs with the objective of discovering patterns in the data set coming from administered questionnaires. Table 3.6 shows the list of variables extracted from questionnaires.

Ward's method establishes relationships between the observations but does not determine the appropriate number of clusters for the data set. The research team anticipated one needed clustered groups of different sizes (provided the type of data in the sample) and this adds complexity in selecting a "good cutting" mechanism for Ward's method. In order to obtain the number and composition of clusters we will use JAM (Chapter 2). This method proved to improve hierarchical clustering analysis under situations where clusters vary in density and size. Once farmer's groups are established, we identify patterns in each of the clusters formed.

3.3.3 Classification Model

The Agricultural Census (2012) is the only source providing information about the agricultural activity distribution around Puerto Rico. This information only includes what

Table 3.6: List of variables extracted from Farmers Needs Questionnaire

Attributes	Needs
Education level Cows: Aggregated value Other animals: Aggregated value Cereals: Aggregated value Vegetables and melons: Aggregated value Fruits: Aggregated value Tubers: Aggregated value Spices and beverage crops: Aggregated value Legumes: Aggregated value Sugar crops: Aggregated value Other crops: Aggregated value	Nutrition and supplementation Forage Reproduction Animal management and control Diseases prevention and management Water irrigation Soil analysis Plague management Erosion control Fertilization Post-harvest management Crop varieties Business accounting Marketing Sustainable energy Available funds Obtaining certifications and registrations Waste management Continuous education Product processing
Total Attributes 11	Total Needs 20

throughout this document is labeled "attributes" (e.g. importance of product category, education level). This is why we decided to create a classification model predicting farmer's cluster using only farmer's attributes as predictors. We train the classification model with the Farmer's Needs Questionnaire clustering outcomes.

Once one defines the clusters for the questionnaire observations, one can use the cluster number of each questionnaire observation as response and the attributes as predictors. We test three different classification methods, the traditional recursive partitioning tree [Breiman et al., 1984], logistic regression and evolutionary tree [Grubinger et al., 2011]. Finally, we are ready to compare the performance of the three classification methods and choose the best, then we use the model with the Census observations.

3.4 Results

Estimation of farmer's needs for the Census data set

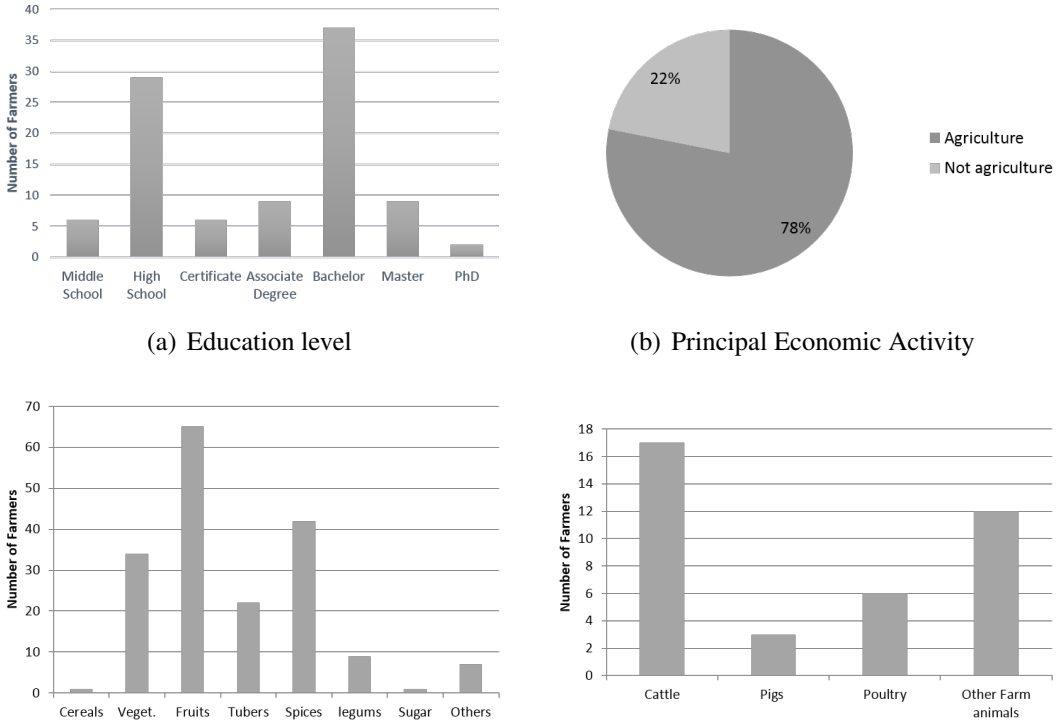
First and foremost, it is important to establish the representativeness of the data collected. A sample profile, explained in Section 3.3.1, was designed so that all "product types" will be well represented in the sample. Table 3.7 shows each product type with corresponding target (i.e. objective) in the sample and the total surveyed. The "result" column shows an upward arrow where the sample surpassed the target and an equal sign where the sample and the target are the same. Observe that the sample composition fulfilled sample requirements.

Table 3.7: Sample product profile

Product	Total surveyed	Objective	Result
Cattle	17	11	↑
Swine	3	2	↑
Poultry	6	6	=
Aquiculture	0	0	=
Other farm animals	12	7	↑
Cereals	1	0	↑
Vegetables and melons	34	8	↑
Fruits	65	37	↑
Oilseeds crops	0	0	=
Tubers	22	6	↑
Spices and beverage crops	42	18	↑
Leguminous crops	9	3	↑
Sugar crops	1	0	↑
Other crops	7	2	↑

Farmer's attributes in the sample were well diverse. Figure 3.2 shows the most relevant attributes. From the figure it is interesting to note that the education level of the participants in the sample was high, specifically 64 out of 99 had a college-level degree. For the

majority of the sample (78%) agriculture was their main economic activity. In terms of crops category, watermelons, fruits and spices were the products most represented and on animal production cattle was dominant.



(a) Education level

(b) Principal Economic Activity

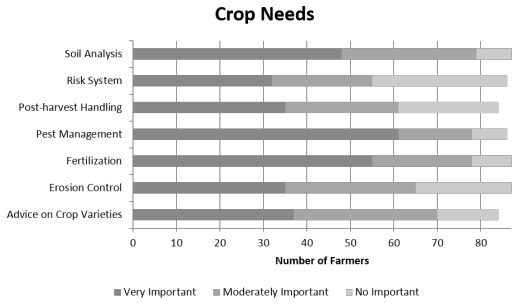
(c) Crop Production

(d) Animal Production

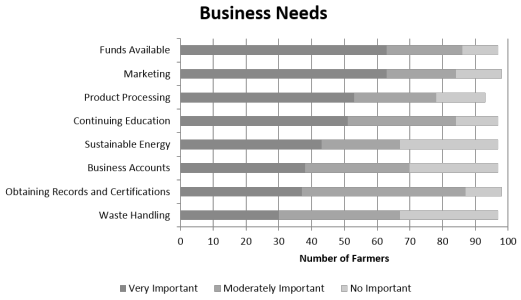
Figure 3.2: Farmers' attributes: Sample Profile

In terms of overall aggregated needs, Figure 3.3 shows the summary of needs for the sample of 99 farmers. The graphs should be interpreted as follows, the darker color indicates the number of farmers that answered a "Very important" for that category. The number that indicated "Moderately important" and "Not important" are shown with a medium contrast and light contrast color, respectively. In the case of crop-related needs, Figure 3.3(a), the most needed services are instruction on pest management, soil analysis and fertilization. For business-related needs, Figure 3.3(b), are funding availability, marketing and product processing. In animal-related needs, Figure 3.3(c), the most needed are nutri-

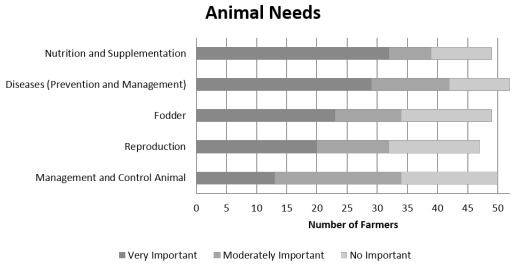
tion and supplementation and disease prevention and control. Interestingly, observe that across needs, the highest ranked are business needs. This challenges the idea that farmer's needs are primarily focused on product-related issues, which was one of the concepts this study is challenging.



(a) Crop-related needs



(b) Business-related needs



(c) Animal-related needs

Figure 3.3: Farmers' needs: Sample Profile

One can predict the cluster of each Census observation using the classification model described in Section 3.3.3. Then, estimate the needs of each Census observation using the average needs of K-nearest neighbors within the predicted cluster. With this methodology we are able to estimate the farmer's needs for the Census data set.

Next, we transform and pre-process the data as explained in Section3.3.1 to able to begin with the clustering analysis.

Clustering analysis

The clustering analysis starts with the construction of Ward's hierarchical clustering. This method gives a dendrogram as output. Then we use JAM, Chapter 2, to define groups out of the Ward's dendrogram. We use Algorithm 3 to define the different grouping options. For the test bed. Table 3.8 shows the different grouping options after applying Algorithm 3 with the corresponding parameters that provided the result, i.e. corresponding Absolute Jump Threshold and Relative Jump Threshold. The algorithm was constructed in R software (R Project for Statistical Computing). Technical notes on the software and computer used are included in Appendix C.

Input : FarmersData /*$Data\ from\ farmer's\ questionnaire$.*/

Output : ClustResult

Require: A ← 0.99,0.98,...,0.90 /*$Sequence\ of\ absolute\ jump\ thresholds$.*/

R ← 0.99,0.98,...,0.90 /*$Sequence\ of\ relative\ jump\ thresholds$.*/

D ← hclust(d=dist(FarmersData),method = "ward.D2") /*$hclust\ defined\ as\ the\ function\ in\ R\ software$.*/;

ClustResult ← ∅;

for *each* $A_i \in A$ **do**

  for *each* $R_j \in R$ **do**

    $ClustResult_{ij}$ ← JAM(A_i, R_j, D);

    /*ClustResult stores all the grouping outputs computed*/;

  end

end

Return ClustResult;

Algorithm 3: Clustering algorithm applied to Farmer's data set.

A desirability function, was defined in order to determine which grouping is the best

Table 3.8: Grouping options for the Farmer's data

Clusters	Absolute Jump Threshold	Relative Jump Threshold
2	0.99	0.98
4	0.98	0.95
6	0.95	0.92
7	0.93	0.97
8	0.93	0.95
10	0.93	0.92
11	0.9	0.94
13	0.92	0.93
15	0.92	0.92
16	0.9	0.93

option. This desirability function takes into account the following performance measures:

SSW: Sum of squares distances within clusters

SSB: Sum of squares distances between clusters

R: Ratio of **SSB** divided by **SSW**

P: Penalty based on the number of clusters

Letting i be the number of clusters in that solution and D_i the desirability of solution with i clusters, the desirability function will be defined by Equation 3.1.

$$D_i = (\min(SSW)/SSW_i + SSB_i/\max(SSB) + R_i/\max(R))/3 - P * (i/\max(i)) \quad (3.1)$$

Note from Equation 3.1 that the smaller the SSW compared to the solution set the more desirable the cluster is. That is the analogous of saying the more "compact" are the groups in that solution. Also, note that the higher the SSB compared to other solutions, the more desirable that particular solution is. This is equivalent of desiring that the groups are far-off each other. There is also a penalty that penalize higher number of clusters. This fact helps with the concept that the higher the number of groups in the cluster, the higher the

probability of small-sized clusters.

Table 3.9 shows the results of the desirability function applied to the case. The highest cluster size tested was 10 given that the sample size in our case is 99 and a cluster size higher than 10 will result in small size groups. Observe from the table that 4 cluster solution provided the highest desirability. It is worth mentioning that changing the penalty from 0.45 to 1, still gives the same solution; hence the size chosen for the rest of the results is 4.

Table 3.9: Performance measures for the grouping options and desirability with penalty=0.5

N-Cluster	SSW	SSB	Ratio	Desirability, D_i
2	352.72	92.97	0.26	0.35
4	305.79	139.9	0.46	0.37
6	291.94	153.75	0.53	0.25
7	291.74	153.95	0.53	0.16
8	274.25	171.44	0.63	0.16
10	260.40	185.29	0.71	0.05

It is interesting and important to note that this grouping result was obtained with a non-horizontal cut to the dendrogram. That is, this grouping solution could have never been obtained with traditional dendrogram cutting methods. This supports choosing **JAM** as dendrogram analysis method for our data set. Next, we show the Results the 4 clusters form.

