

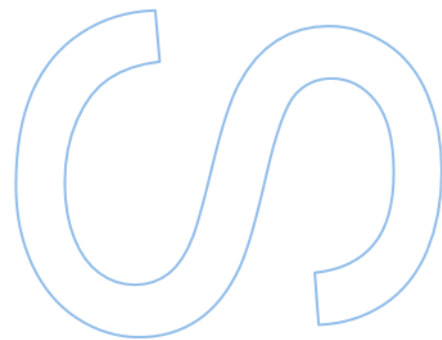
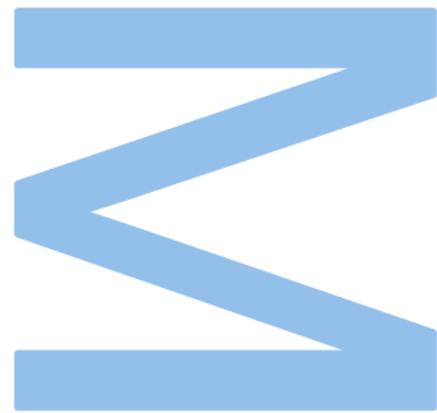
Unravelling Paediatric Hospitalizations: A New Perspective on Paths' Dissimilarities

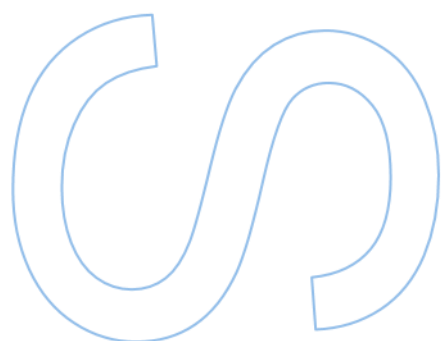
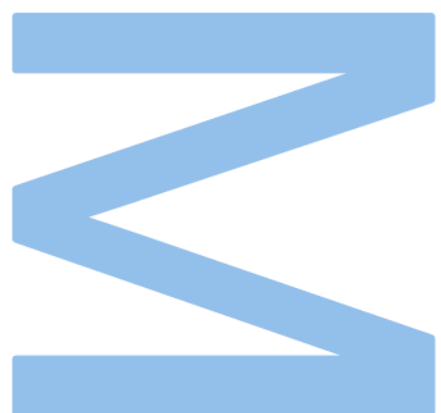
Daniel Cordeiro

Mestrado em Estatística Computacional e Análise de Dados
Departamento de Matemática da Faculdade de Ciências da Universidade
do Porto
2023

Orientador

Ana Rita Pires Gaio
Professora Auxiliar, Faculdade de Ciências da Universidade do Porto





“ For with God, nothing shall be impossible. ”

Luke 1:37

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Ana Rita Gaio, whose guidance and encouragement have been invaluable throughout the course of this research project. Her persistent efforts pushed me not only to excel in my report but also in various other academic pursuits.

I am also extremely thankful to Professor Ana Azevedo for her mentorship during my internship. Her kindness and support greatly contributed to my academic experience.

Special thanks go to Centro Hospitalar Universitário de São João for welcoming me into their institution and providing an environment conducive to research. I would like to extend my gratitude to the staff of Centro de Epidemiologia Hospitalar, whose cooperation and readiness to help have made this project possible.

To my friends, who have been a constant source of comfort, peace, and joy, I owe a debt of gratitude. Your encouragement has made this journey significantly more bearable.

I must also extend my heartfelt gratitude to Mafalda. Her consistent support throughout my master's journey has been indispensable. Always there to assist with any questions or challenges I faced, Mafalda has proven to be a steadfast friend on whom I know I can rely.

To my family, your unyielding faith in me has been my rock. Your endless love and support, especially from my mom, have given me the strength to persevere.

Finally, I owe my deepest gratitude to a very special person, Patricia. Your unwavering support in both good times and bad has been a source of constant inspiration. You have motivated me to become a better person and have been instrumental in helping me overcome my procrastination tendencies.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto

Departamento de Matemática

MSc. Computational Statistics and Data Analysis

Unravelling Paediatric Hospitalizations: A New Perspective on Paths' Dissimilarities

by [Daniel CORDEIRO](#)

The efficient management of patient flows in healthcare settings remains a complex and critical issue. This report aims to cluster hospitalization paths — characterized by a sequence of pairs consisting of physical services patients go through and their respective lengths of stay. This clustering will identify key patterns and offer insights into effective hospital resource allocation.

Initially, a comprehensive descriptive analysis was conducted to gain insights into the data landscape. Subsequently, we focused on modeling the length of stay for each service, considering factors such as Physical Service, Entry Reason (urgent/scheduled/transfer), Entry Day (weekend/weekday), Responsible Service (equal/different from physical service), and Hospital Team (CHUSJ/CMIN/HPH). General linear models with different residual variances per stratum of physical service were required. Our results indicate that the length of stay in all services was significantly higher than that of the Short-Term Unit. Factors like weekend admissions, being treated by the CMIN team as opposed to CHUSJ, and instances where the responsible service differed from the physical service were associated with a shorter length of stay.

Two clustering methodologies were explored: one based on hospital resource variables and another on novel dissimilarities between paths. They were described using hospital variables in table format, directed graphs, and Average Silhouette Width, and compared with the Adjusted Rand Index. Both approaches yielded well-defined clusters.

Although the findings are restricted to the pediatric ward of Centro Hospitalar Universitário de São João, they provide a foundational basis for future research and potential broader applications in healthcare settings. More generally, the developed dissimilarities could be applicable to any scenarios that require grouping sequences of paired data.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto

Departamento de Matemática

Mestrado em Estatística Computacional e Análise de dados

Decifrando Hospitalizações Pediátricas: Visão sobre Dissemelhanças entre Trajetória

por [Daniel CORDEIRO](#)

A gestão eficiente dos fluxos de pacientes em ambientes de saúde continua a ser uma questão complexa e crítica. Este relatório visa agrupar trajetos de hospitalização — caracterizados por uma sequência de pares que consistem em serviços físicos pelos quais os pacientes passam e as suas respetivas durações de estadia. Este agrupamento irá identificar padrões-chave e fornecer informações sobre a alocação eficaz de recursos hospitalares.

Inicialmente, foi realizada uma análise descritiva abrangente para obter informações sobre o panorama dos dados. Posteriormente, focámo-nos na modelação da estadia para cada serviço, considerando fatores como Serviço Físico, Motivo de Entrada (urgente/programada/transferência), Dia de Entrada (fim de semana/dia útil), Serviço Responsável (igual/diferente do serviço físico) e Equipa Hospitalar (CHUSJ/CMIN/HPH). Foram necessários modelos lineares gerais com diferentes variâncias dos resíduos por estrato de serviço físico. Os nossos resultados indicam que a estadia para todos os serviços foi significativamente superior à da Unidade de Curta Duração. Fatores como entrada ao fim de semana, ser tratado pela equipa CMIN em oposição à CHUSJ, e casos em que o serviço responsável diferia do serviço físico, estavam associados a uma menor estadia.

Foram exploradas duas metodologias de agrupamento: uma baseada em variáveis de recursos hospitalares e outra em dissemelhanças entre trajetos. Ambas as abordagens resultaram em grupos bem definidos, e foram descritas usando variáveis hospitalares, gráficos orientados, Largura Média de Silhueta, e com o Índice Rand Ajustado.

Embora as conclusões estejam limitadas à ala pediátrica do CHUSJ, elas fornecem uma base fundamental para futuras pesquisas e potenciais aplicações mais amplas em ambientes de saúde. De forma geral, as dissemelhanças desenvolvidas podem ser usadas em cenários que requerem agrupamento de sequências de dados emparelhados.

Contents

Acknowledgements	iii
Abstract	v
Resumo	vii
Contents	ix
List of Figures	xi
1 Introduction	1
1.1 Problem	1
1.2 Definitions	2
1.3 Data Origin	3
2 Exploratory Data Analysis	7
2.1 Data Manipulation and Cleaning	7
2.2 Descriptive Analysis	9
2.3 Length of Stay	12
3 Modelling Lengths of Stay	17
3.1 Understanding Generalized Least Squares	17
3.2 Problem	18
3.2.1 Data Preparation	18
3.2.2 Associations with Log(LoS)	18
3.2.3 Model Building	21
3.2.4 Model Selection	21
3.2.5 Model Evaluation	24
3.2.6 Multiple Comparisons	26
4 Clustering Hospital Resources	29
4.1 Problem	29
4.1.1 Metric vs Dissimilarity	30
4.1.2 Metric	30
4.1.3 Dissimilarity	31
4.1.4 Gower	32
4.1.4.1 Interval-Scaled, Ratio-Scaled and Ordinal Variables	33

4.1.4.2	Categorical Variables	33
4.1.4.3	Dichotomous Variables	33
4.1.4.4	The daisy function	34
4.2	Centroids-based Clustering (Partitioning methods)	35
4.3	Connectivity-based Clustering (Hierarchical clustering)	35
4.3.1	Linkage	36
4.3.2	Ward's Method	36
4.3.3	Lance-Williams	37
4.3.4	The hclust function	38
4.4	Distribution-based Clustering	39
4.5	Results and Discussion	39
5	Clustering Hospitalization Paths	47
5.1	Problem	47
5.2	Dissimilarities	48
5.2.1	Additive Dissimilarity	51
5.2.2	Product Dissimilarity	52
5.2.3	Jaccard-like Dissimilarity	52
5.2.4	Complex Dissimilarity	53
5.2.4.1	Service term	54
5.2.4.2	LoS term	55
5.2.4.3	LoS term - Equal Services	55
5.2.4.4	LoS term - Similar Services	55
5.2.4.5	Needleman term	57
5.3	Results and Discussion	59
6	Future Work	67
6.1	Time Series	67
6.2	Modelling of the Total Length of Stay	67
6.3	Markov Models	68
A	Algorithms	69
B	Dissimilarities Values	73
	Bibliography	79

List of Figures

1.1	Organogram	4
1.2	CHUSJ Map	5
2.1	Absolute Frequencies by Physical Service	8
2.2	Absolute Frequencies by Responsible Service	8
2.3	Barplot of the Number of Patients per number of Hospitalizations	9
2.4	Number of Paths with n Number of Physical Services	10
2.5	Barplot of Sex Distribution	11
2.6	Barplot of Age Category Distribution	12
2.7	Histogram of LoS Distribution, for each physical service	13
2.8	Left: Histogram of LoS of Nursery, Right: Histogram of LoS of Short-Term Unit	14
2.9	More frequent paths, corresponding to 80% of all episodes (LoS=Length of Stay (in hours); TLoS = Total Length of Stay(in hours))	14
2.10	More frequent paths, corresponding to 80% of all episodes excluding Most Frequent(LoS=Length of Stay; TLoS = Total Length of Stay)	15
2.11	Graphical representation of all 3133 paths	16
3.1	Boxplot of LoS per Physical Service	19
3.2	Boxplot of LoS per Left: Entry Reason; Center: Hospital Team; Right: Responsible Equal Physical Indicator	20
3.3	Boxplot of LoS per Far Left: Age Category; Center Left: Sex; Center Right: District; Far Right: Nationality	20
3.4	Boxplot of LoS per Left: Weekend Indicator; Right: Season	20
3.5	Residuals vs Fitted Values	24
3.6	Histogram of the Residuals	25
3.7	QQ-plot of the Residuals	25
3.8	Boxplot of the Residuals by Physical Service, Left: Different Error Variance Right: Same Error Variance	25
4.1	Average Silhouette Width using the Gower dissimilarity matrix	41
4.2	Dendrogram for the Gower dissimilarity matrix using Ward's method	41
4.3	AIC, BIC, and ICL for various number of components (clusters)	42
5.1	Min, Max and Range for each Service	49
5.2	Min, Max and Range for all paths	49
5.3	Graph of the paths X_{1097} (dark blue) and X_{1649} (orange)	50
5.4	Table of the paths X_{1097} (dark blue) and X_{1649} (orange)	50
5.5	Paths X_{1097} (dark blue) and X_{1649} with Std. LoS	50

5.6	Paths X_{1097} (dark blue) and X_{1649} (orange) with Std. LoS and equal length	51
5.7	Physical Services' Dissimilarity Matrix - D	54
5.8	The 7 clusters of the Additive Dissimilarity represented by directed graphs	63
5.9	The 7 clusters of the Jaccard_A Dissimilarity represented by directed graphs	64
5.10	The 7 clusters of the Complex Dissimilarity represented by directed graphs	65
B.1	Additive: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Righth: Within Cluster dissimilarity to medoid per number of cluster	74
B.2	Additive': Left: histogram of dissimilarity values; Center: ASW per number of cluster; Righth: Within Cluster dissimilarity to medoid per number of cluster	74
B.3	Product: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Righth: Within Cluster dissimilarity to medoid per number of cluster	75
B.4	Product': Left: histogram of dissimilarity values; Center: ASW per number of cluster; Righth: Within Cluster dissimilarity to medoid per number of cluster	75
B.5	Jaccard_A: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Righth: Within Cluster dissimilarity to medoid per number of cluster	76
B.6	Jaccard_P: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Righth: Within Cluster dissimilarity to medoid per number of cluster	76
B.7	Complex: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Righth: Within Cluster dissimilarity to medoid per number of cluster	77
B.8	Complex': Left: histogram of dissimilarity values; Center: ASW per number of cluster; Righth: Within Cluster dissimilarity to medoid per number of cluster	77

Chapter 1

Introduction

1.1 Problem

The effective management of pediatric wards is a crucial aspect of healthcare. At the Centro Hospitalar Universitário de São João (CHUSJ), the relocation of its pediatric ward on November 16, 2021, represents a significant opportunity. This change facilitates better space utilization thereby contributing to the improved functioning of pediatric care services at CHUSJ. The hospital staff has posed specific research queries aimed at enhancing its services and patient management. These include the following:

- How frequently do patients get hospitalized in the pediatric ward?
- What is the age distribution of these patients?
- Which physical services are most frequently utilized?
- What is the Length of Stay (LoS) distribution across different services, and what factors affect it?

The above questions led to a broader research problem: How can we holistically understand and optimize patient flow in the pediatric ward of CHUSJ to improve the quality of healthcare delivery? Additional sub-questions that emerged during the research include:

- How can graphical visualization techniques, like directed graphs, improve our understanding of hospitalizations?

- Can clustering methods based on resource utilization provide insights into patient paths?
- Will the introduction of novel dissimilarity metrics offer a more nuanced understanding of these patient paths?

To address these research questions, our study followed a two-phased approach. The first phase involved an exhaustive descriptive analysis, which also included the graphical visualization of hospitalizations through directed graphs, to answer the initial queries posed by the CHUSJ pediatric ward staff. In the second phase, we expanded our methodology to include generalized least squares (GLS) modeling, and clustering algorithms. This approach helped us to identify unique types of hospitalizations and to uncover patterns that could be critical for resource management.

Our research not only holds implications for the CHUSJ pediatric wards but also for hospital care in a broader context. We found that various socio-demographic and service-related factors differ between patient paths, drawing attention to areas that could benefit from equitable care strategies.

These findings are expected to contribute to best practices in pediatric care, affecting both hospital management strategies and resource allocation for improved patient outcomes. The novel approaches developed here for understanding patient paths offer potential for further research and broader application in other settings, healthcare, and others.

1.2 Definitions

Services Services can either be physical or responsible. Physical services refer to the actual location where a patient is housed, whereas responsible services determine the department (medical specialty) that oversees the patient's care. For instance, a patient might be assigned to the responsible service of "Pediatric Surgery" but be accommodated in the physical service location of "Pediatrics" due to possible space constraints in "Pediatric Surgery". Some responsible services, like "Neurosurgery" for example do not have a corresponding physical service location for children and pediatric patients overseen by these services are allocated to particular physical services, like "Pediatric Surgery" or "Pediatrics" in this case. If not specified otherwise, it is assumed that the term "service" refers to a physical service.

Episode An episode begins with a patient's admission to the hospital and concludes once he or she exits the hospital.

Path A path is composed of the physical services a patient goes through in a given episode and the respective time spent in each service.

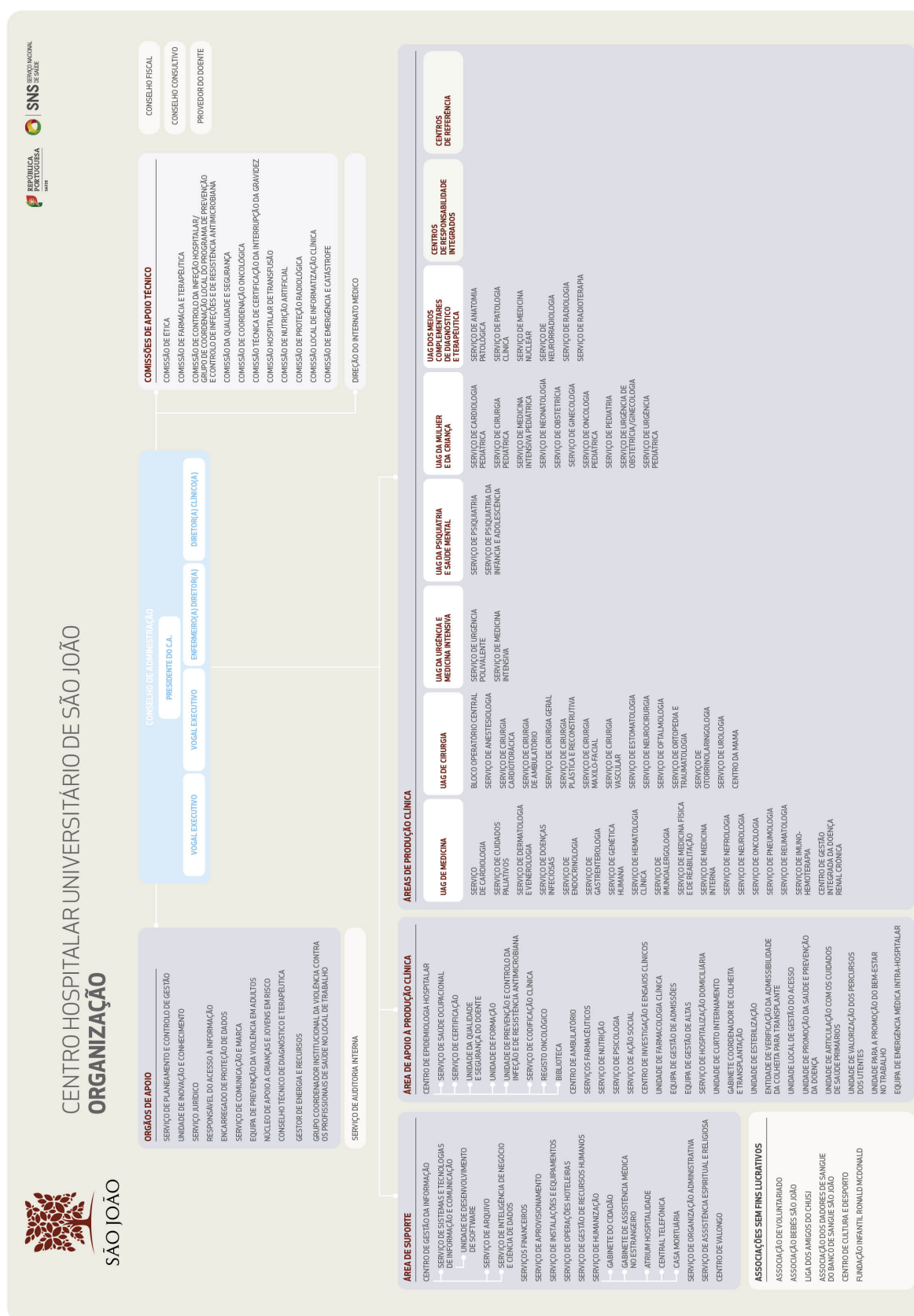
Simple and Compound Paths A path is said to be simple if it contains only one physical service, and compound otherwise (goes through more than a single physical service).

Types of Admission There are three types of hospital admissions: Normal, Urgent and Scheduled. "Normal" is used exclusively for newborns who have just been born in the hospital, indicating that their presence there is expected. The terms "Urgent" and "Scheduled" are self-explanatory, referring to urgent admissions and planned admissions, respectively.

Types of Exit There are three reasons for a hospital exit following the hospitalization: Discharge, Hospital, and Deceased. "Discharge" indicates that the patient was sent home or to a residential institution and may have a follow-up appointment in the future. "Hospital" means the patient has been transferred to another hospital. "Deceased" signifies that the patient has passed away.

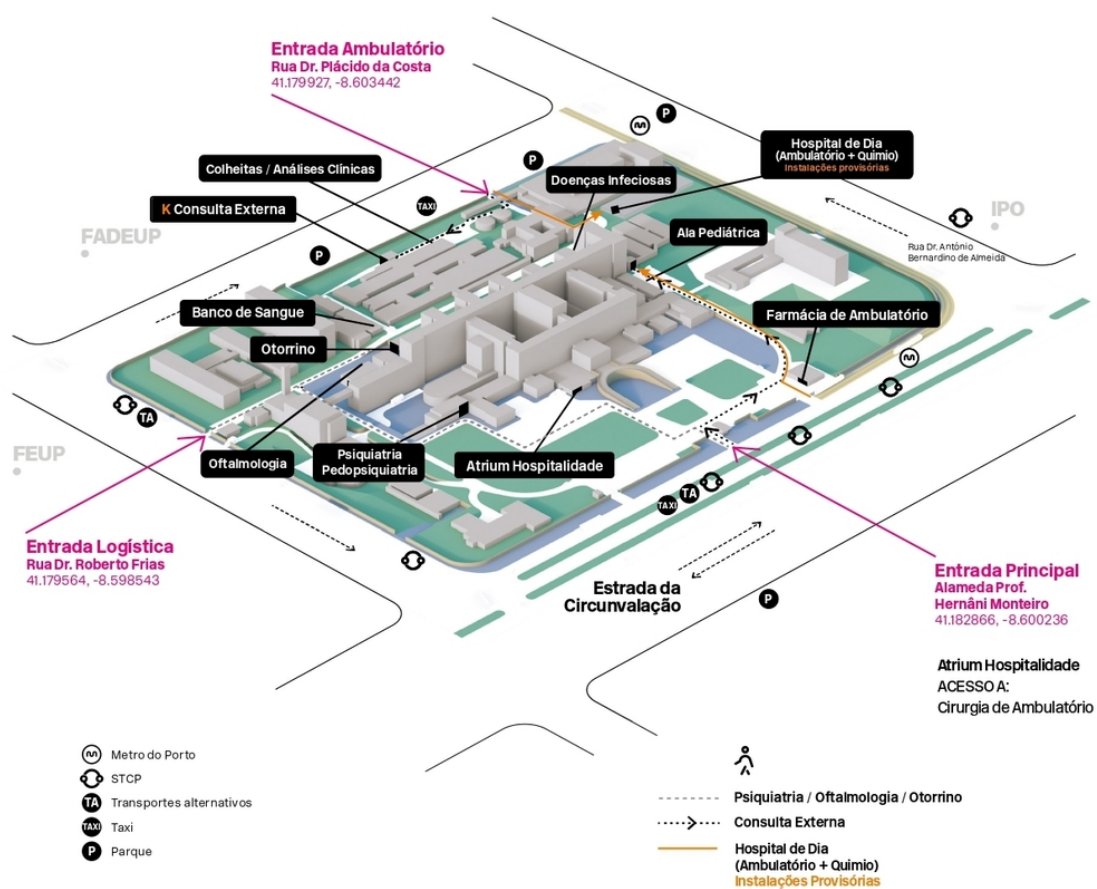
1.3 Data Origin

The data utilized in this analysis were sourced from the administrative database SONHO (Sistema Integrado de Informação Hospitalar), and were given to me in multiple password - encrypted CSV files to ensure confidentiality. Prior to any analytical procedures, all data were pseudonymized to uphold the privacy of the individuals involved. The hospitalizations included in this research belong to patients aged ≥ 18 years, from December 2021 to November 2022. Most of these occurred in services that pertain to the 'Unidade Autónoma de Gestão da Mulher e Criança' (UAG MC), providing a focused lens through which we evaluated patient paths within this particular segment of the healthcare facility. However, surgical departments which do not belong to this area but who are responsible for a relevant number of hospitalizations of children in the physical space of the pediatrics wards, such as for example "Neurosurgery", "Orthopedics" and "Otorhynolaryngology" were also considered.





K Consulta Externa
ACESSO A:
 • K2 - K13
 • Centro de Mama



Centro Hospitalar Universitário de São João
 Porto (Sede) Alameda Professor Hernâni Monteiro 4200-319 Porto T+ 351 225 512 100 geral@chsj.min-saude.pt www.chusj.pt

FIGURE 1.2: CHUSJ Map

Chapter 2

Exploratory Data Analysis

2.1 Data Manipulation and Cleaning

The data manipulation phase of this study presented significant challenges that required an extensive time commitment. Due to the complexity and nuanced characteristics of the paths dataset, a multitude of functions and iterative loops were developed in R to adequately reformat and clean the data for subsequent analysis. While this step was pivotal for ensuring the reliability of the findings, it was labor-intensive, consuming a considerable portion of the project timeline.

Numerous instances of incorrect data inputs were observed and subsequently eliminated. For example, within the category of ‘physical services,’ there were cases where the input did not actually correspond to a physical service but rather to a ‘responsible service’ like “Pediatric Cardiology”. Additionally, there were instances where the specified physical service was not part of the pediatric ward, among other inconsistencies. These entries were removed based on the medical staff’s decision to exclude them, assuming they were errors. Out of the 7320 instances in which services were used, 69 instances were from services that were designated for removal. For our analyses, we focused solely on the first 9 out of 27 physical services—those that were most frequently used and correctly entered in Figure 2.1: “Nursery”, “Pediatric Surgery”, “Pediatrics”, “Short-Term Unit”, “Pediatric Intensive Care Medicine”, “Neonatology”, “Pediatric Oncology”, “Otorhinolaryngology”, “Cardiothoracic Surgery”.

Recem Nascidos (Berc) - Obst.	Cirurgia Pediatrica	Pediatria
2086	1449	1130
Unidade De Curta Duracao - Pediatria	M.Intensiva Pediatrica	Neonatologia Internamento
1101	414	352
Oncologia Pediatrica	Otorrinolaringologia	Cirurgia Cardiotoracica
329	310	80
Cirurgia Plastica E Reconst.	Obstetricia Internamento	Infecologia
18	14	7
Neurologia - Ume	Cirurgia Geral Internamento	Urologia Internamento
5	3	3
Cardiologia Internamento	Cirurgia Vascular	Ginecologia Internamento
2	2	2
M.Intensiva Neurocriticos	Medicina Homens	Ortopedia
8	2	2
Unidade Hospitalizacao Domiciliar	Cardiologia Pediatrica	Gastroenterologia
2	1	1
M.Intensiva	Neurocirurgia	Oftalmologia
1	1	1

FIGURE 2.1: Absolute Frequencies by Physical Service

The same approach was applied to the responsible services, where we selected the first 12 out of 30 services in Figure 2.2, based on frequency and services suitable for individuals under 18 years, as we did for the physical services: "Nursery", "Short-Term Unit", "Pediatric Surgery", "Pediatrics", "Pediatric Oncology", "Pediatric Intensive Care Medicine", "Otorhinolaryngology", "Neonatology", "Orthopedics", "Pediatric Cardiology", "Pediatric Neurosurgery", "Cardiothoracic Surgery".

Recem Nascidos (Berc) - Obst.	Unidade De Curta Duracao - Pediatria	Cirurgia Pediatrica
2087	1088	757
Pediatria	Oncologia Pediatrica	M.Intensiva Pediatrica
743	445	412
Otorrinolaringologia	Neonatologia Internamento	Ortopedia
369	349	340
Cardiologia Pediatrica	Neurocirurgia Pediatrica	Cirurgia Cardiotoracica
240	211	109
Oftalmologia	Cirurgia Plastica E Reconst.	Int Estomatologia
37	31	29
Int Cirurgia Maxilofacial	Obstetricia Internamento	Cirurgia Vascular
27	14	7
Neurologia - Ume	Cirurgia Geral Internamento	Medicina Homens
5	3	3
Cardiologia Internamento	Ginecologia Internamento	M.Intensiva Neurocriticos
2	2	2
Unidade Hospitalizacao Domiciliar	Urologia Internamento	Infecologia
2	2	1
M.Intensiva	Neurocirurgia	Neurologia Adultos
1	1	1

FIGURE 2.2: Absolute Frequencies by Responsible Service

We then selected only the paths for which all physical and responsible services encountered were among those we had chosen to focus on. This approach resulted in a total of 5521 paths.

If an admission or exit was labeled as 'Transfer,' this indicated that a path was truncated. In other words, the patient was already in the hospital before the starting date of our dataset, or had not yet left the hospital by the ending date. All paths with an admission or exit labeled as 'Transfer' were excluded from our analysis, bringing the total number of paths down to 5450.

Upon presenting the data to the medical staff, they identified a potential error in the exit variable. They requested the update of the variable's values using another dataset,

which was deemed to be accurate. As the two datasets contained differing information about the paths, merging them resulted in a reduced total number of paths to 5404.

Among those 5404 paths, some belong to the same patient, as any given patient could have experienced more than one hospitalization within a year. In total, there are 4760 unique patients represented in these paths.

2.2 Descriptive Analysis

Bellow is the bar plot of hospitalization frequencies, delineating the number of patients who experienced a single hospitalization, two hospitalizations, and so forth.

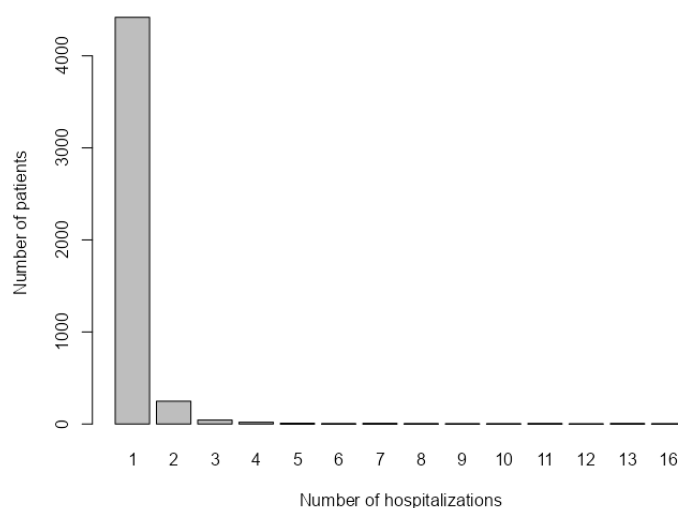


FIGURE 2.3: Barplot of the Number of Patients per number of Hospitalizations

and the respective table:

Hospitalizations	1	2	3	4	5	6	7	8	9	10	11	12	13	16
Patients	4418	248	43	19	6	3	5	3	2	2	4	1	4	2

TABLE 2.1: Table of the Number of Patients per Number of Hospitalizations

In the dataset, 4418 (92.8%) of patients were hospitalized only once, 248 (5.2%) were hospitalized twice, and 94 (1.9%) experienced more than two hospitalizations.

During our analysis of graphs like Figure 2.11, we observed instances where a patient was transferred from a physical service to the same physical service. In such cases, we opted to consolidate these multiple entries into a single visit and summed the respective

LoS. For example, if a patient was moved to *PedOnc* for 5 hours and subsequently transferred again to *PedOnc* for an additional 2 hours, we altered the data to indicate a single visit to *PedOnc* for a total of 7 hours.

A bar plot and table of the frequencies of the number of services used for the paths were as follows:

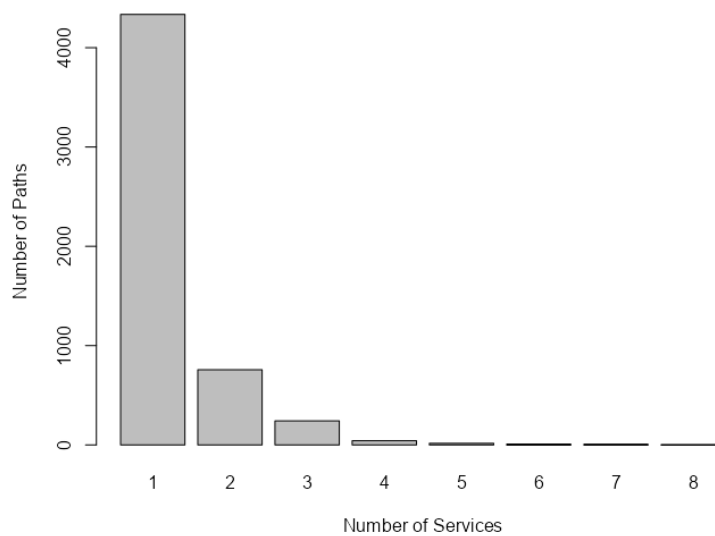


FIGURE 2.4: Number of Paths with n Number of Physical Services

Services	1	2	3	4	5	6	7	8
Path	4335	757	242	42	16	6	5	1

TABLE 2.2: Physical Services Frequency

Among all paths, 4335 (80.2%) were simple and the remaining 1069 (19.8%) were compound paths that included visits to multiple physical services. Upon further breakdown, 757 (14%) went through two physical services, while 312 (5.8%) involve visits to more than two physical services.

Out of the 4335 simple paths, 1811 (or 41.8%) were attributed to newborns who passed only through the nursery. The average LoS for this service in these paths was 71 hours, with a median LoS of 65 hours. When considering both admission and exit, 1800 of these paths had a 'Normal Admission', passed through the nursery, and were followed by discharge. The remaining 11 paths involved different types of admissions and exits.

Out of the 4760 unique patients represented in these paths, 4757 had complete information regarding their sex, date of birth, and region of residence. Of these, 2738 (57.6%) were male.

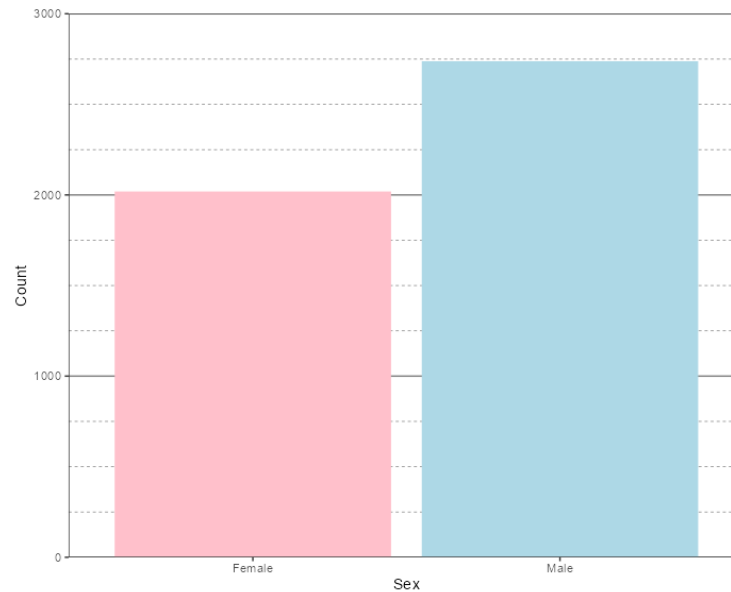


FIGURE 2.5: Barplot of Sex Distribution

The age category for each patient was determined by calculating the time difference between their date of birth and their admission date to the hospital. Notably, some patients experienced multiple hospitalizations and could, as a result, belong to different age categories for different paths. This situation occurred 73 times, meaning that 73 patients transitioned from one age category to another during separate hospitalizations.

The age categories were established by the medical staff and corresponded to the following cutoffs:

- ≤ 30 days : Newborn
- $]30 \text{ days} - 1 \text{ year}[$: Baby
- $[1 \text{ year} - 5 \text{ years}[$: Toddler
- $[5 \text{ years} - 11 \text{ years}[$: Child
- ≥ 11 years : Teenager

Of the 73 patients who experienced more than one hospitalization within the year under study and had a change in their age category, 53 transitioned from 'Newborn' to 'Baby', 8 from 'Baby' to 'Toddler', 9 from 'Toddler' to 'Child', and 3 from 'Child' to 'Teenager'.

Excluding these 73 patients, the remaining patients had the following age category distribution:

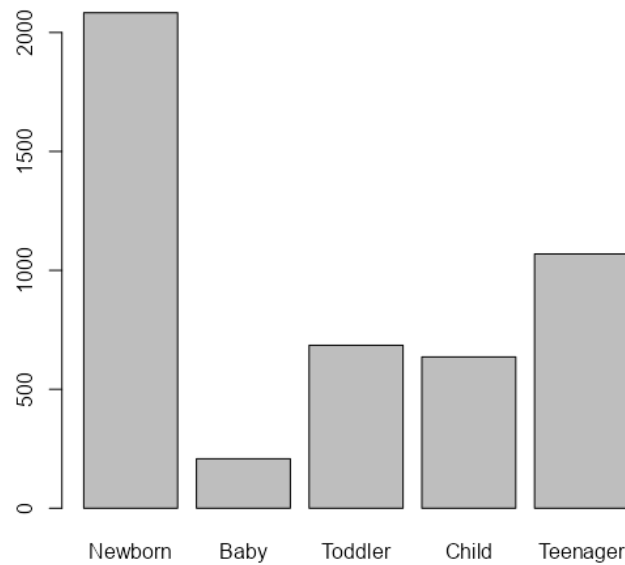


FIGURE 2.6: Barplot of Age Category Distribution

2.3 Length of Stay

Length of Stay (LoS) is a key variable in our analysis, and we have conducted an extensive investigation into its various aspects, particularly its empirical distribution across all physical services.

All LoS distributions of the physical services were right-skewed, regardless of the service itself. This skewness translates to the fact that most patients experienced relatively short times at each service, while a smaller number encountered considerably longer durations. This pattern has important implications for resource allocation and patient flow management within the hospital setting, as it indicates that while the majority of cases may be handled expeditiously, there exists a subset of cases that require an extended period of engagement, potentially straining the available resources.