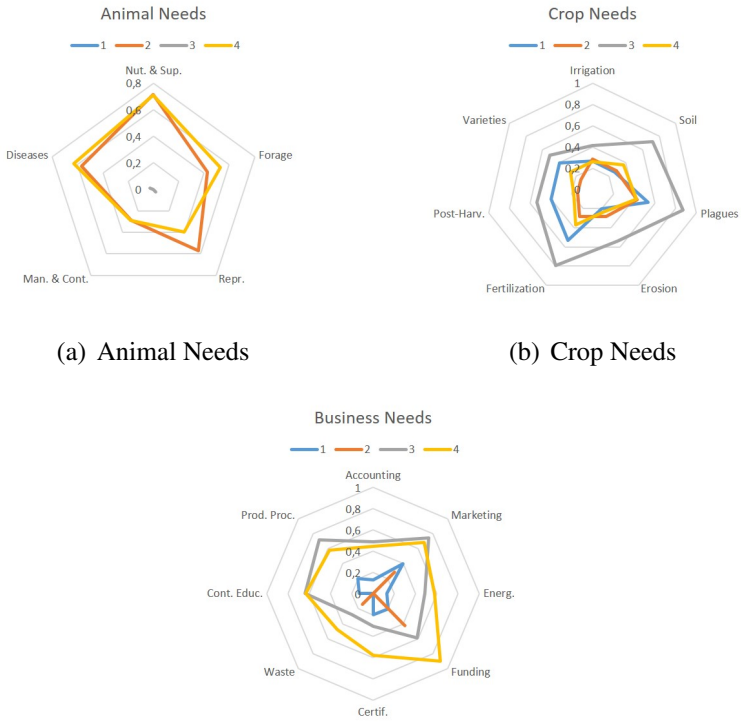
Table 3.10 shows a summary of the demographic variables for each cluster. Observe that Cluster 1 is a medium size cluster (15 observations) comprised of crop farmers with the lowest education degree. Cluster 2 is a really small cluster (7 observations), it is the group with highest educated degree, and they are mainly dedicated to crops. Cluster 3 is a big cluster (39 observations) of crop farmers with average education level (close to the total sample average education level). Cluster 4 is a big cluster (38 observations) with

Table 3.10: Summary of demographic variables of the 4 clusters

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Size	15	7	39	38
Farmers with post-secondary grade	47%	86%	69%	63%
Farmers with animal products	0%	29%	0%	82%
Farmers with crop products	100%	86%	100%	50%

average education level and farmers dedicated to either animal production (50%), crop production (18%) or both (32%).

Figure 3.4 is a radar chart showing the proportion of farmers having each need in each cluster. To construct the radar chart, the only farmer's counted were the ones that answered "very high" to that particular need. In Cluster 1 (blue) farmers do not have animal needs, this make sense because farmers in this group do not have animal production. Their main crop needs are advice in fertilization and plagues followed by advice in crop varieties and post-harvesting practices. They also need advice in marketing. Cluster 2 (red) has farmers with high animal needs and low crop needs, although these farmers are mainly crop farmers. This is the highest educated cluster, so this can be the reason why they want to learn about other farming activities. One could argue that it is because they feel they can learn and then diversify their production. They also have low business needs except for marketing and funding. In Cluster 3 (grey) we have crop farmers with no animal needs, high needs in crop advice, especially fertilization, soil analysis and plagues. For business needs they have high need in marketing, funding, product processing and continuous education. In Cluster 4 (yellow) farmers have high need for advice related to animal production and medium-low need in crop production. They have low business needs except for marketing and funding. Is worth mentioning that this results could be



(a) Animal Needs

(b) Crop Needs

(c) Business Needs

Figure 3.4: Farmer's needs for advice in each cluster

used as starting point to go deep into the study of farmers profiles in Puerto Rico. This results set the ground for the classification model to predict the cluster of an observation given the demographic variables (i.e. attributes).

Classification

In this part, three classification models were studied. The models are used to predict the cluster of an observation, using only demographic variables as predictors. Then we use the best model to allocate census observations in clusters. Next, see the performance measures for these three methods using leave one out, accuracy in Table 3.11 and Confusion matrices in Figure 3.5.

See from Table 3.11 and 3.5 that logistic regression outperform both Recursive parti-

Table 3.11: Accuracy of tested methods

Model	Accuracy
rpart (population priors)	64%
rpart (equal priors)	28%
logistic regression	65%
evolutionary tree	70%

tioning models, because it has more accuracy and it predicts in all the classes. Then, the selection is down to logistic regression and evolutionary tree. Evolutionary tree predicts with more accuracy, 70%, but does not predict any observation in Class 2. On the other hand, logistic regression has 64% accuracy and makes predictions in all classes. We will be using the model to predict the cluster of census observation (municipality), and each census observation will be an average of many farms (typically more than 100 and up to 1,000). Then, it is really difficult that a municipality average has the characteristics of Cluster 2, a very high educated cluster with farmers with high Animal needs and low crop needs, although these farmers are mainly crop farmers. This is why we decide to continue with the evolutionary tree. Evolutionary tree is the one having more accuracy, and it is not really harmful for our test bed that it does not make any Cluster 2 prediction. Next, we will use this model to predict the cluster of each census observation and estimate their needs.

Estimation of Puerto Rican farmer's needs

In order to estimate Puerto Rican Farmer's needs, we use census observations as described in Section 3.3.1 to feed the model chosen in 3.4. Once we predict the cluster of the Census observations we need to predict their needs. We predict the needs of a census observation as a weighted average of the 5 nearest neighbors (individual farmer observations) needs inside the predicted cluster. This way we predict the Puerto Rican Farmer' needs for each

Prediction \ Real Observations	1	2	3	4
1	5	2	4	2
2	0	2	4	4
3	9	2	29	4
4	1	1	2	28

(a) Reg. Log.

Prediction \ Real Observations	1	2	3	4
1	5	0	3	0
2	0	0	0	0
3	7	5	32	6
4	3	2	4	32

(b) Evolutionary tree

Prediction \ Real Observations	1	2	3	4
1	4	1	16	7
2	3	0	16	13
3	8	3	7	1
4	0	3	0	17

(c) Rpart with equal priors

Prediction \ Real Observations	1	2	3	4
1	0	0	5	0
2	0	0	0	0
3	12	5	31	6
4	3	2	3	32

(d) Rpart with population priors

Figure 3.5: Confusion matrixes for classification methods.

municipality.

First we use the evolutionary tree (classification model chosen in Section 3.4) to predict the clusters of each census observation. Next we show the results of this classification in Table 3.12. These results show that all the census observations are located in clusters 3 or 4. Since clusters 3 and 4 are the bigger ones, the sum of both clusters contains 77 farmers, this is the 78 % of our sample. Then, it is normal that census observations, being an average of 100 to 1,000 farmers, have a pattern similar to these clusters. Additionally, cluster 2 is an extreme case of highly educated farmers, so we could predict that an average of farmers in a municipality (census observations) will not follow that pattern. We checked the rpart (population priors) model and the logistic regression, these two models made similar predictions, locating all the census observations also in clusters 3 or 4. So, all these reasons validate the results in Table 3.12. Next, we predict the needs of the census observations.

We calculate the needs of a municipality as a weighted average of the 5 nearest farmers inside the cluster predicted for the municipality. We use as weight the inverse of the distance between the census and questionnaire observations. The result are presented in table form in Table 3.13 and Table 3.14. Table 3.13 shows farmers' need score prediction (column) for each municipality (row) from 1 to 39 and Table 3.14 for each municipality from 40 to 78. The higher the score in these tables the greater the necesity for advise for that particular need in that particular municipality.

To validate these results we analyzed the average of the predicted needs across municipalities, which is presented in Table 3.15. The highest animal need is Nutrition and supplementation, and the lowest is Animal management and control. The highest crop needs are Pests management followed by Fertilization, and the lowest is Irrigation systems. The highest business needs are Funds available followed by Marketing, and the lowest is Waste management. All these patterns are repeated in the average needs of questionnaire observations, supporting the approach taken.

3.4.1 Conclusions

One the main purposes Puerto Rico's Agricultural Extension Service is giving instructional service to farmer's around the island. In this chapter we developed and implemented a methodology to estimate farmer's needs for this instructional service. These needs were estimated at a municipality level.

Since, knowledge of services needed by the clients was not readily available. The first step of this study was to deliver a farmer's needs questionnaire to more than 100 farmers. Then, we characterize farmer's profile using Ward's hierarchical clustering [Jr., 1963] and JAM (method described in chapter 2). Using clustering technique to characterize clientele or customers is an approach used in previous studies [Govind et al., 2008, Shih and Liu,

2003]. After characterization, we used classification models to predict which would be the cluster of each Puerto Rican municipality using the Puerto Rico Agricultural Census 2012. This classification models where trained with the data from questionnaire and then used for prediction with municipalities data. Finally, we predicted needs using a weighted average of 5-nearest neighbours inside the predicted cluster.

Results from this chapter on farmer's needs by municipality were verified by experienced agriculture specialists that have direct contact with Puerto Rican farmers. Some of these specialists are active members of Puerto Rico Agricultural Extension service. This chapter most important contribution is the development of a model that uses farmer-level needs characterization combined with municipal proles to estimate municipality-level needs. This methodology could be replicated in similar cases to predict regional needs using individual observations when few data is available.

Table 3.12: Predicted cluster for each municipality

Municipality	**Cluster**	**Municipality**	**Cluster**
Adjuntas	3	Juncos	4
Aguada	4	Lajas	4
Aguadilla	4	Lares	3
Aguas Buenas	4	Las Marías	3
Aibonito	4	Las Piedras	4
Aasco	3	Loíza	4
Arecibo	4	Luquillo	4
Arroyo	4	Manatí	4
Barceloneta	4	Maricao	3
Barranquitas	4	Maunabo	4
Bayamón	4	Mayagüez	3
Cabo Rojo	4	Moca	4
Caguas	4	Morovis	4
Camuy	4	Naguabo	4
Canóvanas	4	Naranjito	4
Carolina	4	Orocovis	4
Cataño	4	Patillas	4
Cayey	4	Peñuelas	4
Ceiba	4	Ponce	3
Ciales	4	Quebradillas	4
Cidra	4	Rincón	4
Coamo	4	Río Grande	4
Comerío	4	Sabana Grande	4
Corozal	4	Salinas	4
Culebra	4	San Germán	4
Dorado	4	San Juan	3
Fajardo	4	San Lorenzo	4
Florida	4	San Sebastián	4
Guánica	4	Santa Isabel	4
Guayama	3	Toa Alta	4
Guayanilla	4	Toa Baja	4
Guaynabo	4	Trujillo Alto	4
Gurabo	4	Utuado	3
Hatillo	4	Vega Alta	4
Hormigueros	4	Vega Baja	4
Humacao	4	Vieques	4
Isabela	4	Villalba	4
Jayuya	3	Yabucoa	4
Juana Díaz	4	Yauco	3

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16	N17	N18	N19	N20
M1	0.0	0.0	0.0	0.0	0.0	2.0	2.6	3.0	2.2	2.6	2.0	2.6	2.2	2.6	1.8	2.6	2.0	1.8	2.4	2.8
M2	3.0	2.8	2.4	2.4	2.4	2.2	2.4	2.8	2.0	2.8	2.2	2.0	2.6	3.0	2.4	3.0	2.8	2.0	2.6	3.0
M3	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M4	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M5	2.8	2.8	2.4	2.4	2.4	1.6	2.0	2.4	1.8	2.4	1.8	1.6	2.4	3.0	2.2	3.0	2.8	1.8	2.4	2.8
M6	0.2	0.2	0.6	0.2	0.6	2.4	2.8	3.0	2.6	2.6	2.6	2.6	1.8	2.2	2.6	2.8	2.4	1.8	2.4	2.6
M7	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M8	2.6	2.0	2.0	1.8	2.0	1.8	2.0	2.0	1.2	1.8	1.4	2.0	2.2	2.8	2.6	2.8	2.8	2.2	2.2	3.0
M9	3.0	2.6	2.8	2.6	2.8	1.4	1.6	1.8	1.2	1.8	1.4	1.2	2.6	3.0	1.8	3.0	3.0	2.2	2.8	2.8
M10	2.8	2.8	2.2	2.2	2.2	2.0	2.4	2.8	2.0	2.8	2.2	2.0	2.4	3.0	2.4	3.0	2.8	1.8	2.4	2.8
M11	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M12	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M13	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M14	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M15	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M16	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M17	2.8	2.8	2.0	1.7	2.3	0.3	1.0	1.0	0.3	1.0	0.3	0.3	1.7	2.5	2.0	2.8	2.2	1.5	2.8	2.2
M18	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M19	2.6	2.3	2.4	2.0	1.8	1.9	2.4	2.4	1.2	2.1	1.4	2.1	2.0	2.8	2.3	2.8	2.8	1.9	2.1	3.0
M20	3.0	2.8	2.4	2.4	2.2	2.0	2.4	2.8	1.8	2.8	2.0	2.0	2.4	3.0	2.2	2.8	2.8	2.0	2.4	3.0
M21	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M22	2.8	2.8	2.4	2.4	2.4	1.6	2.0	2.4	1.8	2.4	1.8	1.6	2.4	3.0	2.2	3.0	2.8	1.8	2.4	2.8
M23	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M24	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M25	2.8	2.6	2.4	2.2	2.6	1.4	1.7	1.9	0.8	1.7	0.6	0.8	1.8	2.4	1.8	3.0	2.6	1.8	2.8	2.4
M26	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M27	2.6	2.4	2.4	2.0	1.8	1.8	2.4	2.4	1.2	2.2	1.4	2.0	2.0	2.8	2.2	2.8	2.8	1.8	2.2	3.0
M28	2.6	2.4	2.4	2.0	1.8	1.8	2.4	2.4	1.2	2.2	1.4	2.0	2.0	2.8	2.2	2.8	2.8	1.8	2.2	3.0
M29	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M30	0.2	0.2	0.6	0.2	0.6	2.4	2.8	3.0	2.6	2.6	2.6	2.6	1.8	2.2	2.6	2.8	2.4	1.8	2.4	2.6
M31	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M32	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M33	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M34	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M35	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M36	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M37	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M38	0.2	0.2	0.6	0.2	0.6	2.2	2.6	3.0	2.4	2.6	2.4	2.6	1.6	2.2	2.2	2.2	2.2	1.6	2.4	2.8
M39	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8

Table 3.13: Predicted farmer's needs in municipalities 1 to 39. The notation key is in Appendix D, tables D.1 and D.2.