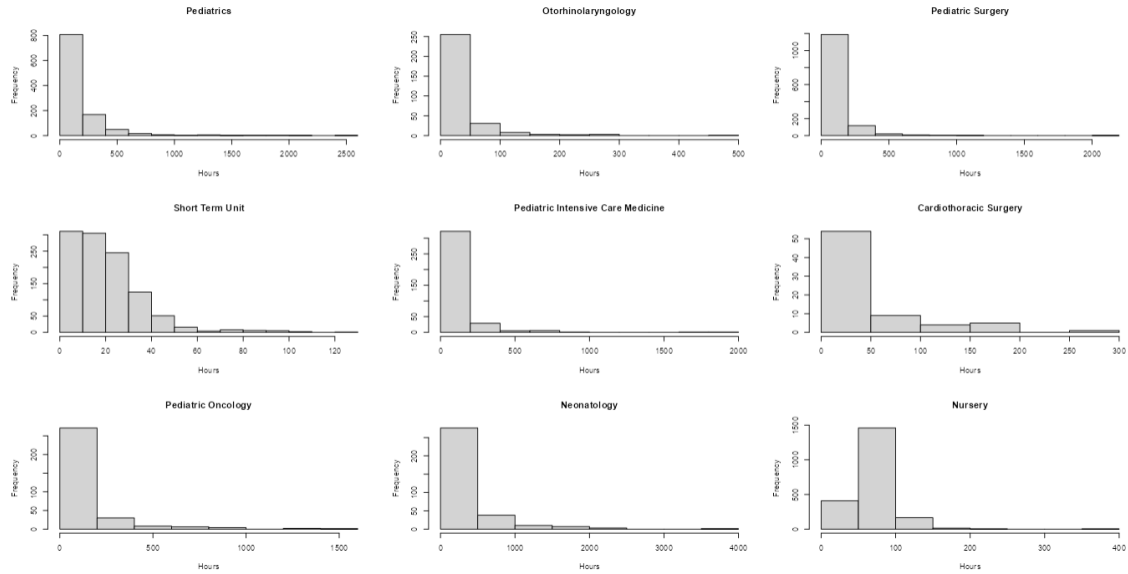


FIGURE 2.7: Histogram of LoS Distribution, for each physical service

In the course of analyzing LoS distributions for the various hospital services, distinctive patterns emerged. For example, the Short-Term Unit (STU) had a clear-cut LoS distribution that aligned with its design to provide rapid treatment; 70% of cases registered stays under 24 hours, and 95% exited within 48 hours. In the Nursery, it was observed that 80% of all instances had a LoS ranging between 2 to 5 days, aligning closely with the hospital's common practice that newborns generally stay for a minimum of 48 hours. Notably, the Nursery data exhibited an initial spike in instances within the first 4 hours, after that, there were virtually no instances in the Nursery until the 48-hour mark. The episodes that occurred in this initial 4-hour window likely indicated complications with the newborns that necessitated their immediate transfer to specialized services for further care. It is worth mentioning that the only newborn cases included in our posterior clustering analysis pertained to patients with compound paths. These unique LoS distribution patterns not only validate the role and function of each service but also highlight the nuances that exist in patient flow through these services.

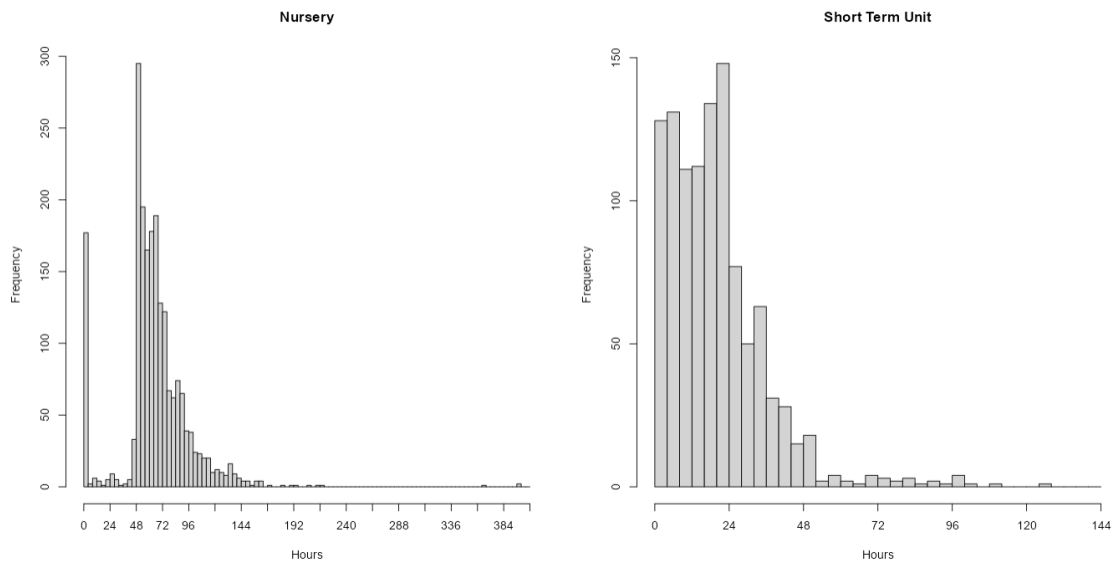


FIGURE 2.8: Left: Histogram of LoS of Nursery, Right: Histogram of LoS of Short-Term Unit

In the dataset under study, it is noteworthy that a high degree of concentration exists among the observed hospital paths, indicating a certain level of uniformity in service utilization. Specifically, 80% of all episodes can be encompassed by just 10 paths, including both the admission and exit types as part of the path.

Admission	Service1	Service2	Service3	Discharge	Mean LoS 1	Mean LoS 2	Mean LoS 3	Count	CumSum	Mean TLoS
Normal	NUR	NA	NA	Discharge	71	NA	NA	1800	33,31%	71
Scheduled	PedSur	NA	NA	Discharge	69	NA	NA	490	42,38%	69
Urgent	PED	NA	NA	Discharge	172	NA	NA	424	50,22%	172
Urgent	STU	NA	NA	Discharge	21	NA	NA	401	57,64%	21
Urgent	STU	PedSur	NA	Discharge	14	93	NA	335	63,84%	107
Scheduled	ORL	NA	NA	Discharge	34	NA	NA	241	68,30%	34
Urgent	PedSur	NA	NA	Discharge	142	NA	NA	239	72,72%	142
Scheduled	PedOnc	NA	NA	Discharge	98	NA	NA	186	76,17%	98
Urgent	STU	NA	NA	Hospital	23	NA	NA	160	79,13%	23
Scheduled	PED	NA	NA	Discharge	85	NA	NA	152	81,94%	85

FIGURE 2.9: More frequent paths, corresponding to 80% of all episodes (LoS=Length of Stay (in hours); TLoS = Total Length of Stay(in hours))

However, when we isolate the dataset to remove the simple newborn paths, accounting for 1800 out of 5404 total episodes, the concentration changes marginally but still remains high. Now, 80% of the remaining 3604 episodes are captured by 12 paths. This finding highlights the presence of standard routes that patients frequently navigate through the hospital. These pathways could be the focal points for optimizing resource allocation and patient flow within the institution.

Admission	Service1	Service2	Service3	Discharge	Mean LoS 1	Mean LoS 2	Mean LoS 3	Count	CumSum	Mean TLoS
Scheduled	PedSur	NA	NA	Discharge	69	NA	NA	490	13,60%	69
Urgent	PED	NA	NA	Discharge	172	NA	NA	424	25,36%	172
Urgent	STU	NA	NA	Discharge	21	NA	NA	401	36,49%	21
Urgent	STU	PedSur	NA	Discharge	14	93	NA	335	45,78%	107
Scheduled	ORL	NA	NA	Discharge	34	NA	NA	241	52,47%	34
Urgent	PedSur	NA	NA	Discharge	142	NA	NA	239	59,10%	142
Scheduled	PedOnc	NA	NA	Discharge	98	NA	NA	186	64,26%	98
Urgent	STU	NA	NA	Hospital	23	NA	NA	160	68,70%	23
Scheduled	PED	NA	NA	Discharge	85	NA	NA	152	72,92%	85
Normal	NUR	NEO	NA	Discharge	8	414	NA	125	76,39%	422
Urgent	STU	PED	NA	Discharge	28	187	NA	97	79,08%	215
Normal	NUR	NEO	PED	Discharge	11	301	228	52	80,52%	540

FIGURE 2.10: More frequent paths, corresponding to 80% of all episodes excluding Most Frequent(LoS=Length of Stay; TLoS = Total Length of Stay)

The following graph (Figure 2.11) provides a comprehensive visualization of all episodes that will be included in the clustering analysis of Chapters 4 and 5. This graphical representation serves as an essential baseline, offering insights into the distribution and characteristics of the patient paths analyzed. By including only the episodes that meet specific criteria, the graph ensures that the subsequent clustering analysis is both relevant and robust. The visualization offers an initial glimpse into the heterogeneity or homogeneity of the episodes, thus setting the stage for the in-depth analysis that will follow.

In the visual representation of patient flow within the Pediatric Ward, various colored arrows and nodes serve distinct but interconnected roles. The green arrows specifically connect the point of admission to the first service that a patient encounters in his/her path. Conversely, the red arrows indicate the transition from the last service a patient visits to the exit type. Light grey arrows serve to connect one physical service to another, forming the intricate network that constitutes a patient's complete episode within the hospital.

The orientation of these arrows is deliberately designed to indicate the direction of the patients' movement — from which service they originated and to which they proceeded. Additionally, the width of the arrows serves as a visual cue for the volume of episodes that transition from one specific node to another. Correspondingly, the number in the nodes and their size indicates the frequency with which a particular service, admission or exit was engaged.

This method of graphical representation not only provides an intuitive way to grasp the complexities and nuances of patient flow, as well as the utilization of services, and the transitions that occur between different stages of a hospital visit, but it also enables immediate visual identification of high-traffic areas and bottlenecks within the system. As

a result, this visualization offers valuable insights into resource allocation and operational efficiencies, serving as a powerful tool for both analysis and decision-making.

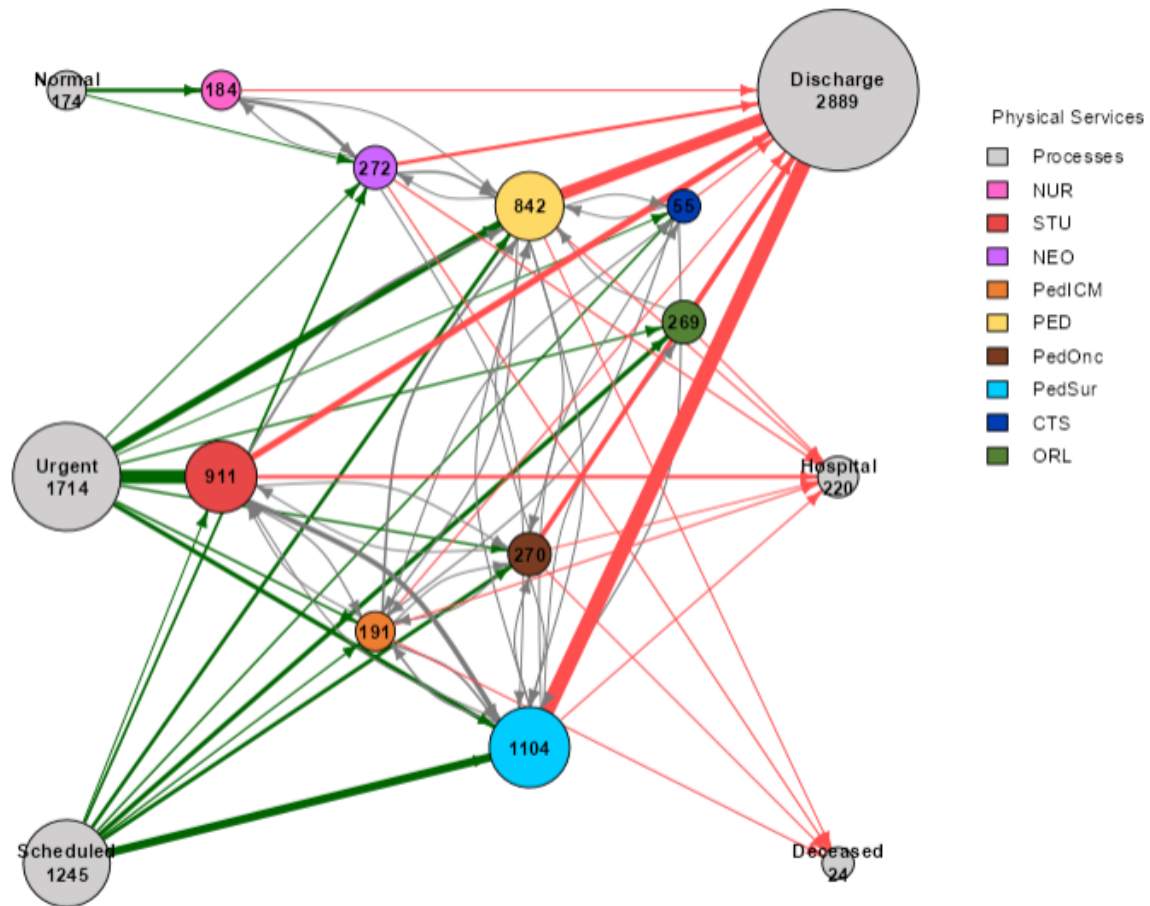


FIGURE 2.11: Graphical representation of all 3133 paths

Chapter 3

Modelling Lengths of Stay

3.1 Understanding Generalized Least Squares

The Generalized Least Squares (GLS) method is an extension of the Ordinary Least Squares (OLS) method, primarily used for linear regression analysis. Both methods aim to find the best-fitting line through the observed data by minimizing a certain sum of squares of residuals[1] [2] [3].

In a regression setting, while OLS coincides with the maximum likelihood estimates whenever the errors are independent and follow a normal distribution with constant variance, GLS provides maximum likelihood estimates whenever the independence and/or homoscedasticity condition are relaxed.

In particular, in GLS, the error terms are transformed based on a specified variance-covariance matrix, which can account for various forms of heteroscedasticity or correlation. This is particularly useful when dealing with heteroscedastic data where the variability of the dependent variable varies across levels of the independent variables, or is dependent on the response variable, for example. This makes GLS a more flexible and robust approach when dealing with complicated datasets, where OLS assumptions may not apply.

GLS models can be implemented in various programming languages, including R, Python, and others. In R, the 'nlme' package provides suitable functions to create GLS models and is often preferred for its flexibility and ease of use.

3.2 Problem

The objective of this analysis was to devise a prediction model for the LoS of a patient in a particular physical service, informed by data available at the point of patient admission. The initial dataset incorporated various features, including patient's physical service, reason for entrance into the service, season of admission, sex, age group, hospital team on duty, district, and nationality. However, not every data point was relevant for analysis. An initial data cleaning process was thus required to improve the predictive model's accuracy and efficiency.

3.2.1 Data Preparation

We restricted the analysis to prevalent patient episodes mirrored in other smaller analyses done in the internship. Episodes with a LoS equating 0 hours (≤ 30 minutes) were identified as outliers and eliminated from the dataset. The primary reason for this decision was the significant skewness in the LoS empirical distribution, and the subsequent requirement for a logarithmic transformation and the meaninglessness of a 0 LoS for our purposes.

We initially contemplated excluding newborn patients with no complications, who only passed through the Nursery. Yet, removing these cases could unintentionally introduce additional information for the "Nursery" category. To avoid such bias, all patients classified as "Newborn" were removed, including one patient who was not a newborn but was accommodated in the Nursery.

12 patients associated with the Neonatology Service, that were not newborns, were also excluded owing to their scarce occurrence.

3.2.2 Associations with Log(LoS)

The variables chosen to elucidate the variation in Length of Stay (LoS) encompass the following:

- **Physical:** The physical service itself.
- **Entry Reason:** Reason for entry into the service.
- **Team_Hospital:** The hospital team in charge at the time of entry into the service.

- **R_equal_P:** An indicator of whether the responsible service coincided with the physical service, where 1 denotes yes.
- **Age_Category:** Categorical representation of the patient's age.
- **Sex:** Sex of the patient.
- **District:** District of residence of the patient.
- **Nationality:** Country of birth of the patient.
- **Weekend:** Indicator of whether the service entry commenced over the weekend, where 1 denotes yes.
- **Season:** The season categorized as Spring_Summer or Autumn_Winter.

Represented bellow are the selected variables and their action on the response variable LoS.

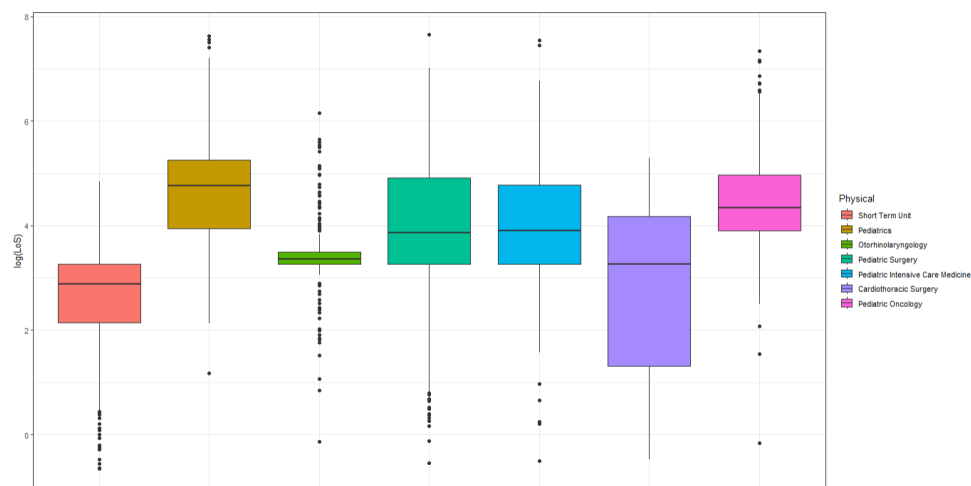


FIGURE 3.1: Boxplot of LoS per Physical Service

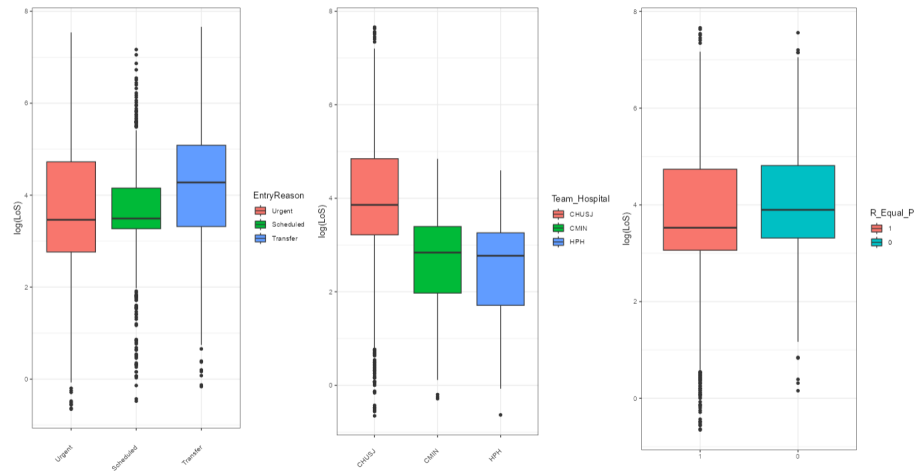


FIGURE 3.2: Boxplot of LoS per Left: Entry Reason; Center: Hospital Team; Right: Responsible Equal Physical Indicator

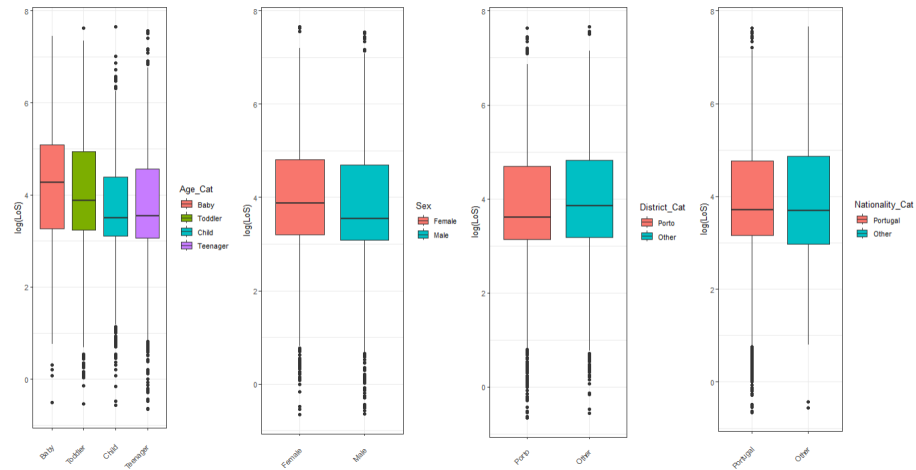


FIGURE 3.3: Boxplot of LoS per Far Left: Age Category; Center Left: Sex; Center Right: District; Far Right: Nationality

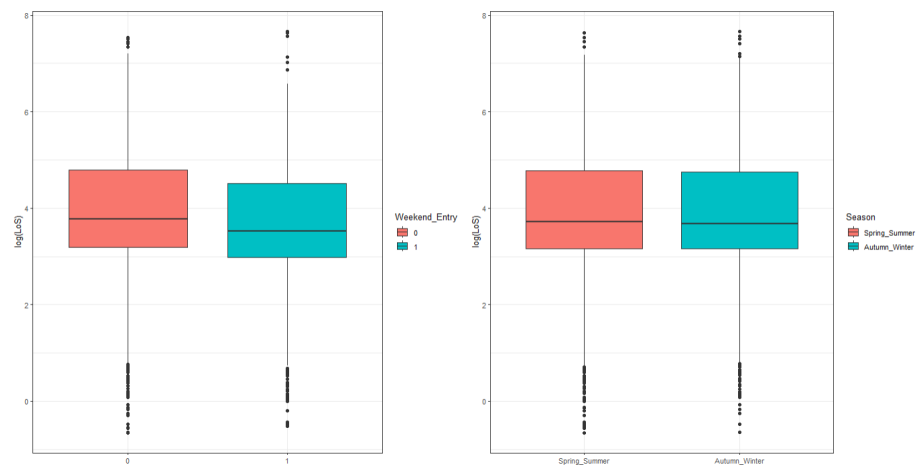


FIGURE 3.4: Boxplot of LoS per Left: Weekend Indicator; Right: Season

3.2.3 Model Building

The fitting of general linear models used the `nlme` package in R. The dependent variable in this model, the patient LoS, was log-transformed in order to obtain an approximately symmetric distribution for the residuals. We also applied a variance transformation function that permitted different variances per stratum of the "Great Group of Diagnosis" hence acknowledging the heteroscedasticity in the data after adequate residuals graphical analyses.

3.2.4 Model Selection

For model comparison, we used the likelihood ratio test with: (1) maximum likelihood (ML) estimation for nested models with the same variance structure and nested linear predictors, (2) restricted maximum likelihood (REML) for nested models with nested variance structures and the same linear predictor. This is because estimation by REML is known to return less biased variance component estimators than estimation by ML. Non-nested models were compared using Akaike's Information Criterion (AIC), where lower values are indicative of a model that better balances fit and complexity.

The final model, referred to as `mod.final`, had the lowest AIC and incorporated the following features: reason for admission, season of admission, Great Group of Diagnosis, a dummy variable representing if the Responsible Service was equal to the Physical Service, a weekend admission indicator, and the hospital team assigned to the day the patient entered the Physical Service:

$$\begin{aligned}
 \log(\text{LoS}) = & \beta_0 + \beta_1 \times \text{Physical}_{\text{Short Term Unit}} + \beta_2 \times \text{Physical}_{\text{Pediatrics}} \\
 & + \beta_3 \times \text{Physical}_{\text{Otorhinolaryngology}} + \beta_4 \times \text{Physical}_{\text{Pediatric Intensive Care Medicine}} \\
 & + \beta_5 \times \text{Physical}_{\text{Cardiothoracic Surgery}} + \beta_6 \times \text{Physical}_{\text{Pediatric Oncology}} \\
 & + \beta_7 \times \text{Reason}_{\text{Scheduled}} + \beta_8 \times \text{Reason}_{\text{Transfer}} \\
 & + \beta_9 \times \text{R_Equal_F}_{\text{No}} + \beta_{10} \times \text{Weekend_Entry}_{\text{Yes}} \\
 & + \beta_{11} \times \text{Team}_{\text{CMIN}} + \beta_{12} \times \text{Team}_{\text{HPH}} + \epsilon
 \end{aligned} \tag{3.1}$$

Where:

- $\beta_0, \beta_1, \dots, \beta_{12}$ are the regression parameters.
- $\text{Physical}_{\text{Short Term Unit}}, \text{Physical}_{\text{Pediatrics}}, \dots, \text{Team}_{\text{HPH}}$ are dummy variables.

The reference categories for the dummy variables used in the models were as follows:

Variable	Reference Category
Physical	PedSur
Reason of Entry	Urgent
Responsible = Physical	Yes
Weekend Entry	No
Team	CHUSJ

TABLE 3.1: Reference Categories for the dummy variables entering the final model

Error term:

$$\epsilon_i = \sigma_{\text{Physical}_i} \times Z_i$$

Where:

- ϵ_i is the error term for the i -th observation.
- $\sigma_{\text{Physical}_i}$ is the standard deviation associated with the level Physical_i of the variable *Physical*.
- Z_i is a standard normally-distributed random variable $Z_i \sim \mathcal{N}(0, 1)$.

In the model there are only significant variables, and the estimates obtained for the model parameters in the table below:

Variable	Value	Std.Error	p-value
(Intercept)	4.45	0.04	0.00
<i>Physical</i>			
Short Term Unit	-1.70	0.06	0.00
Pediatrics	0.51	0.04	0.00
Otorhinolaryngology	-0.45	0.05	0.00
Pediatric Intensive Care Medicine	-0.23	0.07	0.00
Cardiothoracic Surgery	-1.08	0.20	0.00
Pediatric Oncology	0.44	0.06	0.00
<i>EntryReason</i>			
Scheduled	-0.65	0.04	0.00
Transfer	-0.16	0.04	0.00
<i>Responsible = Physical</i>			
No	-0.19	0.04	0.00
<i>Weekend_Entry</i>			
Yes	-0.09	0.04	0.01
<i>Team_Hospital</i>			
CMIN	-0.12	0.06	0.05
HPH	-0.18	0.10	0.06

TABLE 3.2: Final Model Fixed Parameters

Physical Service	Value
Pediatrics	1.000
Otorhinolaryngology	0.782
Pediatric Surgery	1.228
Short Term Unit	1.097
Pediatric Intensive Care Medicine	1.305
Cardiothoracic Surgery	1.849
Pediatric Oncology	1.078

TABLE 3.3: Final Model Variance Parameters

In the context of this report, the Table 3.3 the parameters serve as a quantitative summary of the variance in different Physical Services, as compared to a designated reference group—the Pediatrics (Ped). Specifically, the Ped is assigned a baseline variance of one, against which the variances of other services are scaled. The parameter estimates for each service represent a multiplier of this baseline variance. For instance, the Short Term Unit has a parameter estimate of 1.097, indicating that its variance is identical to that of the Ped. Conversely, the parameter estimate for Cardiothoracic Surgery is 1.849, suggesting its variance is approximately 85% higher than that of the Ped. These estimates offer a comparative understanding of how variance fluctuates across physical services, thereby providing valuable insights into the heterogeneity of the data.

3.2.5 Model Evaluation

Residual diagnostics were conducted to assess the assumptions underlying the final model. We plotted the residuals against the fitted values in Figure 3.5 and examined the distribution of the residuals using a histogram, Figure 3.6, and a QQ-plot, Figure 3.7.

While the QQ-plot was not totally consistent with a normal distribution, the histogram exhibits a fairly symmetric distribution which does not compromise the inference results from the regression model. The plot of residuals against predicted values indicated an approximate constant variance across the range of the data. Taken together, these graphical diagnostics support the choice of this final model.

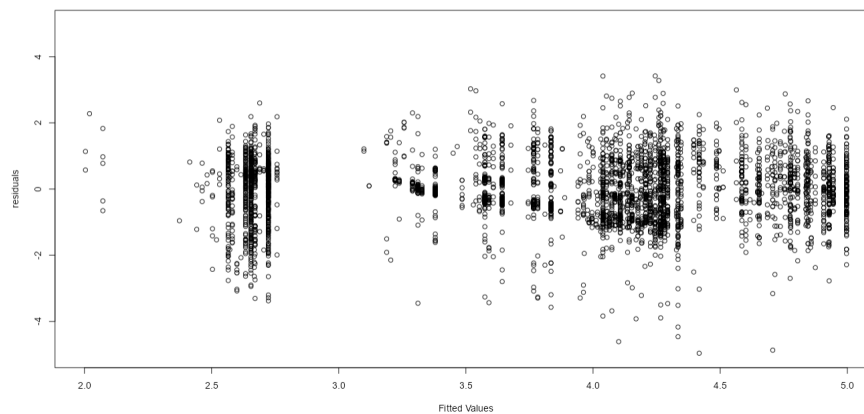


FIGURE 3.5: Residuals vs Fitted Values

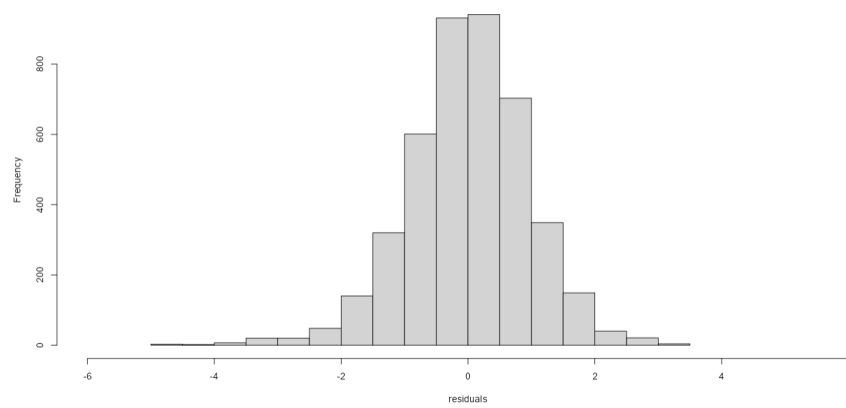


FIGURE 3.6: Histogram of the Residuals

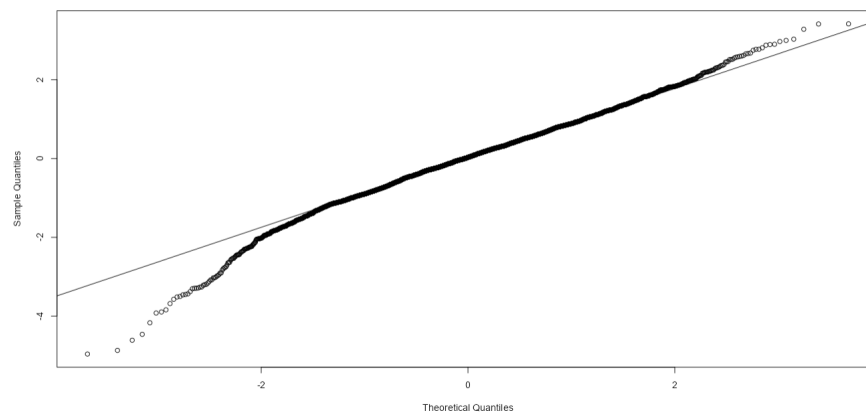
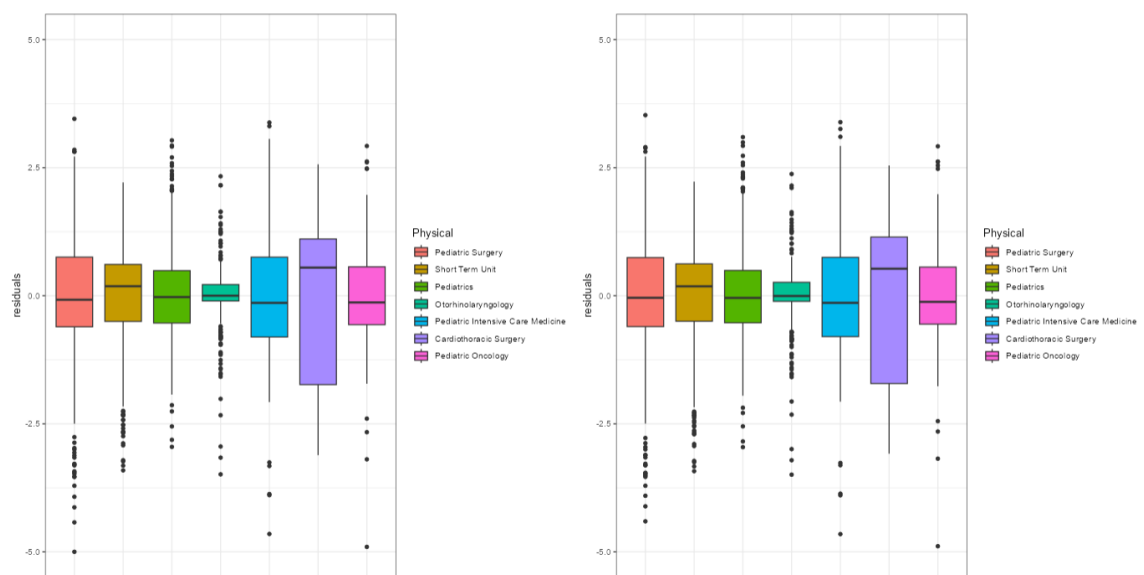


FIGURE 3.7: QQ-plot of the Residuals

FIGURE 3.8: Boxplot of the Residuals by Physical Service, Left: Different Error Variance
Righth: Same Error Variance

Despite what the graphical illustrations might imply, there are noteworthy differences between `final.model`, which accounts for varying error variances across strata of physical services, and `final.model.2`, which does not. Specifically, `final.model` outperforms its counterpart in several key metrics: it has a lower AIC score, the Likelihood Ratio Test using REML was significant, meaning that the simpler model is not sufficient to explain the variation in the data, and has a residual standard error of 0.845 as opposed to 0.955 for `final.model.2`.

3.2.6 Multiple Comparisons

We used the Tukey's method to establish pairwise comparisons between mean $\text{Log}(LoS)$ of the several Physical Services. This is a multiple comparison method that controls the family-wise error rate, ie, the probability of making one or more false discoveries in the multiple tests[4].

The only two comparisons that were not statistically significant were among Pediatrics and Pediatric Oncology, and among Otorhinolaryngology and Pediatric Intensive Care Medicine Table 3.4. Comparisons regarding Cardiothoracic Surgery exhibited a high standard error due to its infrequent utilization.

Contrast	Estimate	SE	p-value
Pediatric Surgery - Short Term Unit	1.699	0.058	< 0.001
Pediatric Surgery - Pediatrics	-0.513	0.041	< 0.001
Pediatric Surgery - Otorhinolaryngology	0.453	0.055	< 0.001
Pediatric Surgery - Pediatric Intensive Care Medicine	0.226	0.069	0.020
Pediatric Surgery - Cardiothoracic Surgery	1.078	0.196	< 0.001
Pediatric Surgery - Pediatric Oncology	-0.444	0.062	< 0.001
Short Term Unit - Pediatrics	-2.213	0.053	< 0.001
Short Term Unit - Otorhinolaryngology	-1.246	0.065	< 0.001
Short Term Unit - Pediatric Intensive Care Medicine	-1.473	0.077	< 0.001
Short Term Unit - Cardiothoracic Surgery	-0.622	0.200	0.039
Short Term Unit - Pediatric Oncology	-2.144	0.071	< 0.001
Pediatrics - Otorhinolaryngology	0.966	0.056	< 0.001
Pediatrics - Pediatric Intensive Care Medicine	0.739	0.069	< 0.001
Pediatrics - Cardiothoracic Surgery	1.591	0.196	< 0.001
Pediatrics - Pediatric Oncology	0.069	0.063	0.928
Otorhinolaryngology - Pediatric Intensive Care Medicine	-0.227	0.077	0.051
Otorhinolaryngology - Cardiothoracic Surgery	0.625	0.196	0.033
Otorhinolaryngology - Pediatric Oncology	-0.897	0.065	< 0.001
Pediatric Intensive Care Medicine - Cardiothoracic Surgery	0.851	0.203	0.001
Pediatric Intensive Care Medicine - Pediatric Oncology	-0.670	0.081	< 0.001
Cardiothoracic Surgery - Pediatric Oncology	-1.522	0.199	< 0.001

TABLE 3.4: Contrast Estimates with Standard Errors and p-values

Chapter 4

Clustering Hospital Resources

4.1 Problem

In this report, the complexity of patient episodes in the hospital was initially visualized through a comprehensive graph. The graph's intricacy prompted us to consider whether there might be a more straightforward way to interpret this complex data. Consequently, the idea of grouping similar patient paths together emerged as a potential solution for simplifying the graph's complexity and extracting meaningful insights from it.

The primary objective of grouping these "patient paths" is to unearth underlying characteristics that might be common among the patients themselves. The expectation is that by identifying these common traits, the hospital can be better prepared to manage similar cases in the future. Specifically, understanding these groups could allow the hospital to anticipate the resources that would be required for a particular type of patient, thereby enabling more efficient allocation of those resources.

To initiate this clustering approach, we opted for a simple methodology that included variables related to the utilization of hospital resources. These variables were specifically chosen by the hospital staff, who possess an insider's perspective and nuanced understanding of the factors that are most relevant in the day-to-day operation of the hospital. By leveraging their expertise, we aimed to ensure that the chosen variables were both meaningful and practical for the analysis.

4.1.1 Metric vs Dissimilarity

Before delving into the core analysis, it is essential to establish the foundational concepts of metric and dissimilarity that will be utilized throughout this study.

4.1.2 Metric

A metric on a set X is a function $d : X \times X \rightarrow [0, \infty)$ that satisfies the following properties for all $x, y, z \in X$:

1. Identity of indiscernibles: $d(x, x) = 0$;
2. Non-negativity: $d(x, y) \geq 0$;
3. Symmetry: $d(x, y) = d(y, x)$;
4. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

An example of a metric is the known Euclidean Distance. Consider three vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)$, $\mathbf{y} = (y_1, y_2, \dots, y_n)$ and $\mathbf{z} = (z_1, z_2, \dots, z_n)$.