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16	N17	N18	N19	N20
M40	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M41	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M42	0.0	0.0	0.0	0.0	0.0	2.0	2.6	3.0	2.4	2.6	2.4	2.6	2.0	2.6	1.8	2.2	2.0	1.8	2.4	2.8
M43	0.0	0.0	0.0	0.0	0.0	2.0	2.4	3.0	2.4	2.8	2.4	2.6	2.4	3.0	1.5	2.2	2.0	1.8	2.6	2.8
M44	3.0	2.8	2.4	2.4	2.4	2.2	2.4	2.8	2.0	2.8	2.2	2.0	2.6	3.0	2.4	3.0	2.8	2.0	2.6	3.0
M45	2.6	2.4	2.4	2.0	1.8	1.8	2.4	2.4	1.2	2.2	1.4	2.0	2.0	2.8	2.2	2.8	2.8	1.8	2.2	3.0
M46	2.8	2.8	2.4	2.4	2.4	1.6	2.0	2.4	1.8	2.4	1.8	1.6	2.4	3.0	2.2	3.0	2.8	1.8	2.4	2.8
M47	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M48	0.0	0.0	0.0	0.0	0.0	2.0	2.6	3.0	2.4	2.6	2.4	2.6	2.0	2.6	1.8	2.2	2.0	1.8	2.4	2.8
M49	3.0	2.8	2.4	2.4	2.2	2.0	2.4	2.8	1.8	2.8	2.0	2.0	2.4	3.0	2.2	2.8	2.8	2.0	2.4	3.0
M50	0.2	0.2	0.6	0.2	0.6	2.4	2.8	3.0	2.6	2.6	2.6	2.6	1.8	2.2	2.6	2.8	2.4	1.8	2.4	2.6
M51	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M52	2.6	2.4	2.2	2.0	1.8	2.0	2.2	2.4	1.6	2.2	1.6	2.2	2.0	2.8	2.6	2.8	2.6	2.0	2.0	3.0
M53	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M54	3.0	2.8	2.4	2.4	2.4	2.2	2.4	2.8	2.0	2.8	2.2	2.0	2.6	3.0	2.4	3.0	2.8	2.0	2.6	3.0
M55	3.0	2.8	2.4	2.4	2.2	2.0	2.4	2.8	1.8	2.8	2.0	2.0	2.4	3.0	2.2	2.8	2.8	2.0	2.4	3.0
M56	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M57	3.0	2.8	2.4	2.4	2.2	2.0	2.4	2.8	1.8	2.8	2.0	2.0	2.4	3.0	2.2	2.8	2.8	2.0	2.4	3.0
M58	0.2	0.2	0.6	0.2	0.6	2.0	2.6	3.0	2.6	2.6	2.6	2.6	1.8	2.2	2.4	2.4	2.2	1.6	2.6	2.8
M59	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M60	2.6	2.2	2.4	2.2	2.0	2.0	2.4	2.6	1.4	2.4	1.6	2.0	2.0	2.8	2.0	2.8	2.8	1.8	2.0	3.0
M61	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M62	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M63	2.8	2.8	2.4	2.4	2.4	1.6	2.0	2.4	1.8	2.4	1.8	1.6	2.4	3.0	2.2	3.0	2.8	1.8	2.4	2.8
M64	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M65	0.2	0.2	0.6	0.2	0.6	2.4	2.8	3.0	2.6	2.6	2.6	2.6	1.8	2.2	2.6	2.8	2.4	1.8	2.4	2.6
M66	3.0	2.8	2.4	2.4	2.2	2.0	2.4	2.8	1.8	2.8	2.0	2.0	2.4	3.0	2.2	2.8	2.8	2.0	2.4	3.0
M67	3.0	2.8	2.6	2.6	2.6	1.8	2.0	2.4	1.8	2.4	1.8	1.6	2.6	3.0	2.2	3.0	2.8	2.0	2.6	3.0
M68	2.8	2.8	2.4	2.4	2.4	1.6	2.0	2.4	1.8	2.4	1.8	1.6	2.4	3.0	2.2	3.0	2.8	1.8	2.4	2.8
M69	2.6	2.4	2.4	2.0	1.8	1.8	2.4	2.4	1.2	2.2	1.4	2.0	2.0	2.8	2.2	2.8	2.8	1.8	2.2	3.0
M70	3.0	2.6	2.6	2.4	2.6	1.2	1.4	1.8	1.2	1.8	1.2	1.0	2.4	3.0	1.8	3.0	2.8	2.0	2.6	2.8
M71	2.6	1.8	2.0	1.8	2.0	1.4	1.6	1.6	1.0	1.4	1.0	1.6	2.0	2.8	2.6	2.8	2.8	2.4	2.0	2.6
M72	0.0	0.0	0.0	0.0	0.0	2.0	2.6	3.0	2.6	2.6	2.6	2.6	2.0	2.6	2.2	2.4	2.2	1.8	2.4	2.6
M73	2.6	2.4	2.4	2.0	1.8	1.8	2.4	2.4	1.2	2.2	1.4	2.0	2.0	2.8	2.2	2.8	2.8	1.8	2.2	3.0
M74	2.8	2.8	2.4	2.4	2.4	1.6	2.0	2.4	1.8	2.4	1.8	1.6	2.4	3.0	2.2	3.0	2.8	1.8	2.4	2.8
M75	2.6	2.2	2.4	2.0	1.9	2.1	2.2	2.4	1.4	2.0	1.4	2.2	2.1	2.8	2.6	2.8	2.6	2.1	2.0	2.7
M76	2.8	2.8	2.4	2.4	2.4	2.2	2.6	3.0	2.4	3.0	2.4	2.2	2.4	3.0	2.4	3.0	2.8	2.0	2.4	2.8
M77	3.0	2.8	2.4	2.4	2.2	2.0	2.4	2.8	1.8	2.8	2.0	2.0	2.4	3.0	2.2	2.8	2.8	2.0	2.4	3.0
M78	0.2	0.2	0.6	0.2	0.6	2.0	2.6	3.0	2.6	2.6	2.6	2.6	1.8	2.2	2.4	2.4	2.2	1.6	2.6	2.8

Table 3.14: Predicted farmer's needs in municipalities 40 to 78. The notation key is in Appendix D, tables D.1 and D.2.

Table 3.15: Average of the predicted needs across municipalities. From 0 (No Need) to 3 (High Need)

Need definition	**Need value**
Funds available	2.87
Marketing	2.86
Product processing	2.85
Obtaining certifications	2.69
Pests management	2.49
Nutrition and supplementation	2.45
Continuous education	2.45
Fertilization	2.40
Accounting	2.30
Forage	2.28
Sustainable energy	2.18
Soil Analysis	2.16
Reproduction	2.13
Diseases prevention and management	2.06
Animal management and control	2.01
Waste management	1.93
Post harvest management	1.84
Crop varieties	1.82
Irrigation systems	1.81
Erosion control	1.79

Chapter 4

Assignment Model

4.1 Introduction

This chapter presents an assignment model that provides a solution for common challenges faced by service organizations. In service organization one may encounter a mismatch between resources capabilities and task requirements, especially when organizations offer a wide variety of services. If a traditional assignment model is used in these scenarios, it will often result in a service organization assigning employees (or other resources) to tasks they will not fulfill to a desired 'utility' level. For example, a service organization may want to assign a resource only if the resource is going to perform accordingly to their standards, otherwise this assignment will harm the companys reputation. Additionally, the organization may want to complete a task only if completed according to those same standards. If a traditional assignment model is used on this scenario, the solution may suggest assigning a resource to a task that will be detrimental to the image of the organization. The assignment model proposed in this work will consider the resource-task fit and may result in a solution that suggests not fulfilling all tasks or not assigning a resource at all.

Our work tailors the assignment model to account for common challenges faced by

service organizations. The common challenges faced do not only include the detrimental benefit of fulfilling tasks with unfit resources, but also include an unknown relationship between demand (or tasks) and workload, a myriad of tasks defined by their required skills, and a diverse resource or personnel profile. To the best of our knowledge, no solution has been proposed in literature for assignment problems taking into account all of these challenges. Solutions for assignment problems with combinations of personnel profiles and tasks diversity has been addressed before by Korkmaz et al. [2008], Peters and Zelewski [2007], Wongwien and Nanthavanij [2013]. The closest work to account for this is Caron et al. [1999] that shows that a particular formulation of the assignment problem can lead to both idle workers and unassigned tasks. However, there is no a real application of the assignment problem with a solution proposing the possibility of both idle workers and unassigned tasks simultaneously.

The methodology described in this chapter uses as venue the assignment of Puerto Rico AES's agents to ease its exposition. This work uses input data from a previous study presented in Chapter 3. This input data defines the tasks needed to be performed by the agents; this data is analogous to a study of the demand.

4.2 Related Literature

Extensive literature on "allocation" problems exists under the name of assignment problem (AP). Numerous variations of the AP have been adapted in various fields for over 50 years. An extensive literature review can be found in Pentico [2007], starting from the classical formulation of the problem, which is a forced one-to-one matching minimizing the sum of assignment costs. The review continues exposing the main variations to the problem that appeared like bottleneck, quadratic, etc. Öncan [2007] provides a literature review of the generalized assignment problem (GAP), which is a particular case of the

allocation problem that focuses on cases that include assigning objects to places or agents to tasks. First, they explain the classical formulation of the GAP and expose the main variations of the problem found across time. The primary focus of this review is to explain the main fields of application in which this problem has been applied from scheduling and transportation to telecommunication applications. Niknafs et al. [2013] is a survey focused on personnel assignment problem (PAP). PAP studies the assignment of agents to tasks, which is a particular case of the GAP. They explain how the research in this review was conducted, and then they show the results. Interestingly, this review shows a table with a brief description of problems they found and a table showing the validation approaches these studies conducted, from small examples to random test cases, case studies, experiments and real-world problems. Furthermore, they summarized some work done in multilevel-AP and multidimensional-AP, mentioning multilevel-AP has been used in a few studies and redirecting to some reviews for the ones interested in the multidimensional-AP.

There is work addressing how to take into account the combination of personnel profiles and task diversity. Korkmaz et al. [2008], Peters and Zelewski [2007], Wongwien and Nanthavanij [2013], Elango et al. [2011], Fernandez-Viagas and Framinan [2014] developed a solution for different situations with combination of personnel profiles. Korkmaz et al. [2008], Peters and Zelewski [2007] solve an assignment problem of employees to jobs considering workers skills and preferences. Both are one-to-one forced matching, which implies they do not allow a solution with both idle workers and unassigned jobs/tasks simultaneously. Wongwien and Nanthavanij [2013] proposes a model for workforce scheduling problem considering employees competences, preferences and controlling workers exposure to harmful sources. This model forces all the tasks to be completed. Fernandez-Viagas and Framinan [2014] solve a project scheduling problem considering personnel skills, and the assignment is also forced since all the tasks must be completed.

The particular features that the model in this thesis takes into account include the non-convenience of assigning tasks to unfit personnel, the myriad of tasks defined by their required skills and a diverse resource or personnel profile. To the best of our knowledge, there is no study in literature addressing a solution to an assignment problem with unknown relation between demand (or tasks) and workload.

Regarding the feature of not forcing assignment, we have not found a real application of proposed models that does not force either all the tasks to be completed or the workers to not have idle time. For example, Caron et al. [1999] proposes a model to solve an assignment problem with seniority constraints, which means that higher seniority workers have priority in the task assignment. It is a theoretical paper and the assignment is forced looking for a full employment solution, creating slack jobs (non-real jobs) if necessary. This paper briefly suggests how the classical constraints of an assignment problem could be modified to not force the assignment, allowing idle workers and unassigned tasks simultaneously. However, no work is conducted in this direction later in such paper. Caron et al. [1999] work considers a combination of personnel profiles and task diversity. It can be seen as an unforced assignment, since it is forced creating fictitious jobs in order to meet the seniority constraints. Our approach to not force the assignment is different since we will not be creating fictitious jobs; instead, we will relax both the classical full employment constraint approach allowing idle workers, and the classical constraint that forces all the jobs to be completed.

To the best of our knowledge the situation where the relationship between demand (or tasks) and workload is unknown has not been solved anywhere. In all works found the relationship is known and is an input to the assignment model. Our approach to the unknown relationship is to balance the agents workload. Balancing the workload has been proposed before. Elango et al. [2011] studies an assignment problem in a multi-robot

system. They first apply clustering to the task using k-means, grouping the tasks based on the physical distance between them. Their approach to balance the workload is to include an idle cost in the objective function. This is not applicable in our case since we will not know the workers' capacity, thus we will not be able to calculate an idle cost. However the concept of balancing the workload using clustering as a first step is interesting for our study. The clustering of tasks in our case will be different. The diversity of knowledge required to complete the tasks implies each agent is going to be able to meet only some of the needs; hence we cannot group tasks taking only into account distance between them, as doing so, would not allow to assign the group to an agent due to the different characteristics of the clients needs in each group. Tasks diversity is taken into account in this study. There is no combination of personnel profiles and the assignment is forced since all the tasks must be completed.