The Euclidean Distance between vectors x and y is defined as:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4.1)$$

While the first three properties of the Euclidean Distance are straightforward and require little elaboration, the fourth property presents a more complex challenge that warrants further investigation. Hence, we shall focus our attention on proving this fourth property

We aim to prove that this metric satisfies the triangular inequality. Starting with the inequality, we square both sides to eliminate the square root, which will be equivalent to the original inequality due to the non-negativeness of the distance:

$$(d(x, z))^2 \leq (d(x, y) + d(y, z))^2 \quad (4.2)$$

By expanding and substituting the definition of Euclidean Distance, we get:

$$\sum_{i=1}^n (x_i - z_i)^2 \leq \sum_{i=1}^n (x_i - y_i)^2 + 2\sqrt{\sum_{i=1}^n (x_i - y_i)^2} \sqrt{\sum_{i=1}^n (y_i - z_i)^2} + \sum_{i=1}^n (y_i - z_i)^2 \quad (4.3)$$

Rewriting $\sum_{i=1}^n (x_i - z_i)^2$ as $\sum_{i=1}^n ((x_i - y_i) + (y_i - z_i))^2$ and expanding the square gives:

$$\sum_{i=1}^n (x_i - z_i)^2 = \sum_{i=1}^n (x_i - y_i)^2 + 2 \sum_{i=1}^n (x_i - y_i)(y_i - z_i) + \sum_{i=1}^n (y_i - z_i)^2 \quad (4.4)$$

Rewriting the triangular inequality without the common terms provides:

$$\sum_{i=1}^n (x_i - y_i)(y_i - z_i) \leq \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \sqrt{\sum_{i=1}^n (y_i - z_i)^2} \quad (4.5)$$

Upon squaring both sides, we arrive at:

$$\left(\sum_{i=1}^n (x_i - y_i)(y_i - z_i) \right)^2 \leq \sum_{i=1}^n (x_i - y_i)^2 \sum_{i=1}^n (y_i - z_i)^2 \quad (4.6)$$

By the Cauchy–Schwarz inequality, this last equation holds true, thus confirming that the Euclidean Distance adheres to the triangular inequality.

4.1.3 Dissimilarity

In contrast to metrics, dissimilarities offer a nuanced way of evaluating the differences between points, albeit without strictly adhering to the triangle inequality property as defined by Gower (1986) [5]. To clarify this distinction, we consider an illustrative example that incorporates the variables age, sex, and disease state of a patient. For the purpose of this example, let us consider three patients—X, Y, and Z—represented by their respective vectors.

1. $\mathbf{x} = (37, 1, 0);$
2. $\mathbf{y} = (35, 0, 0);$
3. $\mathbf{z} = (31, 0, 1);$

In this scenario, we can define a dissimilarity function $d(\mathbf{x}, \mathbf{y})$ that calculates the absolute difference in age between two patients, and adds a constant value if their sex and disease states are different. Mathematically, this can be expressed as:

$$d(x, y) = \begin{cases} |x_1 - y_1| + 3 & \text{if } x_2 \neq y_2 \text{ and } x_3 \neq y_3 \\ |x_1 - y_1| & \text{otherwise} \end{cases} \quad (4.7)$$

Under this formulation, the dissimilarity between the 3 combinations of pairs of patients becomes $d(x, y) = 2$, $d(x, z) = 9$, and $d(y, z) = 4$. While this function preserves many properties inherent to a metric, it notably violates the triangle inequality, as $d(x, z) > d(x, y) + d(y, z)$.

The example presented serves to highlight the inherent flexibility and adaptability of dissimilarity measures. Unlike traditional metrics, they can be customized to incorporate additional variables or specific conditions that may be critical in specialized domains. In fields like healthcare, for instance, this allows for the inclusion of nuanced factors that might significantly impact patient pathways or treatment outcomes.

4.1.4 Gower

One of the challenges of clustering data is how to measure the similarity or dissimilarity between objects that have mixed types of variables, such as numerical, categorical, ordinal, or binary. A common solution to this problem is to use the **Gower's distance**, a metric that can handle different scales and missing values. Gower's distance was proposed by John Gower in 1971 [6] and has been widely used in various fields such as ecology [7], biology [8], and social sciences [9].

The Gower's distance is defined as the weighted average of the pairwise dissimilarities between two objects, among all variables. For each variable, the dissimilarity depends on its type and range, and can be adjusted by a weight factor that reflects its importance.

The Gower's distance is a measure of dissimilarity between two observations that have mixed numeric and categorical data. It is useful for clustering or classification problems where the data types are heterogeneous. The Gower's distance between any two p -dimensional vectors i and j (named objects) is given by [6] [10]:

$$d(i, j) = \frac{\sum_{f=1}^p w_f \delta_{ij}^f d_{ij}^f}{\sum_{f=1}^p w_f \delta_{ij}^f} \quad (4.8)$$

where w_f is the weight of the f -th feature (coordinate of the vector)

$$\sum_{f=1}^p w_f = 1 \quad (4.9)$$

δ_{ij}^f is an indicator function that equals 1 if both objects have a valid value for the f -th feature and 0 otherwise,

$$\delta_{ij}^f = \begin{cases} 1 & \text{if } x_{if} \text{ and } x_{jf} \neq \text{NA} \\ 0 & \text{otherwise} \end{cases} \quad (4.10)$$

and d_{ij}^f is the distance between the values of the f -th feature for objects i and j . The dissimilarity for a variable is zero if both objects have missing values for that variable, and one if only one object has a missing value. As such the Gower's distance can deal with incomplete data without imputation or deletion.

4.1.4.1 Interval-Scaled, Ratio-Scaled and Ordinal Variables

For variables of the types Interval-Scaled, Ratio-Scaled or Ordinal, the dissimilarity between objects corresponds to the normalized Manhattan distance, which is achieved by taking the norm of the difference between the values and dividing it by the range of the specific variable. For ratio-scaled variables, a log transformation, $\log(x_f)$, is applied prior to calculating the dissimilarity since they are, by definition, positive and the distinction between 2 and 20 is the same as 20 and 200, while ordinal variables undergo a transformation into ranks before the dissimilarity is determined.

$$d_{ij}^f = \frac{|x_{if} - x_{jf}|}{R_f} \quad (4.11)$$

$$R_f = \max(x_f) - \min(x_f) \quad (4.12)$$

4.1.4.2 Categorical Variables

For any categorical variable, dichotomous or not, the dissimilarity between two objects is very simple: if the values are the same it is 0, otherwise it is 1.

$$d_{ij}^f = \begin{cases} 0 & \text{if } x_{if} = x_{jf} \\ 1 & \text{otherwise} \end{cases} \quad (4.13)$$

4.1.4.3 Dichotomous Variables

For dichotomous variables, the calculation of the Gower distance varies depending on whether the variables are symmetric or asymmetric.

Dichotomous variables consist of two potential outcomes: 1 (positive/present) or 0 (negative/absent). When both states hold equal importance and can be interchangeably coded as 0 or 1, the binary variable is considered symmetric. For instance, the binary variable "Sex," with possible states "Male" and "Female," treats both states with equal significance when calculating dissimilarities.

Conversely, asymmetric binary variables occur when the outcomes are not equally important. A prime example is the presence or absence of a relatively rare attribute, such as a specific disease. While two individuals with the disease share a common attribute, it cannot be inferred that those without the disease also share something in common. In these cases, one outcome holds more significance than the other. The more significant outcome is typically coded as 1 (present), while the less significant outcome is coded as 0 (absent). Due to the higher importance of agreement between two 1's compared to two 0's, the Gower distance incorporates the Jaccard measure for asymmetric binary variables.

To account for this variation in the general formula for the distance between two objects, δ_{ij}^f in equation 4.10 is set to zero when $x_{if} = x_{jf} = 0$, thus ensuring that the unique characteristics of symmetric and asymmetric binary variables are appropriately considered.

4.1.4.4 The daisy function

In R, Gower's distance for mixed data types can be computed using the `daisy()` function from the `cluster` package. The original version of the `daisy()` function is fully described in Chapter 1 of Kaufman and Rousseeuw (1990). The function returns a dissimilarity matrix that can be converted into a similarity matrix by subtracting each dissimilarity value from the maximum possible value (one). The resulting similarity matrix is positive semi-definite, assuming no missing values in the similarity matrix, which indicates that the dissimilarity matrix represents the square of Euclidean distances [6]. This is important for the use of the Ward method in agglomerative hierarchical algorithms in the next sections.

The `daisy()` function allows the user to assign different levels of importance to each variable by specifying weights, in the `weights` argument. Additionally, the function enables users to explicitly specify the type of the variables, ensuring accurate Gower distance calculations.

4.2 Centroids-based Clustering (Partitioning methods)

The Partition Around Medoids (PAM) algorithm is a clustering method with the goal of finding a set of representative objects, known as medoids, that minimize an objective function. In the context of PAM, this objective function is defined as the sum of the dissimilarities between each object in the data set and its nearest medoid. Unlike the popular k-means algorithm, which uses centroids as cluster centers and aims to minimize the square distance of the objects to the nearest center, the PAM algorithm uses actual objects from the data set and minimizes the absolute distance (or dissimilarity) to each nearest medoid. This makes PAM more robust to outliers and noise.

PAM consists of two main steps: build and swap. In the build step, an initial set of K medoids is chosen either randomly or heuristically from the data set. In the swap step, each medoid is replaced by a non-medoid object of the same cluster if doing so improves the objective function. This process is repeated until no further improvement can be achieved or a maximum number of iterations is reached. PAM can handle any kind of dissimilarity measure between objects.

Pseudo-code to the BUILD algorithm as proposed in Kaufman and Rousseeuw (1990) [10] Chapter 2 can be found in the [A](#)

4.3 Connectivity-based Clustering (Hierarchical clustering)

Hierarchical clustering aims to build a hierarchy of clusters by successively splitting (divisive) or merging (agglomerative) clusters based on a measure of similarity or dissimilarity. Divisive clustering starts with a single cluster containing all objects and proceeds by iteratively finding the most dissimilar groups and splitting them apart. The process continues until each object becomes its own singleton cluster (Cluster composed of a single data point). Agglomerative clustering begins with each object as its own singleton cluster and merges the most similar clusters at each step until all objects belong to the same cluster.

Both divisive and agglomerative clustering methods yield a dendrogram, a tree-like graphical representation that encapsulates the nested grouping of objects along with the hierarchical organization of clusters. This dendrogram offers the flexibility to be cut at various levels thus deriving different numbers of clusters, thereby allowing analysts to fine-tune the granularity of their investigation. Typically, the optimal cut is identified at a

point where an horizontal line can intersect the longest vertical lines in the dendrogram, which usually results in clusters that maximize intra-cluster similarity while minimizing inter-cluster similarity.

4.3.1 Linkage

In agglomerative hierarchical clustering, one of the critical aspects is the calculation of distances and dissimilarities between clusters, particularly as clusters evolve from singletons to more complex formations. The question arises: once two singletons are combined into a new cluster, how should the distance between this new cluster and all remaining clusters be calculated? This is where linkage methods come into play. Linkage methods, including simple, complete, average, and Ward's, among others, offer various strategies to determine these distances, each with its own merits and limitations.

For instance, the single linkage method calculates the distance between two clusters by identifying the minimum distance among all pairwise distances between members of the two clusters. On the other hand, the complete linkage method focuses on the maximum distance among these pairwise distances. Each of these methods, therefore, offers a different perspective on cluster proximity and influences the resultant dendrogram in unique ways. The choice of the linkage method can significantly affect the clustering results and, consequently, the insights gleaned from the analysis.

4.3.2 Ward's Method

First introduced in 1963, the Ward's method of agglomerative clustering relies on the minimization of the Error Sum of Squares (ESS) to group clusters. ESS is the sum of squared deviations between members of a cluster and its centroid. It starts at zero—since every data point initially forms its own cluster—and then the approach seeks to make connections that lead to the least increase in ESS. Accordingly, for a cluster A , the ESS is given by:

$$ESS(A) = \sum_{x \in A} ||x - C_A||^2 \quad (4.14)$$

Where C_A represents the centroid of cluster A

The Total ESS is the sum of the ESS values for all clusters. When clusters A and B are merged into a new cluster F ($F = A \cup B$), the change in Total ESS (ΔESS) can be expressed as:

$$\Delta ESS = ESS(F) - (ESS(A) + ESS(B)) \quad (4.15)$$

The centroid for F can be computed as the weighted mean of the centroids of A and B :

$$C_F = \frac{n_A C_A + n_B C_B}{n_A + n_B} \quad (4.16)$$

where n_A and n_B indicate the number of elements in clusters A and B , respectively.

Upon some algebraic manipulations, the distance between any two clusters A and B can be derived as:

$$d(A, B) = \frac{n_A n_B}{n_A + n_B} \|C_A - C_B\|^2 \quad (4.17)$$

Let A and B be arbitrary clusters. This distance measure can be generalized to involve F and G , with F being the newly created cluster from A and B . Algebraic manipulation yields:

$$d(F, G) = \frac{n_A + n_G}{n_F + n_G} d(A, G) + \frac{n_B + n_G}{n_F + n_G} d(B, G) + \frac{n_A + n_B}{n_F + n_G} d(A, B) \quad (4.18)$$

In Ward's method, the employment of a squared Euclidean distance is mandatory, whether directly applied or indirectly squared. However, the updated distance approach to Ward's method offers an advantage: it allows us to focus solely on the distances between objects, eliminating the need to consider their precise locations, which would otherwise be essential for centroid calculation.

4.3.3 Lance-Williams

The Lance-Williams formula is a general framework for agglomerative hierarchical clustering methods, unifying multiple linkage criteria under a single mathematical representation. The formula describes the dissimilarity between a newly formed cluster and the remaining clusters in the data, allowing the clustering algorithm to efficiently update the dissimilarity matrix after each merge.

Given two clusters A and B that are merged to form a new cluster F , the Lance-Williams formula computes the dissimilarity between the new cluster F and another cluster G as follows:

$$d(F, G) = \alpha_A d(A, G) + \alpha_B d(B, G) + \beta d(A, B) + \gamma |d(A, B) - d(A, B)| \quad (4.19)$$

Here, α_A , α_B , β , and γ are coefficients that depend on the specific linkage criterion being used, and $d(A, G)$, $d(B, G)$, and $d(A, B)$ represent the dissimilarities between clusters in the dissimilarity matrix before the merge.

Different linkage criteria can be represented within the Lance-Williams framework by setting the coefficients appropriately:

- **Single linkage:** $\alpha_A = \frac{1}{2}, \alpha_B = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$
- **Complete linkage:** $\alpha_A = \frac{1}{2}, \alpha_B = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}$
- **Average linkage:** $\alpha_A = \frac{|A|}{|A|+|B|}, \alpha_B = \frac{|B|}{|A|+|B|}, \beta = 0, \gamma = 0$
- **Centroid linkage:** $\alpha_A = \frac{|A|}{|A|+|B|}, \alpha_B = \frac{|B|}{|A|+|B|}, \beta = -\frac{|A||B|}{(|A|+|B|)^2}, \gamma = 0$
- **Ward's method:** $\alpha_A = \frac{|A|+|G|}{|A|+|B|+|G|}, \alpha_B = \frac{|B|+|G|}{|A|+|B|+|G|}, \beta = -\frac{|G|}{|A|+|B|+|G|}, \gamma = 0$

By employing the Lance-Williams formula, various agglomerative hierarchical clustering methods can be implemented in a unified and efficient manner, enabling comparisons and the application of different linkage criteria to the clustering process.

4.3.4 The hclust function

In R the `hclust()` function from the `stats(version 3.6.2)` package does an hierarchical agglomerative clustering and implements the Lance-Williams formula with many methods. One particularity is that there is a difference in using `method = "Ward.D"` and `method = "Ward.D2"` the latter starts by squaring the dissimilarity matrix used. (In the details of the R documentation of `hclust()` of). Normally to implement the correct Ward's method using a distance matrix this matrix needs to be squared first. Writing `hclust(dist(x)^2, method = "Ward.D")` is the same as `hclust(dist(x), method = "Ward.D2")`.

In the context of hierarchical clustering, it is crucial to note the variations in Ward's clustering algorithm as implemented in R. As stated in the R documentation for `hclust()`:

Two different algorithms are found in the literature for Ward clustering. The one used by option "ward.D" (equivalent to the only Ward option "ward" in R versions $\geq 3.0.3$) does not implement Ward's (1963) clustering criterion, whereas option "ward.D2" implements that criterion (Murtagh and Legendre 2014). With the latter, the dissimilarities are squared before cluster updating. Note that `agnes(*, method="ward")` corresponds to `hclust(*, "ward.D2")`.

It is essential to choose the appropriate variant based on the specific research question and data characteristics.

Experiment 4 of Murtagh (2014) [11] also confirms this, since `hclust()` function from `stats` with the new version is the same as `hclust.PL()` function from `hclust` and they give the same results. In the paper the version of `stats` hadn't include the "ward.D2" method yet.

Given that our dissimilarity matrix, generated using the `daisy()` function, is already the square of Euclidean distances, as explained by Gower (1971) [6], the appropriate choice for our analysis would be `hclust(dist(x), method = "Ward.D")`.

4.4 Distribution-based Clustering

A Finite Mixture Model (FMM) serves as a powerful tool for modeling heterogeneous populations. The core idea is to represent the observed data as a combination of multiple underlying probability distributions [12] [13]. Mathematically, the probability density function $p(x)$ of observed data x is defined as:

$$p(x) = \sum_{i=1}^K \pi_i f_i(x; \theta_i)$$

where K is the number of mixture components, π_i is the mixing coefficient for the i^{th} component, and $f_i(x; \theta_i)$ is the probability density function of the i^{th} component with parameters θ_i .

The inclusion of concomitant variables, also known as covariates, further refines the model by allowing the mixing coefficients to vary with respect to these variables. This extended form is often utilized when it is believed that the proportions of different subpopulations are not constant but are functions of observable variables. The role of concomitant variables is especially pivotal in applications where the population heterogeneity is influenced by external factors, offering a more nuanced understanding of the data.

4.5 Results and Discussion

As said in section 4.1, in this clustering analysis, the variables were meticulously selected through a collaborative effort with the hospital staff to capture essential aspects of patient pathways. These variables were: "Admission Type", which categorizes episodes as either "Scheduled" or "Urgent"; "Surgery", indicating whether a surgical procedure was performed (Yes/No); "Intensive Care", specifying if the patient required intensive

care treatment (Yes/No); and "Number of Services", which enumerates whether the patient passed through one, two, or more than two services during their hospital stay. Additionally, "Total Length of Stay" (TLoS) was included as a numerical variable. It's important to note that episodes labeled with "Admission Type" as "Normal" were excluded from the analysis, reducing the dataset from 3133 episodes to 2959.

In our analysis, we employed three binary variables. Intuitively, one might consider forming a cluster for every unique combination of these variables, resulting in a total of eight clusters. Consequently, any clustering structure suggesting more than eight as the "optimal" number of clusters will be deemed infeasible. This decision is motivated by two main factors: First, interpretability is crucial in our analysis, and an excessive number of clusters would complicate this task. Second, in the context of this analysis, eight clusters represent the most straightforward and still logical way to delineate the groups.

To construct the dissimilarity matrix, we employed the `daisy()` function. In this function, the "Admission Type" was treated as a symmetric binary variable, while "Surgery" and "Intensive Care" were treated as asymmetric binary variables. The variable "TLoS" (Total Length of Stay) was considered ratio-scaled to account for the unequal clinical importance of hour differences; for example, the difference between 1 and 2 hours is not the same as between 101 and 102 hours. Lastly, the "Number of Services" was transformed into a factor variable with appropriate levels and treated as an ordinal variable in the analysis.

Upon constructing the dissimilarity matrix, we employed two functions for our initial clustering approach: `pam()` from the `cluster` package and `hclust()` from the `stats` package. Our objective was to ascertain the optimal number of clusters.

Utilizing the `pam()` function in conjunction with the Average Silhouette Width (ASW) criterion, the optimal number of clusters appeared to be 7. It's worth noting that the ASW values were closely comparable across varying numbers of clusters.

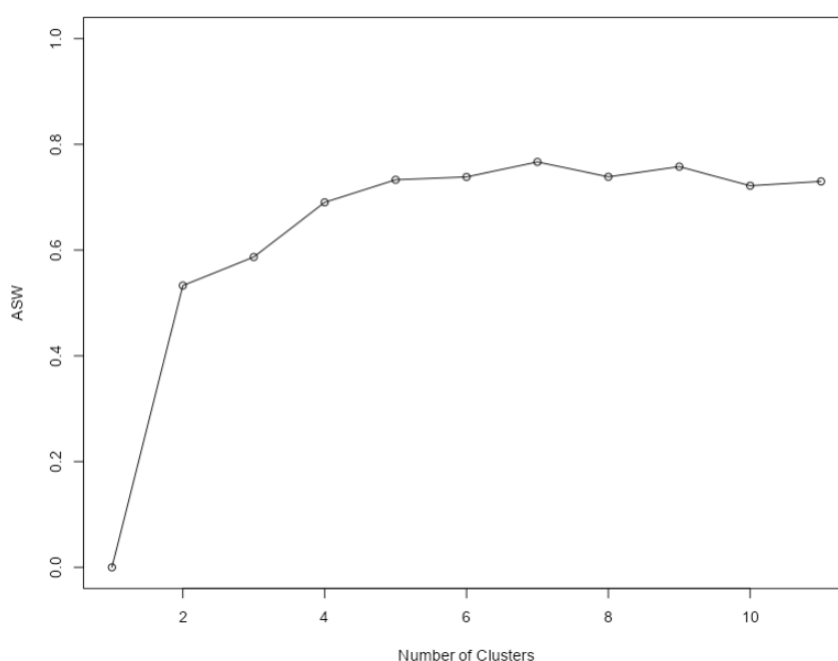


FIGURE 4.1: Average Silhouette Width using the Gower dissimilarity matrix

Doing the Dendrogram with `hclust(, method="Ward.D")` we can see that many numbers of clusters may be used, from 2 to 6 clusters being the most notorious per analysis of the height of vertical lines.

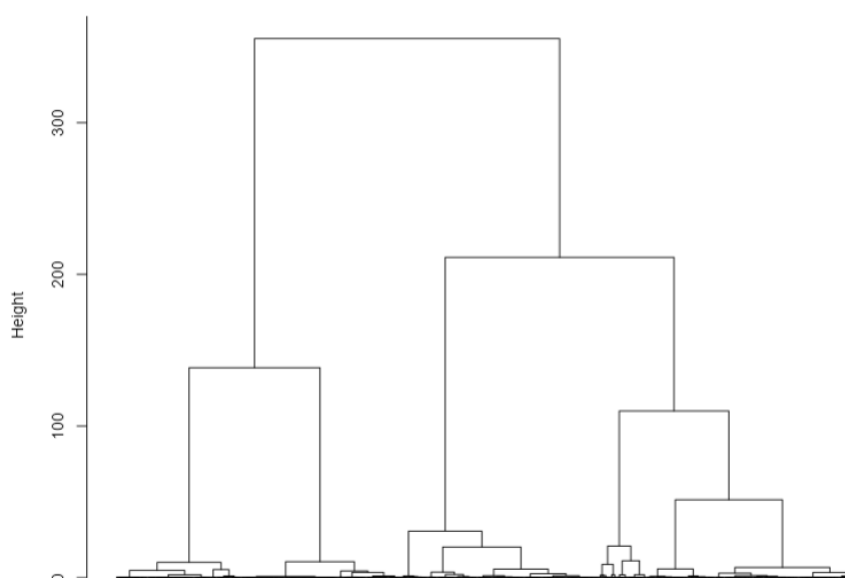


FIGURE 4.2: Dendrogram for the Gower dissimilarity matrix using Ward's method

We constructed five different clustering structures for both PAM and Hclust, ranging from 4 to 8 clusters. To assess the compatibility between the clustering solutions generated by PAM and Hclust, we used the Adjusted Rand Index (ARI). The ARI scores for

the clustering structures, ranging from 4 to 8 clusters, were 0.866, 0.971, 0.965, **0.977**, and 0.901, respectively. Among these, the 7-cluster structure exhibited the highest compatibility between the two algorithms. It's noteworthy that all structures yielded a high ARI.

To conduct a more thorough analysis, we expanded our methods to include Finite Mixture Models (FMM). In order to ensure comparability with the PAM and Hclust algorithms, we abstained from incorporating any concomitant variables, with the notable exception of "Number of Services." For this variable, we employed a Poisson distribution and utilized the actual numbers of services, as opposed to the factor levels of "1," "2," and "+2." To ascertain the optimal number of clusters, we consulted the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Integrated Complete-data Likelihood (ICL). These metrics pointed to a 7-cluster structure, thereby bolstering confidence in the robustness of this configuration, as we can see bellow

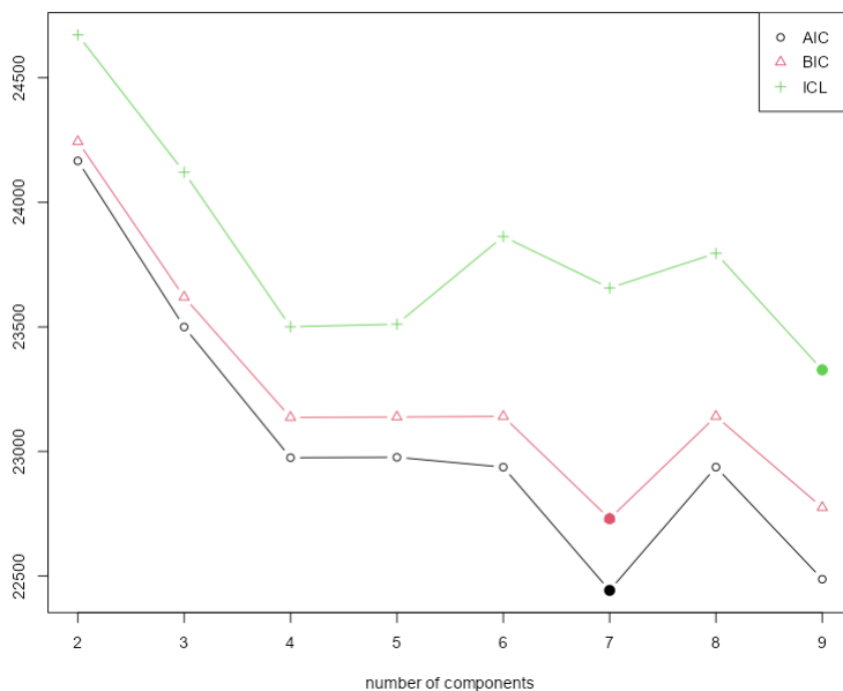


FIGURE 4.3: AIC, BIC, and ICL for various number of components (clusters)

Despite the 7-cluster structure emerging as the optimal choice across all three algorithms—FMM, PAM, and Hclust—the Adjusted Rand Index (ARI) revealed that the FMM model was not highly compatible with the other two algorithms. Specifically, the ARI for the 7-cluster structure between FMM and PAM was 0.385, and between FMM and Hclust, it was 0.397. This suggests that while all three algorithms converge on a similar number

of clusters, the actual clustering assignments shows considerable variation between FMM and the other two.

Below, we present is the 7-cluster structure described by variables of hospital resource use and sociodemographic variables:

PAM	2				7				3				5				6				1				4			
Variable	N	Percent			N	Percent			N	Percent			N	Percent			N	Percent			N	Percent			N	Percent		
Surgery	116				288				644				565				138				514				694			
... No	12	10%			0	0%			0	0%			0	0%			138	100%			514	100%			694	100%		
... Yes	104	90%			288	100%			644	100%			565	100%			0	0%			0	0%			0	0%		
Intensive Care	116				288				644				565				138				514				694			
... No	1	1%			277	96%			643	100%			563	100%			102	74%			500	97%			688	99%		
... Yes	115	99%			11	4%			1	0%			2	0%			36	26%			14	3%			6	1%		
Admission Type	116				288				644				565				138				514				694			
... Scheduled	87	75%			0	0%			644	100%			0	0%			0	0%			514	100%			0	0%		
... Urgent	29	25%			288	100%			0	0%			565	100%			138	100%			0	0%			694	100%		
Number of Services	116				288				644				565				138				514				694			
... 1	1	1%			0	0%			628	98%			565	100%			0	0%			495	96%			694	100%		
... 2	18	16%			283	98%			16	2%			0	0%			130	94%			19	4%			0	0%		
... +2	97	84%			5	2%			0	0%			0	0%			8	6%			0	0%			0	0%		
Weekend	116				288				644				565				138				514				694			
... No	82	71%			215	75%			565	88%			401	71%			98	71%			469	91%			516	74%		
... Yes	34	29%			73	25%			79	12%			164	29%			40	29%			45	9%			178	26%		
Discharge Type	116				288				644				565				138				514				694			
... Discharge	111	96%			284	99%			632	98%			440	78%			137	99%			484	94%			638	92%		
... Deceased	3	3%			0	0%			1	0%			2	0%			0	0%			4	1%			6	1%		
... Hospital	2	2%			4	1%			11	2%			123	22%			1	1%			26	5%			50	7%		
Variable	N	Mean	SD		N	Mean	SD		N	Mean	SD		N	Mean	SD		N	Mean	SD		N	Mean	SD		N	Mean	SD	
TLoS	116	281,0	182,0		288	105,4	102,8		644	69,5	105,1		565	42,5	65,4		138	209,6	145,1		514	85,7	90,9		694	123,3	94,3	

TABLE 4.1: PAM clusters described by Hospital Resources

Hclust	2			7			3			5			6			1			4		
Variable	N	Percent		N	Percent		N	Percent		N	Percent		N	Percent		N	Percent		N	Percent	
Surgery	185			277			644			563			102			500			688		
... No	68	37%		0	0%		0	0%		0	0%		102	100%		500	100%		688	100%	
... Yes	117	63%		277	100%		644	100%		563	100%		0	0%		0	0%		0	0%	
Intensive Care	185			277			644			563			102			500			688		
... No	0	0%		277	100%		644	100%		563	100%		102	100%		500	100%		688	100%	
... Yes	185	100%		0	0%		0	0%		0	0%		0	0%		0	0%		0	0%	
Admission Type	185			277			644			563			102			500			688		
... Scheduled	101	55%		0	0%		644	100%		0	0%		0	0%		500	100%		0	0%	
... Urgent	84	45%		277	100%		0	0%		563	100%		102	100%		0	0%		688	100%	
Number of Services	185			277			644			563			102			500			688		
... 1	24	13%		0	0%		627	97%		563	100%		0	0%		481	96%		688	100%	
... 2	58	31%		272	98%		16	2%		0	0%		101	99%		19	4%		0	0%	
... +2	103	56%		5	2%		1	0%		0	0%		1	1%		0	0%		0	0%	
Weekend	185			277			644			563			102			500			688		
... No	134	72%		208	75%		565	88%		399	71%		73	72%		456	91%		511	74%	
... Yes	51	28%		69	25%		79	12%		164	29%		29	28%		44	9%		177	26%	
Discharge Type	185			277			644			563			102			500			688		
... Discharge	158	85%		274	99%		633	98%		440	78%		102	100%		482	96%		637	93%	
... Deceased	10	5%		0	0%		1	0%		0	0%		0	0%		2	0%		3	0%	
... Hospital	17	9%		3	1%		10	2%		123	22%		0	0%		16	3%		48	7%	
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
TLoS	185	261,7	186,7	277	98,6	92,4	644	69,6	105,1	563	42,1	65,1	102	178,4	103,5	500	85,4	91,3	688	123,8	94,4

TABLE 4.2: Hclust clusters described by Hospital Resources

FMM	5			6			2			3			1			7			4		
Variable	N	Percent		N	Percent		N	Percent		N	Percent		N	Percent		N	Percent		N	Percent	
Surgery	160			908			497			338			110			258			688		
... No	48	30%		203	22%		112	23%		160	47%		109	99%		177	69%		549	80%	
... Yes	112	70%		705	78%		385	77%		178	53%		1	1%		81	31%		139	20%	
Intensive Care	160			908			497			338			110			258			688		
... No	0	0%		908	100%		497	100%		338	100%		92	84%		251	97%		688	100%	
... Yes	160	100%		0	0%		0	0%		0	0%		18	16%		7	3%		0	0%	
Admission Type	160			908			497			338			110			258			688		
... Scheduled	86	54%		4	0%		497	100%		338	100%		62	56%		258	100%		0	0%	
... Urgent	74	46%		904	100%		0	0%		0	0%		48	44%		0	0%		688	100%	
Number of Services	160			908			497			338			110			258			688		
... 1	8	5%		725	80%		493	99%		307	91%		102	93%		254	98%		494	72%	
... 2	49	31%		182	20%		4	1%		30	9%		8	7%		4	2%		189	27%	
... +2	103	64%		1	0%		0	0%		1	0%		0	0%		0	0%		5	1%	
Weekend	160			908			497			338			110			258			688		
... No	114	71%		653	72%		435	88%		300	89%		86	78%		244	95%		514	75%	
... Yes	46	29%		255	28%		62	12%		38	11%		24	22%		14	5%		174	25%	
Discharge Type	160			908			497			338			110			258			688		
... Discharge	147	92%		737	81%		491	99%		324	96%		92	84%		254	98%		681	99%	
... Deceased	6	4%		1	0%		0	0%		3	1%		4	4%		0	0%		2	0%	
... Hospital	7	4%		170	19%		6	1%		11	3%		14	13%		4	2%		5	1%	
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
TLoS	160	293,7	180,6	908	29,3	16,8	497	27,7	2,9	338	172,8	141,8	110	57,4	20,3	258	49,9	3,5	688	182,2	88,4

TABLE 4.3: FMM clusters described by Hospital Resources

In the comparative analysis of these three clustering algorithms — PAM, Hclust, and FMM — their efficacy in categorizing hospital paths revealed distinct strengths and limitations.

PAM excelled in isolating paths involving surgery from those that did not. While Hclust was not as effective in this regard, it adeptly clustered all paths that proceeded through PedICM, also maintaining a clear distinction based on the type of admission.

Both PAM and Hclust generated two distinct clusters comprised solely of simple paths. Interestingly, regardless of the algorithm, there was one of such clusters absorbing most of the “Hospital” exits, even though this variable did not enter the clustering process.

PAM, albeit less effective in clustering PedICM compared to Hclust, managed to form a cluster where the vast majority of paths through PedICM involved more than two services.

When it comes to paths involving exactly two services, Hclust slightly outperformed PAM, capturing over 98% of this variable in the “2 services” category for both clusters.

Regarding TLoS, both PAM and Hclust produced clusters with comparable mean values, e.g., the cluster that grouped Surgery and PedICM exhibited identical mean TLoS in both algorithms, a pattern that held true across all other clusters as well. However, the high standard deviations observed hinder meaningful interpretations of the homogeneity of TLoS between these clusters.

Contrastingly, FMM showed moderate success in clustering PedICM as PAM did and its was very unrestrained with other variables. Notably, FMM seemed more effective in categorizing groups by TLoS, as it produced 4 clusters with consistently low TLoS and 3 with high TLoS, with very low standard deviation values for some clusters with low TLoS.

PAM	2		7		3		5		6		1		4	
Variable	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Sex	116		288		644		565		138		514		693	
... Female	57	49%	100	35%	245	38%	164	29%	67	49%	250	49%	313	45%
... Male	59	51%	188	65%	399	62%	401	71%	71	51%	264	51%	380	55%
Age_Category	116		288		644		565		138		514		693	
... Newborn	0	0%	0	0%	19	3%	2	0%	8	6%	56	11%	30	4%
... Baby	25	22%	7	2%	36	6%	21	4%	19	14%	36	7%	125	18%
... Toddler	37	32%	40	14%	224	35%	98	17%	41	30%	122	24%	206	30%
... Child	24	21%	89	31%	132	20%	171	30%	25	18%	108	21%	128	18%
... Teenager	30	26%	152	53%	233	36%	273	48%	45	33%	192	37%	204	29%
District_Residence	116		288		644		565		138		514		693	
... Porto	59	51%	211	73%	481	75%	431	76%	103	75%	266	52%	517	75%
... Other	57	49%	77	27%	163	25%	134	24%	35	25%	248	48%	176	25%
Country_Birth	116		288		644		565		138		514		693	
... Portugal	112	97%	273	95%	634	98%	537	95%	130	94%	504	98%	670	97%
... Other	4	3%	15	5%	10	2%	28	5%	8	6%	10	2%	23	3%

TABLE 4.4: PAM clusters described by Patient Sociodemographic Variables

Hclust	2		7		3		5		6		1		4	
Variable	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Sex	185		277		644		563		102		500		687	
... Female	88	48%	98	35%	245	38%	164	29%	50	49%	241	48%	310	45%
... Male	97	52%	179	65%	399	62%	399	71%	52	51%	259	52%	377	55%
Age_Category	185		277		644		563		102		500		687	
... Newborn	0	0%	0	0%	19	3%	2	0%	8	8%	56	11%	30	4%
... Baby	37	20%	6	2%	36	6%	21	4%	14	14%	31	6%	124	18%
... Toddler	56	30%	35	13%	224	35%	98	17%	32	31%	120	24%	203	30%
... Child	38	21%	88	32%	132	20%	171	30%	17	17%	104	21%	127	18%
... Teenager	54	29%	148	53%	233	36%	271	48%	31	30%	189	38%	203	30%
District_Residence	185		277		644		563		102		500		687	
... Porto	105	57%	205	74%	480	75%	429	76%	75	74%	262	52%	512	75%
... Other	80	43%	72	26%	164	25%	134	24%	27	26%	238	48%	175	25%
Country_Birth	185		277		644		563		102		500		687	
... Portugal	175	95%	263	95%	634	98%	535	95%	99	97%	490	98%	664	97%
... Other	10	5%	14	5%	10	2%	28	5%	3	3%	10	2%	23	3%

TABLE 4.5: Hclust clusters described by Patient Sociodemographic Variables

FMM	6		1		7		3		2		4		5	
Variable	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Sex	160		907		497		338		110		258		688	
... Female	72	45%	298	33%	190	38%	146	43%	50	45%	135	52%	305	44%
... Male	88	55%	609	67%	307	62%	192	57%	60	55%	123	48%	383	56%
Age_Category	160		907		497		338		110		258		688	
... Newborn	0	0%	4	0%	7	1%	56	17%	11	10%	4	2%	33	5%
... Baby	29	18%	39	4%	27	5%	21	6%	13	12%	19	7%	121	18%
... Toddler	49	31%	163	18%	169	34%	83	25%	28	25%	77	30%	199	29%
... Child	33	21%	258	28%	117	24%	40	12%	30	27%	63	24%	136	20%
... Teenager	49	31%	443	49%	177	36%	138	41%	28	25%	95	37%	199	29%
District_Residence	160		907		497		338		110		258		688	
... Porto	88	55%	680	75%	362	73%	188	56%	73	66%	166	64%	511	74%
... Other	72	45%	227	25%	135	27%	150	44%	37	34%	92	36%	177	26%
Country_Birth	160		907		497		338		110		258		688	
... Portugal	151	94%	862	95%	489	98%	330	98%	108	98%	254	98%	666	97%
... Other	9	6%	45	5%	8	2%	8	2%	2	2%	4	2%	22	3%

TABLE 4.6: FMM clusters described by Patient Sociodemographic Variables

Regarding sociodemographic variables, the most salient difference appeared in two specific clusters across all algorithms. These clusters had a nearly equal distribution of individuals from Porto and other districts. For each algorithm, one of these distinct clusters corresponded to the PedICM group, while the other was exclusively composed of scheduled admissions.