In this thesis, tasks are grouped before modeling the assignment problem. Grouping items, clients, or tasks before solving an assignment problem is an approach found often in the literature [Campbell and Savelsbergh, 2004, Fisher and Jaikumar, 1981, Govind et al., 2008, Yi-Fei Chuang, 2014]. Govind et al. [2008] studies the assignment of bed-days to different types of diseases in the hospital network of Greater Los Angeles area. First, they estimate what will be the disease incidence in each zip code of Greater Los Angeles and, then, they assign the patients to hospitals taking into account each hospital capacity and maximizing sum of patients utility in terms of driving time spent arriving to the hospital (less time, more utility for the patient). Diversity of tasks is considered, but no combination of personnel profiles is taken into account since patients are assigned to hospitals and every hospital can accept any patient regardless what is the particular patients disease. The assignment is forced since the model assigns all patients to hospitals.

Govind et al. [2008], Shih and Liu [2003] propose customer grouping. Grouped in-

formation is useful to allocate company's resources. These cases serve as examples to the usefulness of grouping clients because easily understandable information is obtained. Yi-Fei Chuang [2014] groups interrelated items before allocating them in warehouses. Thus, grouping items, tasks, or clients before solving an assignment problem is an approach that has been used successfully before. We will execute the idea of grouping demand but it will be done as a pre-work to understand demand characteristics for later demand estimation. That is, the assignment model will not be assigning resources to groups in the preliminary work but, will be assigning to the demand estimation resulting from a previous study.

In summary, there is no study addressing solution to an assignment problem with unknown relation between demand and workload. Not forcing the assignment is a consideration only briefly mentioned in the literature at theoretical level and using the objective function with negative weights. However, we will be implementing this feature in the constraints, relaxing the traditional assignment constraint of assigning all resources, not having idle workers and adding constraints to ensure that every assigned task is performed at the desired level. The other two characteristics, combination of personnel profiles and tasks diversity are features found in the literature but not in combination with the first two characteristics.

4.3 Methodology

This methodology gives a solution to resource-task allocation where the objective is to maximize the utility (service output) of the agent to task assignment. The utility includes clienteles' demand, agents' preferences, and agents' expertise. The scenario portrayed is one where there is a mismatch between demand-resource characteristics, and the demand results from an estimation using knowledge from clientele needs in combination with estimated population quantity of specific clientele as exposed in Chapter 3.

The assignment model uses as inputs the characteristics of the demand that were estimated in Chapter 3, and two descriptive variables of the resources, i.e. preference, expertise. We also considered the agents' capacity and how it is affected by the distance between regions assigned to the same agent. The output of the assignment model will be the assignment of resources (agents) to needs in regions/municipalities. In the next sections, we describe the inputs and output of the assignment model.

4.3.1 Municipality Needs Input

We used the result of farmers' needs estimation resulting from Chapter 3 as input for the assignment model. This input will have the format shown in Tables 3.13 and 3.14. The input of farmers' needs per municipality is a level of need i.e. 0 for no need, 1 for low need, 2 for medium need, and 3 for high need, and the quantity of farmers in the corresponding municipality.

4.3.2 Agent's Profile

The resource profile is defined by two descriptive variables. For our test bed, these variables will be:

Expertise: Agent's fit to satisfy a need in terms of technical background, skills, and experience.

Preference: Agent's preference towards working in each programmatic area (refer to the programmatic area description in Chapter 3).

It is important to note that, if we use real data for the variables listed before, we could not guarantee that agents would not be affected in their jobs because of the sensitivity of the information. Hence, for both variables artificial data was generated in order to be

able to test the assignment model and demonstrate potential scenarios, without losing the practical application of the work. Details on this artificial data is presented in Section 4.4.1.

4.3.3 Managerial Decisions

In order to adjust the model to the real-world application, we need to make some decisions about how the model considers the following:

- Penalty to capacity due to agent travel: If an agent is assigned to different municipalities, the model gives a penalty to agent's capacity based on the distance between municipalities.

- Municipality needs: During data collection, we noticed farmers answered "2-medium need" for those needs they do not want advice on but they consider important or they think they could search for advice in the future. Hence, the model will not assign an agent to that level of need. The model will assign 0 to any need bellow 2. Then, assignning an agent to a medium-low need will have no impact in the global utility. In consequence there will bw no incentive for the model to assign an agent to a medium-low need.

4.3.4 Model Output

The model output is an assignment of agents to advice specific needs in municipalities based on the input information and the formulation we developed. Thus, if I have an expert in erosion control the model will intend to assign this agent to the municipalities that being close to each other has high need in erosion. Next, we explain the details of the assignment model.

4.3.5 Unified Model Formulation

This sub-section describes the first optimization model tested. This formulation was executed in MATLAB from MathWorks [1]. For technical information about the software refer to Appendix C. The model built only worked with small sizes of our problem (a maximum of 12 municipalities, 8 agents and 20 needs) due to computational complexity. Program was run in a Dell computer, specification are also in Appendix C. An alternative model was built which could be solved with the resources on-hand. Section 4.3.5 describes the alternative model which executed successfully in MATLAB for the full size problem (78 municipalities, 50 agents and 20 needs).

Details on the assignment model formulation:

The objective of the assignment or optimization model is to find the allocation of resources (agents) to tasks (needs) that maximizes the utility of the service provided. Utility is be defined by a multi-criteria function that includes agents' preferences, location and performance for each task, and also the economic need of the clients. Equation 4.1 provides a generalization of the utility function. Let

$$\max Utility = \sum_{ijk} (X_{ijk} * P_{jk} * Pref_{jk} * F_{i} * N_{ik}) \tag{4.1}$$

where

X_{ijk} is 1 if agent j is assigned to need k in municipality i, and 0 otherwise,

P_{jk} is the performance score of agent j advising need k,

$Pref_{jk}$ is the preference score of agent j advising need k,

[1] MATLAB is a language of technical computing to solve engineering and scientific problems.

F_{i} is the number of farmers in municipality i, and

N_{ik} is the level of need k in tmunicipality i.

Note from Equation 4.1 that the utility is defined by a product of four parameters (P_{jk}, $Pref_{\mathrm{jk}}$, F_{i}, N_{ik}) and the model's variable X_{ijk}. This is something similar to what Nash Jr [1950] proposed in game theory field as solution for the bargaining problem. Here we are looking to assign agents to tasks with a balance of these parameters better than maximizing the sum of the scores which can lead to extremes (i.e. high performance and low preference).

Service Quality & Preferences

As motivated before, the objective is to assign agents to tasks only if we know they are going to perform according to the company's standards. In order to achieve this goal, the model takes into account the agent's knowledge/experience and preferences. The assumption is that assigning agents to tasks the are knowledgeable and also prefer, will lead agents to perform better because they will have tasks related to their background, and they will also be motivated because we are considering their preferences. The model includes a minimum for the knowledge and preference criteria. This minimum is used to assure agents are performing at the minimum desired level by the organization. This parameter can be substituted by any other performance evaluation considered. For the test bed, we are not evaluating the agents due to the sensitivity of the information. However, the model allows for an expertise metric as a input-value ranging from 0 to 1 for each agent advising each need, where 0 means the lowest performance possible and 1 means the highest. The case of agent's preference is analogous to their expertise where the same scale is used.

The model allows for a minimum to be set for both, expertise and preference. In the case of the test bed the minimum for both metrics were set to 0.5. Initially, the model

included these considerations as two separate restrictions adding complexity to the model. Later these restrictions were dropped and added a step into the data processing to reduce the model's computational complexity. The preprocessing consists in assigning a value of 0 to each expertise and preference values lower than 0.5 for every agent. This will result in no utility for assigning agents to a low preference or low expertise task given that the model does not have to assign agents to a task that will not contribute to the utility function.

Model Restrictions

Restriction to avoid redundant assignment

In this assignment problem we have to ensure non-redundancy in the tasks allocation to agents. Equation 4.2 ensures that if an agent j is assigned to a task k in municipality i, no other agent j is assigned to the same task k in the same municipality i.

$$\sum_{j}(X_{\mathrm{ijk}}) \leq 1 \qquad \forall i,k \tag{4.2}$$

Capacity & travel for agents

The traditional personnel assignment problem is a one-to-one matching of agents to job positions. When assigning tasks, agents typically have a defined capacity and each of the tasks consumes part of this capacity. In our case, it would be excessively time consuming (impractical) to define the exact capacity that each task consumes since there is no real information or data in this direction. Furthermore, it will also be hard, if not impossible, to define the exact quantity of tasks that are needed to be performed since there is no record of tasks requested or performed in the past; thus, we need another approach.

This thesis proposes balancing the workload. That is, if we have 3 agents to provide service to 180 farmers, the model looks for a solution in which each agent takes the re-

sponsibility of giving advice to approximately 60 farmers. Now, given that forcing the assignment to exactly 60 farmers will limit the feasible space of our problem, here we explored two options. One option is to include a negative weight in the objective function for overloaded workers .The second option, which is the option chosen, adds a percentage of allowed overload (added 10% of allowed overload for the test bed). Using the same aforementioned example, this option will not allow any agent to give advice to more than 66 (60+0.1*60=66) farmers.

Another consideration for agent's capacity is that an agent giving advice to needs in distant locations will have his capacity reduced due to travel. We calculate this travel establishing one region (municipality) as base region for each agent, and giving a penalty due to travel from this region to other regions where the agent is advising. This penalty is directly proportional to how far the task is from the base region. The penalty we defined is also directly proportional to the quantity of farmers with the need assigned to the farmer in the specific region. The formulation designed for this purpose is provided in Equation 4.3.

$$\sum_{i,k} (F_{\mathrm{i}} * N_{\mathrm{ik}} * X_{\mathrm{ijk}} * (1 + dist_{\mathrm{ji}}/100km)) \leq Capacity_{\mathrm{j}} \qquad \forall j \tag{4.3}$$

were

$dist_{\mathrm{ji}}$ is the distance of agent j to municipality i in kilometers, and

$Capacity_{\mathrm{j}}$ is the maximum capacity allowed to be assigned to agent j.

Note that this restriction, Equation 4.3 is nonlinear, therefore, it increases the computational complexity of our problem. Hence, we decided to create a linear restriction to achieve the same purpose. The result is in Equation 4.4.

$$\sum_{j} (F_{\mathrm{i}} * N_{\mathrm{ik}} * (X_{\mathrm{ijk}} + M_{\mathrm{aj}} + Aux_{\mathrm{ijka}}) * (1 + dist_{\mathrm{ji}}/100km)) \leq Capacity_{\mathrm{j}} \qquad \forall j \tag{4.4}$$

were

M_{aj} is 1 if agent j has region a as base region and 0 otherwise, and

Aux_{ijka} is an auxiliary variable with no real meaning other than modeling purposes. The value of this variable can be either -1 or 0.

Note that $dist_{ji}$ from Equation 4.3 is exchange for $dist_{ai}$ in Equation 4.4. $dist_{ji}$ was used in Equation 4.3 to define distance between agent's j base region and i, so this is a variable. Instead, in Equation 4.4 we used $dist_{ai}$ to define distance between region a and region i, so $dist_{ai}$ is a parameter and not a variable, therefore, Equation 4.4 is a linear restriction.

Equation 4.4 ensures agent j does not surpass his/her capacity with the assignment of tasks (right part of the equation). It also ensures that tasks in region i consume more capacity from the agent j if they are further from region a. (The assignment variables (X_{ijk}, M_{aj} and Aux_{ijka}) are multiplied by (1+ $dist_{ai}$ / (100km)).) Capacity consumed by the assignment of a need is also directly proportional to the number of farmers in the region of the assigned task (F_{ik}). Next, we explain how we control the assignment with the assignment variables (X_{ijk}, M_{aj} and Aux_{ijka}.

We want the three assignment variables (X_{ijk}, M_{aj} and Aux_{ijka}) to sum 1 when an agent j is assigned to need (task) k in region i using a as base region and 0 otherwise. This way we ensure that only the assigned tasks subtract capacity from an agent. In order to ensure this control we add the restriction described in Equation 4.5.

$$X_{ijk} + M_{aj} + Aux_{ijka} \geq 0 \qquad \forall i, j, k, a \tag{4.5}$$

With Equation 4.5 we ensure Aux_{ijka} is 0 if X_{ijk} is 0 and M_{aj} is 0. Aux_{ijka} will be -1 otherwise because the optimization process will make the value of Aux_{ijka} as low as possible in order to free capacity from the agents in Equation 4.4.

Since we want the three assignment variables (X_{ijk}, M_{aj} and Aux_{ijka}) to sum 1 when an agent j is assigned to need (task) k in region i using a as base region and 0 otherwise. We have the following possible situations for the values of assignment variables (X_{ijk}, M_{aj} and Aux_{ijka}):

Case 1 $X_{\mathrm{ijk}} = 1$ & $M_{\mathrm{aj}} = 1$; This means agent j is assigned to need k in region i using a as base region. In this case, Aux_{ijka} will have a value of -1. The sum of the three variables will be 1, and the capacity restriction 4.4 will work properly.