Chapter 5

Clustering Hospitalization Paths

5.1 Problem

In the previous chapter, the limitations of our clustering approach became apparent. Specifically, the clustering was not directly based on the paths themselves, but rather on variables that were indirectly associated with these paths. While this provided some initial insights, it fell short of capturing the true essence and complexity of the patient pathways within the Pediatric Ward.

In light of these limitations, we felt compelled to develop new dissimilarity measures that would be more contextually relevant than those commonly found in existing literature. Although we did integrate one established algorithm, the Needleman-Wunsch algorithm [14], into our more complex dissimilarity measure, most conventional measures were ill-suited for analyzing sequences of physical services and their respective LoS. We looked at dissimilarities developed within the bio-informatics and genetics field but none satisfied our requirements.

In the current chapter, we aim to overcome these limitations by introducing various dissimilarity measures tailored for paths. These measures, which vary in complexity, provide a more nuanced framework for clustering and grouping the paths effectively. Through extensive discussions with the medical staff, it became clear that the key factors contributing to dissimilarities between paths were primarily the services each patient visited. The time spent in each service was considered a secondary factor, while the sequence in which services were visited held minimal relevance. Their intuitive understanding of the hospital operations was invaluable in this process, and a dissimilarity matrix between physical services based on their collective insights was created.

Following the introduction and description of each dissimilarity measure, we employed PAM algorithm to obtain several clustering structures, and used the method of Average Silhouette Width to determine the optimal number of clusters. To further elucidate the characteristics of these clusters, we incorporated additional variables into our analysis. These included patient socio-economic factors and other hospital resources that were not initially included in the definition of the dissimilarities.

Through the inclusion of these new dissimilarity measures and auxiliary variables, we aim to offer a more comprehensive understanding of patient paths. This, in turn, has the potential to optimize resource allocation and improve patient outcomes, fulfilling the overarching objective of this research.

5.2 Dissimilarities

In crafting the dissimilarity measures for this study, one central question emerged: how can we meaningfully compare the time patients spend in different services, when the services themselves have very different Lengths of Stay (LoS)? For instance, a 48-hour stay in the Short-Term Unit (STU) might be considered unusually long, whereas the same duration in the Paediatric Intensive Care Medicine (PedICM) could be considered short-lived. Given the inherent differences in LoS expectations for each service, a straightforward comparison would be misleading.

To address this issue, we implemented a standardization process for the time spent in each service. Initially, we eliminated any paths that included services falling below the 5th percentile or above the 95th percentile in terms of LoS. This helped to mitigate the effects of outliers and focused our analysis on the most common hospital stays. Following this filtration, we standardized the remaining times according to the following formula:

$$t = \frac{t' - Min}{Range} \quad (5.1)$$

In this formula, t' denotes the time a patient spends in a particular service, measured in hours, and Min signifies the minimum Length of Stay (LoS) for the service in question, and $Range$ is defined as the difference between the maximum and minimum LoS for that service. This enabled us to compare LoS across different services on a similar scale.

Beyond the time spent in individual services, we recognized the potential importance of evaluating the Total LoS (TLoS) for each patient path. The TLoS for a path is simply the

sum of all LoS in that specific path, mathematically represented as:

$$T' = \sum t' \quad (5.2)$$

We subjected this TLoS to a similar standardization process to maintain consistency in our analysis. The standardized TLoS ((TLoS_s)) is calculated as follows:

$$T = \frac{T' - \text{Min}}{\text{Range}} \quad (5.3)$$

Here the *Min* and *Range* represent the minimum and range of Total Length of Stay, measured in hours, across all paths.

This time standardization — for individual services and for total paths — enriches the reliability of our dissimilarity measures, thus enhancing the overall quality of the clusters formed. By ensuring that time spent in different services and in paths is comparable, we lay a stronger foundation for subsequent analyses and interpretations.

Here we can see the Min, Max and Range for each Service LoS and for all paths:

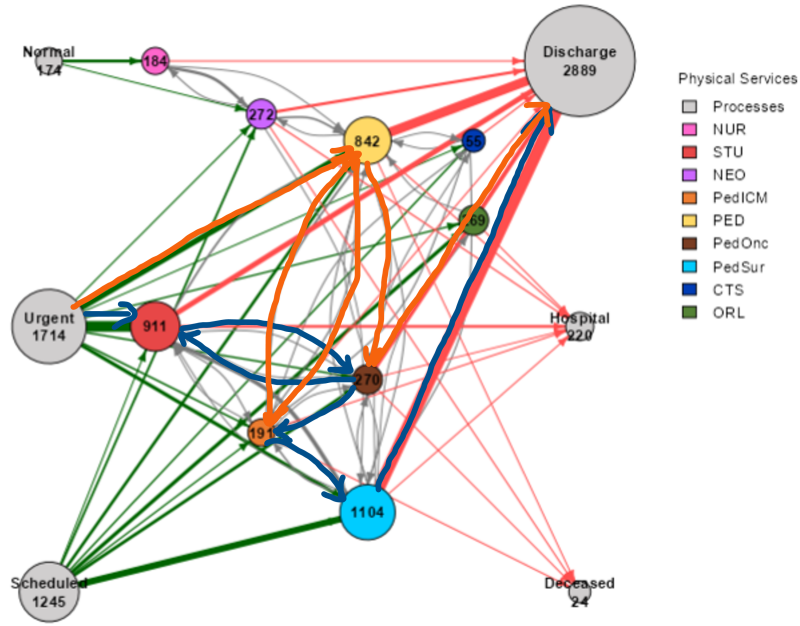
Service	Min	Max	Range
CTS	0,42	167,52	167,10
PedSur	13,73	311,17	297,44
PedICM	17,58	343,90	326,32
Neo	24,43	1104,15	1079,72
PedOnc	24,63	465,18	440,55
ORL	12,35	103,43	91,08
Ped	23,90	498,82	474,92
NUR	0,22	87,52	87,30
STU	2,55	48,60	46,05

FIGURE 5.1: Min, Max and Range for each Service

Paths	Min	Max	Range
All	2,38	1104,52	1102,14

FIGURE 5.2: Min, Max and Range for all paths

To facilitate a clearer understanding of the dissimilarities between hospital paths, we will employ two representative examples from the data, labeled "Orange" and "Blue." In graphical form, these paths are represented as shown in Figure 5.3 just bellow in orange and dark blue.

FIGURE 5.3: Graph of the paths X_{1097} (dark blue) and X_{1649} (orange)

In a table format they are described as follows:

1097	STU	29,98	PedOnc	79,52	STU	37,12	PedOnc	243,23	PedICM	17,60	PedSur	123,12		530,57
1649	Ped	203,72	PedICM	77,37	Ped	70,25	PedOnc	165,98						517,32

FIGURE 5.4: Table of the paths X_{1097} (dark blue) and X_{1649} (orange)

Which can be rewritten with the following standardized LoS:

1097	STU	0,60	PedOnc	0,12	STU	0,75	PedOnc	0,50	PedICM	0,00	PedSur	0,37		0,48
1649	Ped	0,38	PedICM	0,18	Ped	0,10	PedOnc	0,32						0,47

FIGURE 5.5: Paths X_{1097} (dark blue) and X_{1649} with Std. LoS

Each row in these tables corresponds to a single path, with successive columns displaying the Physical Service and the associated LoS in chronological order. The last column indicates the TLoS for that particular path.

Every two paths X_i and X_k will be described in the dissimilarities as follows, except for the Jaccard-like:

$$X_i = ((S_{i1}, t_{i1}), (S_{i2}, t_{i2}), (S_{i3}, t_{i3}), \dots, (S_{in_i}, t_{in_i})) \quad (5.4)$$

$$X_k = ((S_{k1}, t_{k1}), (S_{k2}, t_{k2}), (S_{k3}, t_{k3}), \dots, (S_{kn_k}, t_{kn_k})) \quad (5.5)$$

Or more succinctly:

$$X_i = (S_{ij}, t_{ij})_{j=1,2,3,\dots,n_i} \quad (5.6)$$

$$X_k = (S_{km}, t_{km})_{m=1,2,3,\dots,n_k} \quad (5.7)$$

Here, j indicates the position of the (Service, LoS) pair in a path. S_{ij} and t_{ij} are the j -th Service and standardized LoS of path X_i , respectively. n_i specify the number of (Service, time) pairs in paths X_i . T_i represents the standardized total LoS for path X_i .

In all dissimilarity measures, with the exception of the Jaccard-like approach, paths of differing lengths are artificially made of equal length by appending pairs (Z,0) to the shorter path. This serves to equalize the lengths of the two paths for comparison and introduces a conceptual *ghost service* where a patient stays for zero hours. In the above example this results in:

1097	STU	0,60	PedOnc	0,12	STU	0,75	PedOnc	0,50	PedICM	0,00	PedSur	0,37			0,48
1649	Ped	0,38	PedICM	0,18	Ped	0,10	PedOnc	0,32	Z	0,00	Z	0,00			0,47

FIGURE 5.6: Paths X_{1097} (dark blue) and X_{1649} (orange) with Std. LoS and equal length

5.2.1 Additive Dissimilarity

The custom Additive Dissimilarity we considered is relatively straightforward. We first compared each j -th service between path X_i and path X_k . If the service was the same for both paths, a zero was returned; otherwise, a one was returned. The result was then added to the absolute difference of the standardized LoS (LoS_s) for each j -th service in the paths. We have also introduced a variant that incorporates a slight penalty based on the difference in the TLoS_s between the two paths.

Thus, the Additive Dissimilarity between path X_i and path X_k can be mathematically represented as:

$$D_{ik}^A = \sum_{j=1}^n \alpha D_{ikj}^S + \beta D_{ikj}^t \quad (5.8)$$

$$D_{ik}^{A'} = \left(\sum_{j=1}^n \alpha D_{ikj}^S + \beta D_{ikj}^t \right) + \gamma D_{ik}^T \quad (5.9)$$

where $n = \max(n_i, n_k)$. We chose $\alpha = \beta = 1$ and $\gamma = 0.4$, D_{ikj}^S pertains to the Service comparison, D_{ikj}^t is concerned with the LoS_s of each Service and D_{ik}^T focuses on the TLoS_s.

The individual components of the dissimilarity can be specified as:

$$D_{ikj}^S = \begin{cases} 0 & \text{if } S_{ij} = S_{kj} \\ 1 & \text{otherwise} \end{cases} \quad (5.10)$$

$$D_{ikj}^t = |t_{ij} - t_{kj}| \quad (5.11)$$

$$D_{ik}^T = |T_i - T_k| \quad (5.12)$$

5.2.2 Product Dissimilarity

The Product Dissimilarity, the dissimilarity between two paths is calculated by multiplying the Service term D_{ikj}^S by the square root of the LoS_s term D_{ikj}^t . The square root is employed to prevent the result from approaching zero at a linear rate when two distinct services have nearly identical LoS_s, as a small number would imply that the paths were very similar. Additionally, we adjusted the Service component to take on values of either 0.8 or 1, instead of the conventional 0 or 1. This nuance ensures that two paths with the same service but considerable LoS_s differences aren't considered identical. We determined the optimal constants through trial and error with hypothetical paths.

$$D_{ik}^P = \alpha \left(\sum_{j=1}^n D_{ikj}^S \sqrt{D_{ikj}^t} \right) \quad (5.13)$$

$$D_{ik}^{P'} = \alpha \left(\sum_{j=1}^n D_{ikj}^S \sqrt{D_{ikj}^t} \right) + \gamma D_{ik}^T \quad (5.14)$$

We considered the choices $\alpha = 1$ and $\gamma = 0.4$

The individual components of the dissimilarity can be specified as:

$$D_{ikj}^S = \begin{cases} 0.8 & \text{if } S_{ij} = S_{kj} \\ 1 & \text{otherwise} \end{cases} \quad (5.15)$$

$$D_{ikj}^t = |t_{ij} - t_{kj}| \quad (5.16)$$

$$D_{ik}^T = |T_i - T_k| \quad (5.17)$$

5.2.3 Jaccard-like Dissimilarity

In contrast to the previous dissimilarities that strictly accounted for the sequence of services, the Jaccard-like dissimilarity focus primarily on the presence or absence of services across THE two paths, incorporating the TLoS_s either additively or multiplicatively.

For this dissimilarity, a path X_i is represented simply as the set of all its services: $X_i = \{S_{ij}\}_{j=1, \dots, n_i}$. The equations for the Jaccard-like dissimilarity are as follows:

$$D_{ik}^{JA} = \alpha D_{ik}^S + \gamma D_{ik}^T \quad (5.18)$$

$$D_{ik}^{JP} = D_{ik}^S \sqrt{D_{ik}^T} \quad (5.19)$$

where $\alpha = \gamma = 0.5$

The Service component D_{ik}^S is calculated using:

$$D_{ik}^S = \frac{1}{2} \left(\frac{abs(n_1 - n_k)}{C} + \left(1 - \frac{|X_i \cap X_k|}{|X_i \cup X_k|} \right) \right) \quad (5.20)$$

We took $C = 6$, representing the largest difference in path length within our dataset, (although this can be adjusted, based on other data). The D_{ik}^T is the same as before.

This dissimilarity aims for dissimilarity scores to fall within the range of 0 to 1. The square root term serves the same purpose as in the Multiplicative Dissimilarity, preventing the measure from approaching zero too quickly.

5.2.4 Complex Dissimilarity

Building on earlier discussions, we now introduce a highly nuanced dissimilarity measure—named by us as the Complex Dissimilarity — that synthesizes all three essential factors elaborated at the beginning of this chapter. These elements are: the specific services a patient went through, the time spent at each service (LoS_s), and the sequential order in which these services were availed.

Arguably, the most innovative aspect of this model is the integration of a dedicated dissimilarity matrix for each service, D . This step was enabled through close consultation with the hospital staff, whose specialized knowledge about the hospital's inner workings, particularly in the pediatric ward, was invaluable in quantifying the nuanced differences between services.

	CTS	PedSurg	PedICM	NEO	PedOnc	ORL	Ped	Nue	STU	Z
CTS	0,00	0,25	0,50	0,50	0,50	0,50	0,75	1,00	1,00	1,00
PedSurg	0,25	0,00	0,75	0,50	0,50	0,25	0,75	1,00	1,00	1,00
PedICM	0,50	0,75	0,00	0,25	0,75	1,00	0,75	1,00	1,00	1,00
NEO	0,50	0,50	0,25	0,00	0,50	0,75	0,75	1,00	1,00	1,00
PedOnc	0,50	0,50	0,75	0,50	0,00	0,75	0,50	1,00	1,00	1,00
ORL	0,50	0,25	1,00	0,75	0,75	0,00	0,50	1,00	1,00	1,00
Ped	0,75	0,75	0,75	0,75	0,50	0,50	0,00	1,00	1,00	1,00
Nue	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00	1,00	1,00
STU	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00	1,00
Z	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00

FIGURE 5.7: Physical Services' Dissimilarity Matrix - D

The Complex Dissimilarity consists of three principal components: D_{ik}^S related to the services, D_{ik}^t concerning the LoS_s at each service, and D_{ik}^N which captures the order of the services. The LoS-related term is bifurcated into two sub-components: D_{ik}^{t1} and D_{ik}^{t2} .

$$D_{ik}^C = \alpha D_{ik}^S + \beta D_{ik}^t + \gamma D_{ik}^N \quad (5.21)$$

$$D_{ik}^{C'} = \alpha' D_{ik}^S + \beta D_{ik}^t + \gamma D_{ik}^N \quad (5.22)$$

Where $\alpha = \frac{1}{n}$ and $\alpha' = \frac{1}{n_i + n_k}$, $\beta = 0.5$ and $\gamma = 0.2$

5.2.4.1 Service term

In the first term, D_{ik}^S , we sum all possible combinations of the dissimilarity between services of the two paths being compared. Any overlapping services are then eliminated from this calculation to avoid artificially inflating the dissimilarity measure. The rationale here is straightforward: the more service combinations, the greater the dissimilarity

meaning that there are many different services between the two paths. If both paths feature the same set of services, this component should naturally gravitate towards zero. To normalize the impact of this term, especially for paths with a higher number of services, a scaling constant - either $\frac{1}{n}$ or $\frac{1}{n_i+n_k}$ is applied.

$$D_{ik}^S = \sum_{j=1}^{n_i} \sum_{m=1}^{n_k} d(S_{ij}, S_{km}) - \sum_{\substack{1 \leq j \leq n_i \\ 1 \leq m \leq n_k \\ S_{ij}, S_{km} \in I}} d(S_{ij}, S_{km}) \quad (5.23)$$

Here: $d(S_{ij}, S_{km})$ is the dissimilarity between service S_{ij} in X_i and S_{km} in X_k , and $I = \{S_{ij}\}_{j=1,2,3,\dots,n_i} \cap \{S_{km}\}_{m=1,2,3,\dots,n_m}$

5.2.4.2 LoS term

We divided the LoS-based component of our dissimilarity measure, D_{ik}^t , it into two distinct parts, D_{ik}^{t1} and D_{ik}^{t2} , to capture the nuances of time spent in each service:

$$D_{ik}^t = D_{ik}^{t1} + D_{ik}^{t2} \quad (5.24)$$

5.2.4.3 LoS term - Equal Services

The term, D_{ik}^{t1} , compares each standardized Length of Stay (LoS_s) of the common services in both paths X_i and X_k , based on the order of their appearances. For example, looking at Figure 5.6, *PedOnc* appears twice in X_i and once in X_k , so we will only compute the absolute value of the difference of LoS_s of *PedOnc* $j = 2$ in X_i and *PedOnc* $m = 4$ in X_k , i.e., $|t_{i2} - t_{k4}|$, because they appear in this order. The *PedOnc* $j = 4$ in X_i hasn't another *PedOnc* in X_k to be compared to. Similarly *PedICM* appears once in both paths in position $j = 5$ and $m = 2$ so we will compute $|t_{i5} - t_{k2}|$.

$$D_{ik}^{t1} = \sum_{(j,m) \in Q} \text{abs}(t_{ij} - t_{km}) \quad (5.25)$$

Here Q is a set containing pairs of indices (j, m) representing the positions of the occurrences of each common service in order, in paths X_i and X_k . The algorithm to create Q is in the Appendix A

5.2.4.4 LoS term - Similar Services

After calculating the dissimilarity term associated with "Equal Services" D_{ik}^{t1} , we incorporate a term based on "Similar Services", D_{ik}^{t2} . This term aims to account for the

LoS_s between services that, while not identical, share substantial similarities according to a predefined service dissimilarity matrix, D . The steps for calculating this term are as follows:

1. **Remaining Standardized LoS:** After accounting for "Equal Services," focus on the remaining services in paths X_i and X_k .
2. **Identify Most Similar Services:** Using D , find the most similar pairs of services between X_i and X_k .
3. **Calculate Absolute Differences:** For the most similar pair of services, compute the absolute difference of their respective standardized LoS value.
4. **Tie-breaking Rules:** In cases where there are equally similar pairs of services, the following tie-breaking rules are applied:
 - (a) **Sum of Indices:** The pair with the smallest sum of indices $j + m$ is chosen.
 - (b) **Swapped Indices:** If the sum is the same and the indices are swapped (e.g., $(j, m) = (3, 4)$ and $(j, m) = (4, 3)$), the pair with the smallest j is chosen. (This is arbitrary because it doesn't influence the dissimilarity i.e. $d(X_i, X_k) = d(X_k, X_i)$)
 - (c) **Non-swapped Indices:** If the sum is the same but the indices are not swapped (e.g., $(j, m) = (1, 5)$ and $(j, m) = (2, 4)$), the pair with the smallest individual index, either j or m , is chosen.