Case 2 $X_{\mathrm{ijk}} = 1$ & $M_{\mathrm{aj}} = 0$; This means agent j is assigned to need k in region i but using as base region other than a. In this case, Aux_{ijka} will have a value of -1. The sum of the three variables will be 0 and the capacity restriction 4.4 will work properly.

Case 3 $X_{\mathrm{ijk}} = 0$ & $M_{\mathrm{aj}} = 1$; This means agent j is not assigned to need k in region i but uses a as base region. In this case, Aux_{ijka} will have a value of -1. The sum of the three variables will be 0 and the capacity restriction 4.4 will work properly.

Case 4 $X_{\mathrm{ijk}} = 0$ & $M_{\mathrm{aj}} = 0$; This means agent j is not assigned to need k in region i and a is not his base region. In this case, Aux_{ijka} will have a value of 0 due to restriction 4.5. The sum of the three variables will be 0 and the capacity restriction 4.4 will work properly.

Finally, we need to define the capacity for each agent. We decided to add a 10%, this will allow overload for agents. Since travel would limit the agents capacity we will add an additional 20%. This additional 20% will allow agents to travel an average of 20km to attend each of the tasks in our model and still fulfill all of the farmer needs if they meet the minimum requirements, Equation 4.6 includes these considerations. In conclusion, we were able to define a capacity restriction that takes into account travel of the agents in this

four dimensional assignment problem.

$$Capacity_{\mathrm{j}} = (F_{\mathrm{i}} * N_{\mathrm{ik}}) * 1.3/(total\ number\ of\ agents) \tag{4.6}$$

With this, we have finished the first linear programming model formulation for the purpose of this study. As stated before, this model was executed in MATLAB, and only worked with small sizes of our problem (a maximum of 12 municipalities, 8 agents and 20 needs) due to computational complexity. Next, we describe the alternative model, this model was executed successfully in MATLAB for the full size problem (78 municipalities, 50 agents and 20 needs).

Alternative Model Formulation

The linear model described in Section 4.3.5 is not suitable for the purpose of this study due to its computational complexity. Therefore, we decided to split the model into two separate linear programming models, then iterate between them until a good solution is found. The first model assigns agents to needs in specific municipalities without changing the agent's base region. The second model assigns a base region to each agent given the output of the first model. Since we divided the problem into two models, we can not assure the global output will be optimal. In Section 4.4, we show how the results given by these two models compare to the global optimal solution given by the unified model described in Section 4.3.5. Next, we describe this alternative model formulation.

Alternative model formulation: Inputs and output

This model has the same inputs and output as the unified model, please see Section 4.3.1. Next, we describe the iterative process we used with these two models.

Alternative model formulation: Initialization

First step in this optimization process is to find an initial solution. The approach we used to create this initial solution is first assign agents to tasks (needs), since we only have resources and tasks but not initial positions for the agents, we run a simplified model similar to what was shown in 4.3.5 without taking into account the effect of travel in agents' capacity.

The objective function for this first step is 4.7. We consider skills and preferences as in 4.3.5. We avoid redundant assignments using the restriction shown in 4.3.5. For the capacity restriction we do not consider the effect of travel on capacity. The capacity restriction used is shown in Equation 4.8. This model generates an initial distribution of tasks to agents. Next, we use this output in a second model.

$$max\ Utility = \sum_{ijk}(X_{\mathrm{ijk}} * P_{\mathrm{jk}} * Pref_{\mathrm{jk}} * F_{\mathrm{i}} * N_{\mathrm{ik}}) \tag{4.7}$$

$$\sum_{i,k}(F_{\mathrm{i}} * N_{\mathrm{ik}} * X_{\mathrm{ijk}}) \leq Capacity_{\mathrm{j}} \qquad \forall j \tag{4.8}$$

The second model uses as input the distribution of tasks assigned to agents. This model assigns a base region to each agents so it minimizes the travel for the agents based on the task assignment.

The objective function 4.9 has one decision variable, M_{aj}, and four parameters F_{i}, N_{ik}, D_{ai}, and X_{ijk}. The definition of all these symbols were previously explained in 4.3.5. This model includes one restriction to ensure each agent is assigned to one and only one base

region.

$$min\ z = \sum_{ijka}(X_{\mathrm{ijk}} * F_{\mathrm{i}} * N_{\mathrm{ik}} * M_{\mathrm{ai}}) \tag{4.9}$$

$$\sum_{a}(M_{\mathrm{aj}}) = 1 \qquad \forall j \tag{4.10}$$

Once we executed the second model we have an initial assignment of agents to tasks in municipalities and a base region to each agent. Is worth mentioning that this first solution is not considering the travel effect on capacity. Therefore, these results will never be used as proposed assignment. The purpose of this solution is to have a starting point so we can iterate later. Next, we describe the models we will use to iterate.

Alternative model formulation: Iterations

The models we present here use as input the initial solution generated by the models in the previous Section 4.3.5. Figure 4.1 shows a flowchart of the iterative process.

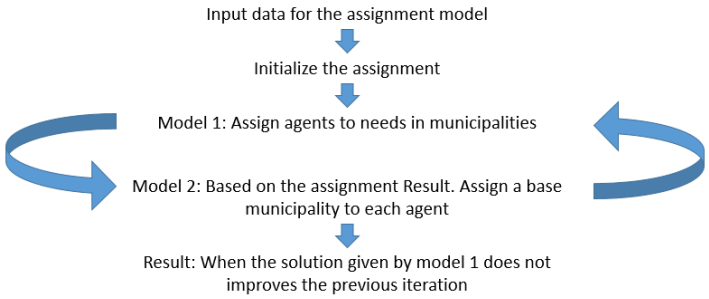


Figure 4.1: Restriction to ensure agents are assigned to one base region

Now we describe the first model we use in this iterative process. The model is the same described in Section except for the capacity restriction. The capacity restriction used is the same of the unified model, please see 4.3.5.

The second model is exactly the same described for the second step in Section . In order to see how this model performs see results in Next section.

4.4 Results

Primarily, we tested the functionality of the optimization models on different scenarios with dummy data, see Section 4.4.1 for details. We tested that our model satisfies the requisites we designed it for. First, we tested that agents are only assigned to tasks (farmer's needs) if they meet the minimum requirements for expertise and preferences. Second, no agent is assigned to a low need as defined in methodology. Third, the model should assign agents to tasks allowing idle agents and unassigned tasks simultaneously. Finally, agents cannot exceed the capacity restriction, taking into account assigned tasks and travel.

Additionally, in methodology we decided to use an alternative formulation, consisting in dividing the optimization model into two different models. Then we use the iterative model to find the solution to our assignment problem. Consequently, we cannot assure that the solution given by this iterative process is a global optimum. Thus, we need to evaluate the output of the iterative model. We decided to carry out this evaluation by comparing the results given by the unified model (global optimum) and the iterative model. Since the unified model is computationally complex, we can only run the model for a maximum size of 12 municipalities and 8 agents. Then, we decided to compare both modeling options up to this size. We explain how we performed this evaluation and discuss the results in Section 4.4.2.

4.4.1 Optimization Model Evaluation: Functionality Tests

In order to test the functionality of the optimization models we created a base scenario. This base scenario describes the inputs for the optimization model. Then, we modified the input of this base scenario to test the different characteristics the model has to fulfill. The base scenario created has a small size, so we can interpret the results easily. We chose a size of 5 agents, 5 needs, and 5 municipalities. Tables 4.1, 4.2, 4.3, and 4.4 describe the base scenario. We tested a total of three functionalities:

- Functionality 1: Agents should only be assigned to tasks only if we know they will perform according to the organization's minimum standards.
- Functionality 2: The assignment model should allow solutions with idle workers and unassigned tasks, simultaneously, if there is a mismatch between agents capabilities and tasks required skills.
- Functionality 3: capacity should restrict the quantity of tasks an agent can handle, and capacity of agents should be affected by travel between regions.

Please see the inputs of the assignment model for the base scenario in tables 4.1, 4.2, 4.3, and 4.4. Next, we present the results on the three functionality tests we performed.

Service Quality: Scenario 1

We want to assign agents to tasks only if we know they will perform according to the organization's standards. Thus, the model we created should not assign a task to an agent that does not fulfill the minimum requirements. These requirements include both preferences and expertise for each particular need (0.5 for each in this case). Therefore, we used the base scenario where there is a mismatch between preferences and skills for some particular

Table 4.1: Input 1: Expertise and preferences for agents in scenario 1

		Need 1	Need 2	Need 3	Need 4	Need 5
Agent 1	Expertise	0.7	0	0.5	0	0.1
	Preferences	0.7	0.6	0. 5	0.4	0.2
Agent 2	Expertise	0.1	0.7	0. 5	0.4	0.5
	Preferences	0.4	1	0.1	1	0.5
Agent 3	Expertise	0	0.7	0.4	0.6	0.7
	Preferences	0.4	0.8	0.7	0.3	0.6
Agent 4	Expertise	0.1	0.4	0. 4	0.8	0
	Preferences	0.3	0.4	0.4	0.7	0.6
Agent 5	Expertise	0.9	0.6	0.3	0. 1	0.2
	Preferences	0.9	0.7	0. 5	0.4	0

Table 4.2: Base Scenario: Number of farmers per municipality

	Municipality 1	Municipality 2	Municipality 3	Municipality 4	Municipality 5
Number of farmers	68	114	93	158	154

Table 4.3: Base Scenario: Distance between municipalities

	Municipality 1	Municipality 2	Municipality 3	Municipality 4	Municipality 5
Municipality 1	0	88	32	71	48
Municipality 2	88	0	25	55	47
Municipality 3	32	25	0	66	68
Municipality 4	71	55	66	0	15
Municipality 5	48	47	68	15	0

Table 4.4: Base Scenario: Distance between municipalities

	Need 1	Need 2	Need 3	Need 4	Need 5
Municipality 1	2	3	3	1	3
Municipality 2	2	2	3	1	1
Municipality 3	3	2	3	2	0
Municipality 4	1	0	2	3	1
Municipality 5	3	0	0	3	1

needs. Then, we analyzed the model output to confirm that agents were not assigned to those specific needs where the mismatch occurs. This scenario has 5 agents, 5 needs, and

5 municipalities. Please see Section 4.4.1 for the input details.

We used the base scenario as scenario 1 (case 1) to test that agents assigned to tasks always fulfill the minimum requirements (0.5) for both expertise and preferences. Table 4.5 presents the assignment results for this particular scenario (case 1). Results show agent 5 was assigned to need 1 in municipalities 1-3-5 and agent 1 to need 1 in municipality 2. Agent 2 was assigned to need 2 in municipalities 1-2-3. Agent 3 was assigned to need 3 in municipalities 2-4 and agent 1 was assigned to need 3 in municipalities 1-3. Agent 4 was assigned to need 4 in municipalities 3-4-5. Finally, agent 3 was assigned to need 5 in municipality 1. Thus, agents were only assigned to needs fulfilling minimum requirements (0.5) for both expertise and preferences. However, we need more information to assure this model is considering the minimum requirements for expertise and preferences properly. Therefore, we run the model with the exact same scenario removing from the model the service quality minimum requirements (case 2). This means we will not process the data as explained in Section 4.3.5. The final result was exactly the same except for need 3. Tables 4.10 present the assignment results for this particular scenario in need 3. Results show agent 1 was assigned to need 3 in municipalities 1-3 and agent 3 to need 3 in municipalities 2-4. The only agent fulfilling both requirements for need 3 is agent 1. Then, the model should assign agent 1 to need 3 and not agent 3 to need 3. Even though agent 3 would have better performance for this need if we did not establish a minimum requirement. Therefore, results show our model (case 1) is considering minimum requirements for service quality correctly (assigning agent 1 to need 3 and not agent 3).

Next, we present a scenario we used to test that our model allows idle agents and unassigned tasks simultaneously.

Table 4.5: Scenario 1: Assignment results for need 1

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	0	0	0	1
Municipality 2	1	0	0	0	0
Municipality 3	0	0	0	0	1
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	1

Table 4.6: Scenario 1: Assignment results for need 2

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	1	0	0	0
Municipality 2	0	1	0	0	0
Municipality 3	0	1	0	0	0
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	0

Table 4.7: Scenario 1: Assignment results for need 3

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	1	0	0	0	0
Municipality 2	1	0	0	0	0
Municipality 3	1	0	0	0	0
Municipality 4	1	0	0	0	0
Municipality 5	0	0	0	0	0

Table 4.8: Scenario 1: Assignment results for need 4

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	0	0	0	0
Municipality 2	0	0	0	0	0
Municipality 3	0	0	0	1	0
Municipality 4	0	0	0	1	0
Municipality 5	0	0	0	1	0

Not forcing the assignment: Scenario 2

Scenario 2 was designed to test if our model allows for idle workers and unassigned tasks simultaneously. Forcing the assignment could be detrimental to the company's reputation.

Table 4.9: Scenario 1: Assignment results for need 5

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	0	1	0	0
Municipality 2	0	0	0	0	0
Municipality 3	0	0	0	0	0
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	0

Table 4.10: Base Scenario 1.2: Assignment results for need 3

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	1	0	0	0	0
Municipality 2	0	0	1	0	0
Municipality 3	1	0	0	0	0
Municipality 4	0	0	1	0	0
Municipality 5	0	0	0	0	0

Especially in service organizations, where the client is in close contact with company's personnel. Therefore, we modified the input of the base scenario to create scenario 2, see Table 4.11. In this scenario, agents do not fulfill the minimum requirements for a particular need. Additionally, there is one low qualified agent with low preferences, not fulfilling minimum requirements for service quality in any particular need. Then, we analyzed the model output to confirm that agents were not assigned to those specific needs where the mismatch occurs.

We created scenario 2 to test the assignment is not forced. Tables 4.12, 4.13, 4.14, 4.15, and 4.16 present the assignment results for this particular scenario. Results show needs 3 and 5 are not assigned to any agent. We can see in Table 4.11 that there is no agent fulfilling minimum requirements (0.5) in both expertise and preferences for those particular needs. Therefore, it is desirable that no agent is assigned to those needs. Additionally, agent 3 is not assigned to any need. We can see in Table 4.11 that agent 3 does not fulfill minimum requirements for any particular need. Then, not assigning agent 3 to any need

Table 4.11: Input 1: Expertise and preferences for agents in scenario 1

		Need 1	Need 2	Need 3	Need 4	Need 5
Agent 1	Expertise	0.7	0	0.1	0	0.1
	Preferences	0.7	0.6	0.4	0.4	0.2
Agent 2	Expertise	0.1	0.7	0.5	0.4	0.1
	Preferences	0.4	1	0.1	1	0.5
Agent 3	Expertise	0	0.7	0	0.6	0.4
	Preferences	0.4	0	0.7	0.3	0.6
Agent 4	Expertise	0.1	0.4	0.4	0.8	0
	Preferences	0.3	0.4	0.4	0.7	0.6
Agent 5	Expertise	0.9	0.6	0.3	0.1	0.2
	Preferences	0.9	0.7	0.5	0.4	0

is what we expected from the model output. However, Table 4.11 shows agent 1 does fulfill minimum requirements for need 1 (0.7 expertise and 0.7 preference). Still, agent 1 is almost completely idle. If we go into more detail, we can see agent 5 has 0.9 expertise and 0.9 preference for need 1. Thus, agent 5 is outperforming agent 1. This is the reason agent 1 is almost idle in this scenario.

Table 4.12: Scenario 1: Assignment results for need 1

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	0	0	0	1
Municipality 2	1	0	0	0	0
Municipality 3	0	0	0	0	1
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	1

Table 4.13: Scenario 1: Assignment results for need 2

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	1	0	0	0
Municipality 2	0	1	0	0	0
Municipality 3	0	1	0	0	0
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	0

Table 4.14: Scenario 1: Assignment results for need 3

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	0	0	0	0
Municipality 2	0	0	0	0	0
Municipality 3	0	0	0	0	0
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	0

Table 4.15: Scenario 1: Assignment results for need 4

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	0	0	0	0
Municipality 2	0	0	0	0	0
Municipality 3	0	0	0	1	0
Municipality 4	0	0	0	1	0
Municipality 5	0	0	0	1	0

Table 4.16: Scenario 1: Assignment results for need 5

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	0	0	0	0
Municipality 2	0	0	0	0	0
Municipality 3	0	0	0	0	0
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	0

In conclusion, results showed the model is not forcing any assignment. Needs can be unassigned if no agent is fulfilling minimum requirements. Additionally, agents will be idle if not fulfilling service quality requirements for any need. Finally, agents could be idle if they are outperformed by other agents. Then, we tested no assignment is forced in our optimization model.

Capacity & travel for agents : Scenario 3

Scenario 3 was designed to test if the capacity and travel restriction of our model works correctly. We tested two characteristics. First, we reduce agents capacity, so agents should be assigned to fewer tasks. Second, we increase distances between municipalities, this should also affect agents capacity to perform tasks. For details about how the capacity restriction works in our model, please see Section 4.3.5.

In scenario 3 (case 1) we use the base scenario increasing the capacity by 1,000 times its original value. This would be equivalent to not having any capacity restriction. Results show there is a change in the agents assigned to need 1, please see Table 4.17. Agent 5 was assigned to need 1 in municipalities 1-2-3-5. In the original solution to the base scenario agent 5 was assigned to need 1 in municipalities 1-3-5, and agent 1 was assigned to need 1 in municipality 2. Since, agent 5 has the best skills and preferences for need 1, this is what we could expect if we do not have capacity restriction.

Table 4.17: Scenario 1: Assignment results for need 1

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	0	0	0	0	1
Municipality 2	0	0	0	0	1
Municipality 3	0	0	0	0	1
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	1

In scenario 3 (case 2), we use the base scenario reducing capacity to an 80% of its original value. Results show there is a change in the agents assigned to Need 1, please see Table 4.18. Agent 5 was assigned to need 1 in municipalities 3-5 and agent 1 was assigned to need 1 in municipalities 1-2. In the original solution to the base scenario agent 5 was assigned to need 1 in municipalities 1-3-5 and agent 1 was assigned to need 1 in municipality 2. In this case we have the opposite of case 1. Agent 5 has better skills and

preferences for need 1 than agent 1. Thus, reducing capacity makes agent 5 to be assigned to less tasks and be substituted by agent 1. In conclusion, reducing capacity works properly for our model

Table 4.18: Scenario 1: Assignment results for need 1

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	1	0	0	0	0
Municipality 2	1	0	0	0	0
Municipality 3	0	0	0	0	1
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	1

In scenario 3 (case 3), we modified the distance matrix from the base scenario, reducing distance between municipalities. Table 4.19 shows the new distance matrix. Results show there is change in the assignment of need 1. Table 4.20 shows the results of the assignment for need 1. In this case, agent 5 was assigned to need 1 in municipalities 2-3-5 and agent 1 was assigned to need 1 in municipality 1. In the original solution to the base scenario agent 5 was assigned to need 1 in municipalities 1-3-5 and agent 1 was assigned to need 1 in municipality 2. Agent 5 has better skills and preferences for need 1 than agent 1. Thus, reducing the distance between municipalities allowed agent 5 to be assigned to municipality 2 instead of municipality 1. Looking deeper at the base scenario input, we realized municipality 1 and 2 have the same value for need 1. However, municipality 2 has 114 farmers and municipality 1 has 68 farmers. Therefore, reducing distances between municipalities allowed agent 5 to be assigned to more farmers in demanding need 1, increasing overall utility of the assignment. In conclusion, we tested our model to ascertain that it considers properly the effect of distance in capacity.

Finally, in scenario 3 (case 4), we reduced capacity to zero. Results showed no agent was assigned to any need. Matrixes are all zeros for the 5 needs. Therefore, we do not show the output matrixes for this third case. In conclusion, our model showed capacity

Table 4.19: Base Scenario: distance between municipalities

	Municipality 1	Municipality 2	Municipality 3	Municipality 4	Municipality 5
Municipality 1	0	8.8	3.2	7.1	4.8
Municipality 2	8.8	0	25	55	4.7
Municipality 3	3.2	2.5	0	66	6.8
Municipality 4	7.1	5.5	66	0	1.5
Municipality 5	4.8	4.7	6.8	1.5	0

Table 4.20: Scenario 1: Assignment results for need 1

	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5
Municipality 1	1	0	0	0	0
Municipality 2	0	0	0	0	1
Municipality 3	0	0	0	0	1
Municipality 4	0	0	0	0	0
Municipality 5	0	0	0	0	1

and travel restriction defined in Section 4.3.5 fulfills the requirements we designed it for. Next, we evaluate the output of iterative model, comparing the results with the output of the unified model (optimal).

4.4.2 Comparison between the two modeling options

As explained before, we have a limit on the size data for the unified model due to computational complexity. Therefore, we were able to compare both models only up to a size of 12 municipalities and 8 agents. However, the study of demand for farmers results in an estimation of farmer's needs across the 78 Puerto Rican municipalities. Thus, we need a smaller version of the problem for the purpose of this comparison. We decided to take a random sample of the municipalities in order to create a smaller version of the problem and carry out the comparison of both modeling options. Accordingly, we did the same to reduce the number of agents. Next, we go into more detail on the scenarios tested in this

comparison and the metrics selected for the comparison.

We performed this comparison at four different sizes of the problem. First, 12 municipalities and 8 agents, the maximum possible to execute the unified model. Second, 11 municipalities and 7 agents, the contiguous lower size maintaining the proportion of agents and municipalities of the full problem. Accordingly, the next size is 9 municipalities and 6 agents. Finally, the smallest size we tested is 8 municipalities and 5 agents. These are the scenarios we tested in the comparison. Next, we explain the metrics we used to compare the outputs of both modeling options.

Table 4.21: Comparison between performance of both methods. Utility is an average of 100 replicas for each sample. *m* represents number of municipalities and *a* represents number of agents

		Size1 (m=8 a=5)	Size2 (m=9 a=6)	Size3 (m=11 a=7)	Size4 (m=12 a=8)
Utility	Iterative model	63,873	88,826	104,529	119,305
	Unified model	64,883	89,344	105,351	120,201
Utility/municipality	Iterative model	7,984	9,870	9,503	9,942
	Unified model	8,110	9,927	9,577	10,017

There is very little difference in utility between both modeling options. Average *utility*/*municipality* of all 400 replicas (100 replicas per size and 4 sizes) for the iterative model is less than 1% lower than the average *utility*/*municipality* for the unified model (optimal solution). Then, the iterative model is performing very well up to size of 12 municipalities and 8 agents. Next, we run the iterative model for the full size problem (78 municipalities and 50 agents). Utility was 931,121, *utility*/*municipality* was 11,937. The 11,937 *utility*/*municipality* is in the 79th percentile of all 400 replicas. One would expect that some of the replicas have greater *utility*/*municipality* than the full size problem, because a random sample will sometimes select municipalities with very high needs and/or

an extremely good fit between agents' expertise and municipalities' needs. Therefore, two points support our iterative model is giving a good solution. First, the results presented in Table 4.21 show our model performs very similar to the optimal solution for small sizes. Second, the full size solution is in the 79th percentile of all 400 replicas. One would expect some of the random samples for the small size problem would have very high needs or extremely good 'fit' between agents and municipalities. This extremely good 'fit' would lead to very high $utility/municipality$ results not comparable to the full size problem where we have 'unfit' resources and municipalities with low needs. Therefore, being in the 79th percentile of the 400-replica's optimal solution shows our iterative model is performing well for the full size problem.

4.5 Conclusion

There is a lot of work in the field of linear programming models for assignment of resources to tasks. Still some of the difficulties faced by service organizations were not addressed before. Especially when the personnel profile is diverse along with a variety of tasks needed to be performed. These situations lead to mismatch between resources' capabilities and tasks required expertise. Traditional assignment models force the assignment, thus no task is uncompleted or no agent has idle time. This often lead to agents assigned to tasks where they perform under the company's standards. This study presents a model allowing idle time for resources and uncompleted tasks simultaneously if there is no possible match between them.

We tested the model on our test bed, the Puerto Rico Agricultural Extension Service. This test bed is a 4 dimension assignment problem, it assigns agents to needs in regions, and it also assigns agents to base regions. 4-dimension assignment problems are known to be np-hard. Thus, we were only able to run the full (unified) model in small size versions

of the full size problem. Then, we developed an iterative model, dividing the full model in 2 different models. Each model uses as input the other model output and we stop the iterative process when there is no improvement in the solution.

The iterative model was tested in different scenarios to test functionality. Additionally, it was compared with the unified model up to the maximum size. The execution was performed using MATLAB (see Appendix C for the Technical Specifications of the software). Results show our iterative model passed the functionality tests and performed very well compared to the unified model (optimal solution). In conclusion, we were able to develop an assignment model addressing a solution for some of the common challenges faced by service organizations using as test bed the Puerto Rico Agricultural Extension Service.

Chapter 5

Conclusion

This thesis contributes in two different fields of knowledge. First, we made a contribution in clustering analysis. We developed the "Jump Analysis Method". A method that improves clustering extraction from Ward's dendrograms when groups have different sizes and densities. Second, a novel methodology to assign resources to tasks for complex service-systems, where mismatch between personnel profile and task requirements is considered. A contemporary assignment-model would force the assignment of all resources or all demand, and in either case often resulting in resources assigned to tasks they will perform poorly. This research used as test bed the Puerto Rico Agricultural Extension Service.

Is well known that Ward's hierarchical clustering method tend to fail in scenarios where groups have different sizes and densities. We hypothesized that performance of Ward's hierarchical clustering can improve in these scenarios by using a multilevel cut on Ward's dendrograms. We developed Jump Analysis Method (JAM), a novel technique for analyzing dendrograms generated with Ward's hierarchical clustering method. This technique requires simple parameter tuning, and only basic statistical knowledge in order to be implemented. Additionally, this approach allows users to cut a dendrogram at dierent lev-

els unlike traditional horizontal cuts (i.e."stopping rules"). This methodology was tested with success for dierent scenarios including our test bed. Chapter 2 presents this analysis technique and various tests to proof the utility of the contribution. Chapter 3 presents a study of demand for our test bed where JAM is used along with other statistical analysis techniques. Results validated our hypothesis, showing this method can improve Ward's hierarchical clustering grouping results in scenarios where groups have different sizes and densities.

In the development of the test bed case, we realized there was no record of service requests made by clientele. Therefore, a study of demand was conducted in chapter 3. We studied farmer's needs in Puerto Rico. We first designed a questionnaire, challenging a common assumption that farmers will have only needs related to the agricultural technique. This assumption was made in the previous studies of farmer's needs we have found. We included the possibility of farmers having business related needs, such as marketing or business accounting. Results showed farmer's in Puerto Rico have their highest needs in business related areas, validating our questionnaire design approach. Specifically, farmers have a very high need for advise in marketing their products, product processing, and finding capital trough incentives and funding. The highest need related to animal production was advise in nutrition and supplementation. The highest needs related to crop production were plague management and fertilization.

In chapter 4 we presented the resource assignment model contribution. Unlike traditional models, the assignment model solution presented in this study considers the possibility of having both idle workers and unassigned tasks simultaneously, without the use of dummy resources and/or dummy demand. In this model, if assigning a resource to a task may be detrimental (i.e. not in the best interest of the customer), then, there will be unmet demand. This allows the system's management to clearly identify opportunities

for training and recruitment in their system. The organization can be confident no task is performed under their minimum service standards.

The assignment model presented includes other sources of complexity commonly faced by service organizations. First, the myriad of tasks defined by the skills needed to be performed. Second, a diverse resource profile defined by their skills and preferences. Third, the relationship between tasks (needs) and workload is unknown. We have not found a previous study solving an assignment problem in a situation where the relationship between tasks and workload is unknown. Finally, we included in the model that resources' capacity is affected by travel between regions.

The assignment of resources to tasks for our test bed is a 4-dimension assignment problem, because we assign agents to tasks in regions, and we also assign agents to base locations. n-dimension assignment problems are known to be np-hard, when n is greater than 2. Therefore, we were not able to run the full size problem with the computing capacity in the research lab; moreover, we expect that a mid-size service system will not have the capacity to run a full size of such model. Then, we developed an iterative process dividing the model into 2 different assignment models. These two models feed each other and the process ends when the solution does not improve in one iteration. We run various tests for iterative model working under the conditions of our test bed and results were satisfactory.

We compared *utility* results of our model and global optimum (unified model) up to the size of 12 municipalities and 8 agents. 12 municipalities and 8 agents is the maximum size we were able to run for the unified model. We generated 400 random samples of our problem to test performance up to this maximum size. *Utility* results showed our model performed very well. Average *utility* for the iterative model was less than 1% lower than the global optimum. Therefore, this iterative model performed very well up to

this size. Additionally, we used the metric *utility/municipality* to evaluate results on the full size problem. *Utility/municipality* allows us to compare results on different sizes of our problem. Then, we compared the solution of the full size problem with the optimal solution for the 400 random samples generated in the previous test. Solution for the full size problem was in 79th percentile of the 400 random samples optimal solutions. This result shows our iterative model performed well for the full size problem.

Future extension for the assignment model presented in this research is to evaluate assignment results to identify training and recruiting opportunities. The scope of the model presented here is to solve an allocation problem. Some training and recruiting opportunities can be easy to identify from the model results. However, identifying the best training and recruiting opportunities is valuable information for any organization, especially in complex scenarios as our test bed. Additional extension is to evaluate performance of the iterative model in different allocation problems. The performance evaluations conducted were related to the test bed and some functionality tests. Considering different problems would require some adjustments to the model and could validate the iterative approach as a general solution for these problems.

In the case of the Jump Analysis Method (JAM). We developed a methodology to analyze dendrograms generated with Ward's linkage. An extension would be to improve metrics for the *jump* and *relative jump*. This metrics can consider the number of observations under each merge point. The *relative jump* metric combined with multilevel cut showed a new way to look at dendrograms. Something similar can be implemented in dendrograms generated with other linkage metrics, not only Ward's linkage.

Bibliography

4-H. Positive youth development and mentoring organization, 2015. URL `http://www.4-h.org/`.

J. Almeida, L. Barbosa, A. Pais, and S. Formosinho. Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 87(2):208–217, 2007. doi: 10.1016/j.chemolab.2007.01.005.

S. Barreto, C. Ferreira, J. Paix£o, and B. S. Santos. Using clustering analysis in a capacitated location-routing problem. *European Journal of Operational Research*, 179(3):968 – 977, 2007. doi: http://dx.doi.org/10.1016/j.ejor.2005.06.074.

P. Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

C. Budayan, I. Dikmen, and M. T. Birgonul. Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy c-means method for strategic grouping. *Expert Systems with Applications*, 36(9):11772–11781, 2009. doi: 10.1016/j.eswa.2009.04.022.

A. M. Campbell and M. W. Savelsbergh. A decomposition approach for the inventory-routing problem. *Transportation science*, 38(4):488–502, 2004.

G. Caron, P. Hansen, and B. Jaumard. The assignment problem with seniority and job priority constraints. *Operations Research*, 47(3):449–453, 1999.

L. Chase, L. Ely, and M. Hutjens. Major advances in extension education programs in dairy production. *Journal of dairy science*, 89(4):1147–1154, 2006.

Y.-F. Chuang, H.-T. Lee, and Y.-C. Lai. Item-associated cluster assignment model on storage allocation problems. *Computers & Industrial Engineering*, 63(4):1171–1177, 2012. doi: 10.1016/j.cie.2012.06.021.

G. M. Downs and J. M. Barnard. Clustering methods and their uses in computational chemistry. *Reviews in computational chemistry*, 18:1–40, 2002.

M. Elango, S. Nachiappan, and M. K. Tiwari. Balancing task allocation in multi-robot systems using k-means clustering and auction based mechanisms. *Expert Systems with Applications*, 38(6):6486–6491, 2011. doi: 10.1016/j.eswa.2010.11.097.

V. Fernandez-Viagas and J. M. Framinan. Integrated project scheduling and staff assignment with controllable processing times. *The Scientific World Journal*, 2014, 2014.

M. L. Fisher and R. Jaikumar. A generalized assignment heuristic for vehicle routing. *Networks*, 11(2):109–124, 1981.

J. Gong and D. W. Oard. Selecting hierarchical clustering cut points for web person-name disambiguation. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 778–779, New York, USA, 2009. doi: 10.1145/1571941.1572124.

R. Govind, R. Chatterjee, and V. Mittal. Timely access to health care: Customer-focused resource allocation in a hospital network. *International Journal of Research in Marketing*, 25(4):294–300, 2008.

T. Grubinger, A. Zeileis, and K.-P. Pfeiffer. evtree: Evolutionary learning of globally optimal classification and regression trees in r. Technical report, Working Papers in Economics and Statistics, 2011.

S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. *SIGMOD Rec.*, 27(2):73–84, June 1998. ISSN 0163-5808. doi: 10.1145/276305.276312. URL http://doi.acm.org/10.1145/276305.276312.

S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. In *Data Engineering, 1999. Proceedings., 15th International Conference on*, pages 512–521. IEEE, 1999. doi: 10.1109/ICDE.1999.754967.

S. M. Hashemi, M. Mokhtarnia, J. M. Erbaugh, and A. Asadi. Potential of extension workshops to change farmers' knowledge and awareness of ipm. *Science of the total environment*, 407(1):84–88, 2008.

S. M. Hashemi, S. M. Hosseini, and C. A. Damalas. Farmers' competence and training needs on pest management practices: Participation in extension workshops. *Crop Protection*, 28(11):934–939, 2009.

G. T. Ho, W. Ip, C. Lee, and W. Mou. Customer grouping for better resources allocation using ga based clustering technique. *Expert Systems with Applications*, 39(2):1979–1987, 2012. doi: 10.1016/j.eswa.2011.08.045.

R. Hu, Y. Cai, K. Z. Chen, and J. Huang. Effects of inclusive public agricultural extension

service: Results from a policy reform experiment in western china. *China Economic Review*, 23(4):962–974, 2012.

J. H. W. Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.

Y. Jung, H. Park, D.-Z. Du, and B. L. Drake. A decision criterion for the optimal number of clusters in hierarchical clustering. *J. of Global Optimization*, 25(1):91–111, Jan. 2003. ISSN 0925-5001. doi: 10.1023/A:1021394316112.

G. Karypis, E.-H. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75, 1999. doi: 10.1109/2.781637.

M. Kaur and U. Kaur. Comparison between k-means and hierarchical algorithm using query redirection. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(7), 2013.

B. R. Kiran, J. Serra, and J. Cousty. Climbing: a unified approach for global constraints on hierarchical segmentation. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 324–334. Springer, 2012. doi: 10.1007/978-3-642-33885-4_33.

İ. Korkmaz, H. Gökçen, and T. Çetinyokuş. An analytic hierarchy process and two-sided matching based decision support system for military personnel assignment. *Information Sciences*, 178(14):2915–2927, 2008.

P. Langfeldera, B. Zhangb, and S. Horvatha. Dynamic tree cut: in-depth description, tests and applications. *November*, 22:2007, 2007.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

T. Manjala et al. An extension officer's perspective on practice change. *Extension Farming Systems Journal*, 5(1):119, 2009.

S. A. Mingoti and J. O. Lima. Comparing som neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, 174(3):1742–1759, 2006. doi: 10.1016/j.ejor.2005.03.039.

J. F. Nash Jr. The bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 155–162, 1950.

C. G. A. Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

A. Niknafs, J. Denzinger, and G. Ruhe. A systematic literature review of the personnel assignment problem. In *Proceedings of the International Multiconference of Engineers and Computer Scientists, Hong Kong*, 2013.

C. F. Olson. Parallel algorithms for hierarchical clustering. *Parallel computing*, 21(8): 1313–1325, 1995. doi: 10.1016/0167-8191(95)00017-I.

T. Öncan. A survey of the generalized assignment problem and its applications. *Infor*, 45 (3):123, 2007.

J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160–1167, 2009. doi: 10.1200/JCO.2008.18.1370.

D. W. Pentico. Assignment problems: A golden anniversary survey. *European Journal of Operational Research*, 176(2):774–793, 2007.

M. L. Peters and S. Zelewski. Assignment of employees to workplaces under consideration of employee competences and preferences. *Management Research News*, 30(2):84–99, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

H. Ruifa, Y. Zhijian, P. Kelly, and J. Huang. Agricultural extension system reform and agent time allocation in china. *China Economic Review*, 20(2):303–315, 2009.

S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576–584. IEEE, 2004. doi: 10.1109/ICTAI.2004.50.

J. Sander, X. Qin, Z. Lu, N. Niu, and A. Kovarsky. Automatic extraction of clusters from hierarchical clustering representations. In *Advances in knowledge discovery and data mining*, pages 75–87. Springer, 2003. doi: 10.1007/3-540-36175-8_8.

Y.-Y. Shih and C.-Y. Liu. A method for customer lifetime value rankingcombining the analytic hierarchy process and clustering analysis. *Journal of Database Marketing and Customer Strategy Management*, 11(2):159–172, 2003.

R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973. doi: 10.1093/comjnl/16.1.30.

Speech and I. P. Unit. Clustering datasets, 2016. URL `https://cs.joensuu.fi/sipu/datasets/`.

S. S. Tax, S. W. Brown, and M. Chandrashekaran. Customer evaluations of service complaint experiences: implications for relationship marketing. *The journal of marketing*, pages 60–76, 1998.

P. Willett. Recent trends in hierarchic document clustering: a critical review. *Information Processing & Management*, 24(5):577–597, 1988. doi: 10.1016/0306-4573(88)90027-1.

T. Wongwien and S. Nanthavanij. Ergonomic workforce scheduling with productivity and employee satisfaction consideration. In *Proceedings of the 4th International Conference on Engineering, Project, and Production Management, Bangkok, Thailand*, 2013.

J.-Y. W. Yi-Fei Chuang, Shui-Hui Chia. Enhancing order-picking efficiency through data mining and assignment approaches. *WSEAS Transactions on Business & Economics*, 11(1):52–64, 2014.

T. Zhang, R. Ramakrishnan, and M. Livny. Birch: An efficient data clustering method for very large databases. *SIGMOD Rec.*, 25(2):103–114, June 1996. ISSN 0163-5808. doi: 10.1145/235968.233324. URL `http://doi.acm.org/10.1145/235968.233324`.

Appendices

Appendix A

Farmer’s need survey

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ
COLEGIO DE INGENIERIA
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

HOJA DE CONSENTIMIENTO INFORMADO

"Formulario para estudio de necesidades de los Agricultores"

Introducción

Este formulario está diseñado para llevar a cabo un **estudio de las necesidades de los agricultores de Puerto Rico**.

Antes de que decida rellenar el formulario, le agradecemos que lea cuidadosamente esta hoja de consentimiento informado. Cualquier duda la puede consultar con la persona que le entregó el formulario o a través de la información de contacto que aparece al final de esta hoja.

Información de contacto

Yo, _____le puedo resolver cualquier duda con el formulario incluso después de haberlo realizado, información de contacto al final de esta hoja.

Propósito

El objetivo de este formulario es analizar una muestra de los agricultores de Puerto Rico para estimar las necesidades de toda la población de agricultores.

Participación voluntaria y derechos

La participación en este formulario es completamente voluntaria, usted tiene el derecho de abandonar la actividad en cualquier momento.

Posibles beneficios y riesgos

El formulario es completamente anónimo. La duración de este formulario es de alrededor de 10 minutos, esta es la única incomodidad que se le ocasionará. Este estudio, de ser utilizado por las agencias que se dedican a apoyar a los agricultores, les permitirá ofrecer un servicio más acorde a las necesidades actuales de la población de agricultores.

Procedimiento

Se le explicará toda la información que esta Hoja de Consentimiento Informado incluye. Más adelante, se le entregará la Hoja de Consentimiento Informado y el formulario. Si usted decide participar, completará el formulario y lo devolverá a quien se lo entrego junto con la hoja de consentimiento firmada, quien a su vez le dejará una copia de la hoja de consentimiento informado firmada por uno de los integrantes del grupo de investigación.

Información de contacto:

Cualquier duda o comentario contactar a Miguel Ángel Ruiz Hernández:
Correo electrónico: miguel.ruiz6@upr.edu
Teléfono: 787-832-4040 extensión 3819 ó 3204 (Departamento de Ingeniería Industrial de la UPRM).

Firma del Participante:_____

Formulario para estudio de necesidades de los Agricultores en Puerto Rico

Propósito del estudio: Conocer las necesidades y/o problemas comunes a los que se enfrentan los agricultores de Puerto Rico con el fin de lograr una mayor eficiencia en las tareas del Servicio de Extensión Agrícola para satisfacer estas necesidades.

Equipo de investigadores: Miguel Ángel Ruiz: correo electrónico: miguel.ruiz6@upr.edu, Betzabé Rodríguez: correo electrónico: betzabe.rodriguez@upr.edu, Saylisse Dávila: saylisse.davila@upr.edu, Guillermo Ortiz: guillermo.ortiz@upr.edu, Viviana Cesaní: vivianai.cesani@upr.edu, Jorge Tirado: jorge.tirado1@upr.edu
Teléfono: 787-832-4040, extensión 3819 ó 3204 (Departamento de Ingeniería Industrial, UPRM)

1. ¿Completó usted este cuestionario anteriormente? ☐ SÍ / ☐ NO	
2. Pueblo: _____	3. Código Postal: _____
4. ¿Es la agricultura su principal actividad económica? ☐ SÍ / ☐ NO	

5-1 Seleccione su máximo nivel educativo:

☐ Intermedia ☐ Superior ☐ Certificación ☐ Grado asociado

☐ Bachillerato ☐ Maestría ☐ Doctorado

5-2 Profesional en que área: _____

6. Tipo e importancia de productos:
(I) Enliste en la siguiente tabla los productos de cultivo o ganadería de su negocio (e.g., leche de vaca, pollo, piñas, aguacates, etc.) (II) Luego **INDIQUE LA IMPORTANCIA** de cada uno de los productos enlistados. Entiéndase por importancia la relevancia en términos de ganancias que significa ese producto para su negocio. (III) Favor señalar en la columna ennegrecida cuál de los productos es el principal de su negocio, en caso de haber dos productos principales igual de importantes escoja ambos.

Ejemplo: Suponga que produce plátanos, tomates y cilantrillo. Suponga también que la parte básica de su negocio son los plátanos (producto principal), el cilantrillo es importante pero no mucho, y por último los tomates son unas pocas matas. A modo de ejemplo quedaría así:

Producto	Extremadamente importante	Muy importante	Moderadamente importante	Poco importante	Sin importancia alguna	Producto principal
6.a) plátanos	X	☐	☐	☐	☐	X
6.b) cilantrillo	☐	☐	X	☐	☐	☐
6.c) tomates	☐	☐	☐	X	☐	☐

Producto	Extremadamente importante	Muy importante	Moderadamente importante	Poco importante	Sin ninguna importancia	Producto principal
6-1_____	☐	☐	☐	☐	☐	☐
6-2_____	☐	☐	☐	☐	☐	☐
6-3_____	☐	☐	☐	☐	☐	☐
6-4_____	☐	☐	☐	☐	☐	☐
6-5_____	☐	☐	☐	☐	☐	☐
6-6_____	☐	☐	☐	☐	☐	☐
6-7_____	☐	☐	☐	☐	☐	☐
6-8_____	☐	☐	☐	☐	☐	☐

6-9 Comentarios: _____

_____

7. Identifique las **áreas en las que NECESITA ASESORAMIENTO, indicando la importancia de este asesoramiento.**

7-1 Técnicas de **producción animal:**

7-1 Técnicas de producción animal	Extremadamente importante	Moderadamente importante	Sin ninguna importancia
7-1.1 Nutrición y Suplementación	☐	☐	☐
7-1.2 Forraje	☐	☐	☐
7-1.3 Reproducción	☐	☐	☐
7-1.4 Manejo y Control animal	☐	☐	☐
7-1.5 Enfermedades (prevención y manejo)	☐	☐	☐
7-1.6 Otro:_____	☐	☐	☐

7-2 Técnicas de **cultivo:**

7-2 Técnicas de cultivo	Extremadamente importante	Moderadamente importante	Sin ninguna importancia
7-2.1 Sistemas de riego	☐	☐	☐
7-2.2 Análisis de suelo	☐	☐	☐
7-2.3 Manejo de plagas	☐	☐	☐
7-2.4 Control de erosión	☐	☐	☐
7-2.5 Abonamiento	☐	☐	☐
7-2.6 Manejo post-cosechas	☐	☐	☐
7-2.7 Asesoramiento sobre variedades de cultivos	☐	☐	☐
7-2.8 Otro:_____	☐	☐	☐

7-3 **Negocio:**

7-3 Negocio	Extremadamente importante	Moderadamente importante	Sin ninguna importancia
7-3.1 Cuentas del negocio	☐	☐	☐
7-3.2 Mercadeo	☐	☐	☐
7-3.3 Energía sustentable	☐	☐	☐
7-3.4 Fondos disponibles (búsqueda de financiación e incentivos)	☐	☐	☐
7-3.5 Obtención Registros y Certificaciones	☐	☐	☐
7-3.6 Manejo de desperdicios	☐	☐	☐
7-3.7 Educación continua	☐	☐	☐
7-3.8 Procesamiento producto (convertir en producto elaborado)	☐	☐	☐
7-3.9 Otro:_____	☐	☐	☐

SERVICIO DE EXTENSIÓN AGRÍCOLA
COLEGIO DE CIENCIAS AGRÍCOLAS

7-4 Comentarios:

8. **Método de asesoramiento**

8-1 Escoja a través de qué **actividad prefiere recibir el asesoramiento, seleccionar únicamente una actividad.**

☐ Visitas a la finca	☐ Seguimiento telefónico	☐ Publicación de artículos	☐ Demostraciones de método en finca
☐ Demostraciones de método en estación experimental	☐ Reuniones para tratar un tema específico	☐ Otro	

8-2 Comentarios:

9. Describa lo que usted como agricultor espera recibir por parte del servicio de extensión agrícola.

_____

_____

_____

_____

_____

_____

_____

_____

_____

Appendix B

IRB Protocols: Farmer's need survey

Comité para la Protección de los Seres Humanos en la Investigación
CPSHI/IRB 00002053
Universidad de Puerto Rico – Recinto Universitario de Mayagüez
Decanato de Asuntos Académicos
Call Box 9000
Mayagüez, PR 00681-9000

2 de febrero de 2015

Sr. Miguel A. Ruiz
Estudiante Graduado
Departamento de Ingeniería Industrial
Facultad de Ingeniería
Recinto Universitario de Mayagüez

Estimada Sr. Ruiz:

Como presidente del Comité para la Protección de los Seres Humanos en la Investigación (CPSHI) he revisado las modificaciones realizadas a su proyecto de investigación titulado "Study of Puerto Rican Farmer's Needs" (Protocolo núm. 20141110). Luego de evaluar su protocolo y demás documentación he determinado sigue siendo exento bajo la Categoría 2 del 45 CFR 46.101(b).

Cualquier cambio al protocolo o a la metodología deberá ser revisado y aprobado por el CPSHI antes de su implantación. El CPSHI deberá ser informado de inmediato de cualquier efecto adverso o problema inesperado que surgiera con relación al riesgo de los seres humanos, de cualquier queja sobre la conducción de esta investigación y de cualquier violación a la confidencialidad de los participantes.

Agradecemos su compromiso con los más altos estándares de protección de los seres humanos en la investigación y le deseamos éxito en su proyecto. Queda de usted,

Atentamente,

Dr. Rafael A. Boglio Martínez
Presidente
CPSHI/IRB
UPR - RUM

Comité para la Protección de los Seres Humanos en la Investigación
CPSHI/IRB 00002053
Universidad de Puerto Rico – Recinto Universitario de Mayagüez
Decanato de Asuntos Académicos
Call Box 9000
Mayagüez, PR 00681-9000

22 de mayo de 2015

Sr. Miguel A. Ruiz
Estudiante Graduado
Departamento de Ingeniería Industrial
Facultad de Ingeniería
RUM

Estimado Sr. Ruiz:

Como presidente del Comité para la Protección de los Seres Humanos en la Investigación (CPSHI) he revisado las modificaciones realizadas a su proyecto de investigación titulado "Study of Puerto Rican Farmer's Needs" (Protocolo num. 20141110). Luego de evaluar su protocolo y demás documentación he determinado sigue exento bajo la Categoría 2 del 45 CFR 46.101(b).

Cualquier cambio al protocolo o a la metodóloga deberá ser revisado y aprobado por el CPSHI antes de su implantación. El CPSHI deberá ser informado de inmediato de cualquier efecto adverso o problema inesperado que surgiera con relación al riesgo de los seres humanos, de cualquier queja sobre la conducción de esta investigación y de cualquier violación a la confidencialidad de los participantes.

Atentamente,

Dr. Rafael A. Boglio Martínez
Presidente
CPSHI/IRB
UPR - RUM

Appendix C

Technical Specifications

TECHNICAL ESPECIFICATIONS

The computer used in this work was a Dell Precision T7810 with the following specifications:

- Installed RAM memory: 64GB
- Processor: Intel® Xeon® Processor E5-2630 (2.3 Ghz)
- System: Windows 10 Professional 64-bit Copyright © 2009 Microsoft Corporation

The R software version used was R version 3.2.1 (2015-06-18). The detailed information given by version command in R was:

```
> version
               _
platform       x86_64-w64-mingw32
arch           x86_64
os             mingw32
system         x86_64, mingw32
status
major          3
minor          2.1
year           2015
month          06
day            18
svn rev        68531
language       R
version.string R version 3.2.1 (2015-06-18)
nickname       World-Famous Astronaut
```

The MATLAB ® software version used in this work was R2015a.

Appendix D

Notation Keys for Tables

Table D.1: Meaning of needs notation used in this document

Notation	Meaning
N1	Nutrition and supplementation
N2	Forage
N3	Reproduction
N4	Animal management and control
N5	Diseases prevention and management
N6	Irrigation systems
N7	Soil Analysis
N8	Pests management
N9	Erosion control
N10	Fertilization
N11	Post harvest management
N12	Crop varieties
N13	Accounting
N14	Marketing
N15	Sustainable energy
N16	Funds available
N17	Obtaining certifications
N18	Waste management
N19	Continuous education
N20	Product processing

Table D.2: Meaning of Municipality notation used in this document

Notation	**Meaning**
M1	Adjuntas
M2	Aguada
M3	Aguadilla
M4	Aguas Buenas
M5	Aibonito
M6	Aasco
M7	Arecibo
M8	Arroyo
M9	Barceloneta
M10	Barranquitas
M11	Bayamón
M12	Cabo Rojo
M13	Caguas
M14	Camuy
M15	Canóvanas
M16	Carolina
M17	Cataño
M18	Cayey
M19	Ceiba
M20	Ciales
M21	Cidra
M22	Coamo
M23	Comerío
M24	Corozal
M25	Culebra
M26	Dorado
M27	Fajardo
M28	Florida
M29	Guánica
M30	Guayama
M31	Guayanilla
M32	Guaynabo
M33	Gurabo
M34	Hatillo
M35	Hormigueros
M36	Humacao
M37	Isabela
M38	Jayuya
M39	Juana Díaz

Table D.3: Meaning of Municipality notation used in this document

Notation	**Meaning**
M40	Juncos
M41	Lajas
M42	Lares
M43	Las Marías
M44	Las Piedras
M45	Loíza
M46	Luquillo
M47	Manatí
M48	Maricao
M49	Maunabo
M50	Mayagüez
M51	Moca
M52	Morovis
M53	Naguabo
M54	Naranjito
M55	Orocovis
M56	Patillas
M57	Peñuelas
M58	Ponce
M59	Quebradillas
M60	Rincón
M61	Río Grande
M62	Sabana Grande
M63	Salinas
M64	San Germán
M65	San Juan
M66	San Lorenzo
M67	San Sebastián
M68	Santa Isabel
M69	Toa Alta
M70	Toa Baja
M71	Trujillo Alto
M72	Utuado
M73	Vega Alta
M74	Vega Baja
M75	Vieques
M76	Villalba
M77	Yabucoa
M78	Yauco