By including this "Similar Services" term, our dissimilarity measure gains an additional layer of sophistication, accounting not only for identical services but also for similar services, thereby offering a more comprehensive understanding of path dissimilarities:

$$D_{ik}^{t_2} = \sum_{(j,m) \in P} \text{abs}(t_{ij} - t_{km}) \quad (5.26)$$

In this equation, P is a set containing pairs of indices (j, m) that correspond to the first occurrences of the most similar services in X_i and X_k . The algorithm to create P is present in [Appendix A](#)

5.2.4.5 Needleman term

The last term of the Complex Dissimilarity, D_{ik}^N , leverages the Needleman Algorithm. This algorithmic addition ensures that even paths with identical services and identical LoS_s aren't deemed completely equivalent (i.e., a score of zero) in our Dissimilarity. It offers a layer of granular differentiation based on the sequence in which services are utilized, thus rounding out our comprehensive approach to measuring dissimilarity between patient paths.

The Needleman-Wunsch algorithm, commonly referred to simply as the Needleman algorithm, is a foundational algorithm in bioinformatics for sequence alignment. Developed by Saul B. Needleman and Christian D. Wunsch in 1970 [14], this algorithm provides a dynamic programming approach for optimally aligning pairs of sequences. It is particularly useful in comparing biological sequences like DNA, RNA, or protein sequences, but its applications are not limited to these areas and can be adapted to any sequence-based analysis.

At the core of the Needleman algorithm is the concept of a similarity matrix, also known as a substitution matrix. This matrix contains the scores of aligning each pair of elements from the sequences under consideration. In bioinformatics, matrices like PAM (Point Accepted Mutation) or BLOSUM (Blocks Substitution Matrix) are often used. The scores in the similarity matrix serve as a quantitative measure of how alike or dissimilar two elements are. For example, in the case of DNA sequences, the similarity matrix could contain the scores for matching 'A' with 'A', 'A' with 'T', and so on. The choice of the similarity matrix often depends on the specific context in which the algorithm is applied.

The Needleman algorithm utilizes dynamic programming to efficiently calculate the optimal alignment score between two sequences. The algorithm initializes a scoring matrix with dimensions based on the lengths of the two sequences and then fills it in a manner that accounts for matches, mismatches, and gaps. To do this, the algorithm compares elements in the sequences using the scores from the similarity matrix and also includes penalties for gaps. The final score of the optimal alignment is found in the last cell of the scoring matrix, and the optimal alignment can be reconstructed by backtracking through this matrix.

The traceback starts at the cell containing the optimal score, the bottom-right cell and works its way back to the starting cell, top-left corner, by following pointers, or "arrows," placed in the matrix. These pointers indicate whether a cell's score was calculated from a

diagonal (match/mismatch), vertical (gap in sequence 1), or horizontal (gap in sequence 2) move. Depending on the scoring and the sequences themselves, it's possible for multiple traceback paths to exist, each corresponding to a different optimal alignment with the same optimal score, making it necessary to choose between them based on additional criteria or preferences.

In the context of measuring dissimilarity between paths within a hospital, the Needleman algorithm has been adapted to suit the unique requirements of the data. The first step involves transforming the existing dissimilarity matrix between services, D , into a similarity matrix $S = 1 - D$.

To facilitate the use of the Needleman algorithm, the acronyms for the hospital services are then mapped to single letters. For example, 'CTS' is mapped to 'A', 'PedSur' to 'B', and so on. Each path can therefore be represented as a string of these mapped letters, succinctly encapsulating the sequence of services that a patient went through.

To align the paths for comparison, each string is padded with the letter 'Z' until it reaches a length of 7, which is the maximum number of services that a path could involve in our data, after the removal of the path that included services with "extreme" LoS already mentioned.

The Needleman score is then computed for each pair of padded strings for the two strings without changing the alignment and for the best alignment. To transform each score into a dissimilarity is straightforward, $1 - \left(\frac{\text{score}}{7}\right)$.

So the Needleman term for our Complex Dissimilarity is the difference between the dissimilarities of both scores:

$$D_{ik}^N = 1 - \left(\frac{\text{score}}{7}\right) - \left(1 - \left(\frac{\text{Score}}{7}\right)\right) \quad (5.27)$$

Where "score" is the score for the not aligned strings and "Score" is the score of the best alignment between the two strings. Resulting in this:

$$D_{ik}^N = \frac{\text{Score} - \text{score}}{7} \quad (5.28)$$

This ensures that completely identical paths have a dissimilarity score of zero, thereby satisfying the intuitive notion that identical paths are not dissimilar and falls between 0 and 1 to be easy to understand and use.

5.3 Results and Discussion

After formulating the various dissimilarities, we applied them to our dataset of 3133 paths to generate individual dissimilarity matrices. These matrices were then subjected to clustering using the PAM algorithm.

To determine the optimal number of clusters for each dissimilarity measure, we employed the Average Silhouette Width (ASW) and the Elbow technique [15].

The ASW values generally showed an increasing trend with the number of clusters, continuing at least up to 12 clusters, however, in some of them there was an almost flat trend, with small oscillations. [B](#)

It's worth noting that the ASW values for the Product Dissimilarity were significantly low, rendering it unsuitable for further discussion or analysis in this context. [B.3](#) and [B.4](#)

We opted for a 7-cluster configuration to maintain consistency with the previous clustering analysis. This choice was further justified as the 7-cluster model exhibited moderate -to-high ASW values (although not the highest) for most dissimilarities. Additionally, upon conducting a cross-analysis of tables featuring different numbers of clusters, we determined that the 7-cluster arrangement was the most stable.

When utilizing the Adjusted Rand Index (ARI) to compare each dissimilarity measure with its corresponding variation, the ARI scores for the Additive, Product, and Jaccard-Like Dissimilarities were 0.985, 0.776, and 0.687, respectively.

In a separate comparison where each dissimilarity measure was juxtaposed with the Complex Dissimilarity, the ARI scores were as follows: for the Additive and Additive' dissimilarities, the scores were 0.821 and 0.820, respectively; for Product and Product' dissimilarities, the values were 0.078 and 0.093. Lastly, for Jaccard_A and Jaccard_P, the scores stood at 0.810 and 0.676, respectively.

A similar description of the one presented previously is given below for the Additive, Jaccard_A and the Complex (with $\alpha = \frac{1}{n}$) dissimilarity:

Additive	5			3			1			6			7			4			2		
Variable	N	Percent		N	Percent		N	Percent		N	Percent		N	Percent		N	Percent		N	Percent	
Surgery	380			750			639			278			191			283			612		
... No	71	19%		254	34%		178	28%		35	13%		181	95%		246	87%		562	92%	
... Yes	309	81%		496	66%		461	72%		243	87%		10	5%		37	13%		50	8%	
Intensive Care	380			750			639			278			191			283			612		
... No	319	84%		711	95%		601	94%		278	100%		183	96%		266	94%		590	96%	
... Yes	61	16%		39	5%		38	6%		0	0%		8	4%		17	6%		22	4%	
Admission Type	380			750			639			278			191			283			612		
... Normal	0	0%		0	0%		0	0%		0	0%		174	91%		0	0%		0	0%	
... Scheduled	44	12%		521	69%		30	5%		242	87%		6	3%		226	80%		176	29%	
... Urgent	336	88%		229	31%		609	95%		36	13%		11	6%		57	20%		436	71%	
Number of Services	380			750			639			278			191			283			612		
... 1	0	0%		695	93%		559	87%		277	100%		0	0%		271	96%		581	95%	
... 2	338	89%		23	3%		72	11%		1	0%		134	70%		9	3%		10	2%	
... +2	42	11%		32	4%		8	1%		0	0%		57	30%		3	1%		21	3%	
Weekend	380			750			639			278			191			283			612		
... No	278	73%		661	88%		466	73%		206	74%		148	77%		231	82%		492	80%	
... Yes	102	27%		89	12%		173	27%		72	26%		43	23%		52	18%		120	20%	
Discharge Type	380			750			639			278			191			283			612		
... Discharge	374	98%		743	99%		472	74%		269	97%		180	94%		256	90%		595	97%	
... Deceased	1	0%		1	0%		3	0%		1	0%		8	4%		5	2%		5	1%	
... Hospital	5	1%		6	1%		164	26%		8	3%		3	2%		22	8%		12	2%	
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
TLoS	380	125,7	117,4	750	89,1	78,3	639	68,2	149,2	278	38,0	32,0	191	319,6	226,4	283	100,5	94,9	612	144,6	107,9

TABLE 5.1: Additive clusters described by Hospital Resources

Jaccard_A	5			3			6			7			1			4			2		
Variable	N	Percent		N	Percent		N	Percent		N	Percent		N	Percent		N	Percent		N	Percent	
Surgery	387			700			552			278			265			234			717		
... No	62	16%		229	33%		122	22%		33	12%		232	88%		211	90%		638	89%	
... Yes	325	84%		471	67%		430	78%		245	88%		33	12%		23	10%		79	11%	
Intensive Care	387			700			552			278			265			234			717		
... No	302	78%		689	98%		547	99%		277	100%		257	97%		228	97%		648	90%	
... Yes	85	22%		11	2%		5	1%		1	0%		8	3%		6	3%		69	10%	
Admission Type	387			700			552			278			265			234			717		
... Normal	0	0%		0	0%		0	0%		0	0%		174	66%		0	0%		0	0%	
... Scheduled	65	17%		489	70%		6	1%		241	87%		81	31%		183	78%		180	25%	
... Urgent	322	83%		211	30%		546	99%		37	13%		10	4%		51	22%		537	75%	
Number of Services	387			700			552			278			265			234			717		
... 1	0	0%		691	99%		545	99%		278	100%		62	23%		229	98%		578	81%	
... 2	307	79%		9	1%		7	1%		0	0%		142	54%		3	1%		119	17%	
... +2	80	21%		0	0%		0	0%		0	0%		61	23%		2	1%		20	3%	
Weekend	387			700			552			278			265			234			717		
... No	282	73%		624	89%		397	72%		208	75%		206	78%		194	83%		571	80%	
... Yes	105	27%		76	11%		155	28%		70	25%		59	22%		40	17%		146	20%	
Discharge Type	387			700			552			278			265			234			717		
... Discharge	381	98%		692	99%		392	71%		265	95%		238	90%		228	97%		693	97%	
... Deceased	2	1%		1	0%		1	0%		1	0%		10	4%		3	1%		6	1%	
... Hospital	4	1%		7	1%		159	29%		12	4%		17	6%		3	1%		18	3%	
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
TLoS	387	132,4	133,5	700	78,3	67,7	552	21,1	10,8	576	32,6	13,1	265	302,8	235,8	234	106,8	95,8	717	160,0	119,6

TABLE 5.2: Jaccard.A clusters described by Hospital Resources

Complex	7		3		5		6		1		4		2								
Variable	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent							
Surgery	276		794		599		275		187		276		726								
... No	28	10%	257	32%	155	26%	36	13%	174	93%	235	85%	641	88%							
... Yes	248	90%	537	68%	444	74%	239	87%	13	7%	41	15%	85	12%							
Intensive Care	276		794		599		275		187		276		726								
... No	273	99%	725	91%	593	99%	275	100%	163	87%	269	97%	650	90%							
... Yes	3	1%	69	9%	6	1%	0	0%	24	13%	7	3%	76	10%							
Admission Type	276		794		599		275		187		276		726								
... Normal	0	0%	1	0%	0	0%	0	0%	149	80%	0	0%	24	3%							
... Scheduled	0	0%	568	72%	5	1%	233	85%	30	16%	220	80%	189	26%							
... Urgent	276	100%	225	28%	594	99%	42	15%	8	4%	56	20%	513	71%							
Number of Services	276		794		599		275		187		276		726								
... 1	0	0%	708	89%	541	90%	267	97%	38	20%	262	95%	567	78%							
... 2	269	97%	28	4%	56	9%	8	3%	119	64%	7	3%	100	14%							
... +2	7	3%	58	7%	2	0%	0	0%	30	16%	7	3%	59	8%							
Weekend	276		794		599		275		187		276		726								
... No	206	75%	689	87%	432	72%	203	74%	150	80%	226	82%	576	79%							
... Yes	70	25%	105	13%	167	28%	72	26%	37	20%	50	18%	150	21%							
Discharge Type	276		794		599		275		187		276		726								
... Discharge	273	99%	791	100%	439	73%	267	97%	152	81%	255	92%	712	98%							
... Deceased	0	0%	1	0%	1	0%	0	0%	17	9%	3	1%	2	0%							
... Hospital	3	1%	2	0%	159	27%	8	3%	18	10%	18	7%	12	2%							
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD						
TLoS	276	89,0	70,5	794	94,9	83,2	599	36,6	61,8	275	37,0	29,3	187	308,3	243,3	276	109,1	109,0	726	174,8	152,1

TABLE 5.3: Complex clusters described by Hospital Resources

It is important to underscore the efficacy of these dissimilarity measures in cluster formation. Remarkably, without explicitly incorporating variables like surgery type, exit type, or even specific hospital departments, these dissimilarity measures were able to distinctly segregate the data into seven well-defined clusters. The only variables implicitly accounted for were the Number of Services and Total Length of Stay (TLoS). This achievement suggests that the formulated dissimilarities possess a high level of sensitivity and robustness, effectively capturing underlying patterns and classifications based on hospital resource utilization. The ability to produce such distinct and meaningful clusters with limited input variables is an encouraging outcome, highlighting the potential utility of these dissimilarity measures in healthcare data analysis.

Additive	5		3		1		6		7		4		2	
Variable	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Sex	380		750		639		278		191		282		612	
... Female	141	37%	295	39%	210	33%	107	38%	77	40%	160	57%	274	45%
... Male	239	63%	455	61%	429	67%	171	62%	114	60%	122	43%	338	55%
Age_Category	380		750		639		278		191		282		612	
... Newborn	4	1%	14	2%	16	3%	6	2%	183	96%	29	10%	37	6%
... Baby	17	4%	61	8%	40	6%	2	1%	1	1%	16	6%	132	22%
... Toddler	62	16%	196	26%	128	20%	126	45%	3	2%	87	31%	166	27%
... Child	104	27%	158	21%	167	26%	81	29%	3	2%	44	16%	120	20%
... Teenager	193	51%	321	43%	288	45%	63	23%	1	1%	106	38%	157	26%
Distrit_Residence	380		750		639		278		191		282		612	
... Porto	266	70%	523	70%	464	73%	220	79%	158	83%	143	51%	439	72%
... Other	114	30%	227	30%	175	27%	58	21%	33	17%	139	49%	173	28%
Country_Birth	380		750		639		278		191		282		612	
... Portugal	362	95%	731	97%	611	96%	272	98%	190	99%	271	96%	596	97%
... Other	18	5%	19	3%	28	4%	6	2%	1	1%	11	4%	16	3%

TABLE 5.4: Additive clusters described by Patient Sociodemographic Variables

Jaccard_A	5		3		6		7		1		4		2	
Variable	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Sex	387		700		552		278		264		234		717	
... Female	143	37%	268	38%	173	31%	109	39%	101	38%	141	60%	329	46%
... Male	244	63%	432	62%	379	69%	169	61%	163	62%	93	40%	388	54%
Age_Category	387		700		552		278		264		234		717	
... Newborn	1	0%	2	0%	2	0%	4	1%	246	93%	0	0%	34	5%
... Baby	18	5%	52	7%	31	6%	4	1%	11	4%	4	2%	149	21%
... Toddler	69	18%	182	26%	99	18%	125	45%	1	0%	87	37%	205	29%
... Child	112	29%	149	21%	152	28%	81	29%	2	1%	40	17%	141	20%
... Teenager	187	48%	315	45%	268	49%	64	23%	4	2%	103	44%	188	26%
Distrit_Residence	387		700		552		278		264		234		717	
... Porto	260	67%	503	72%	408	74%	220	79%	190	72%	116	50%	516	72%
... Other	127	33%	197	28%	144	26%	58	21%	74	28%	118	50%	201	28%
Country_Birth	387		700		552		278		264		234		717	
... Portugal	367	95%	683	98%	528	96%	273	98%	262	99%	224	96%	696	97%
... Other	20	5%	17	2%	24	4%	5	2%	2	1%	10	4%	21	3%

TABLE 5.5: Jaccard_A clusters described by Patient Sociodemographic Variables

Complex	7		3		5		6		1		4		2	
Variable	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
Sex	276		794		599		275		187		275		726	
... Female	95	34%	307	39%	197	33%	108	39%	79	42%	155	56%	323	44%
... Male	181	66%	487	61%	402	67%	167	61%	108	58%	120	44%	403	56%
Age_Category	276		794		599		275		187		275		726	
... Newborn	0	0%	22	3%	3	1%	0	0%	161	86%	27	10%	76	10%
... Baby	6	2%	60	8%	33	6%	1	0%	9	5%	15	5%	145	20%
... Toddler	31	11%	209	26%	121	20%	125	45%	5	3%	85	31%	192	26%
... Child	85	31%	165	21%	160	27%	85	31%	5	3%	43	16%	134	18%
... Teenager	154	56%	338	43%	282	47%	64	23%	7	4%	105	38%	179	25%
Distrit_Residence	276		794		599		275		187		275		726	
... Porto	204	74%	544	69%	443	74%	221	80%	139	74%	136	49%	526	72%
... Other	72	26%	250	31%	156	26%	54	20%	48	26%	139	51%	200	28%
Country_Birth	276		794		599		275		187		275		726	
... Portugal	262	95%	773	97%	573	96%	270	98%	186	99%	265	96%	704	97%
... Other	14	5%	21	3%	26	4%	5	2%	1	1%	10	4%	22	3%

TABLE 5.6: Complex clusters described by Patient Sociodemographic Variables

In examining the sociodemographic variables, a comparable pattern is observed in the “district” variable, albeit solely within one cluster, as seen in the preceding analysis. The district distribution is nearly balanced, with 50% originating from Porto and the remaining 50% from other districts. This pattern is conspicuous in the cluster characterized by the schedule, simple, non-surgical, and non-intensive care paths. Additionally, the inclusion of the nursery compound paths in this analysis led to their collective grouping, particularly in the Additive Dissimilarity.

In the following pages the directed graph illustrates the seven clusters for the three dissimilarities: Additive, Jaccard_A, and Complex.

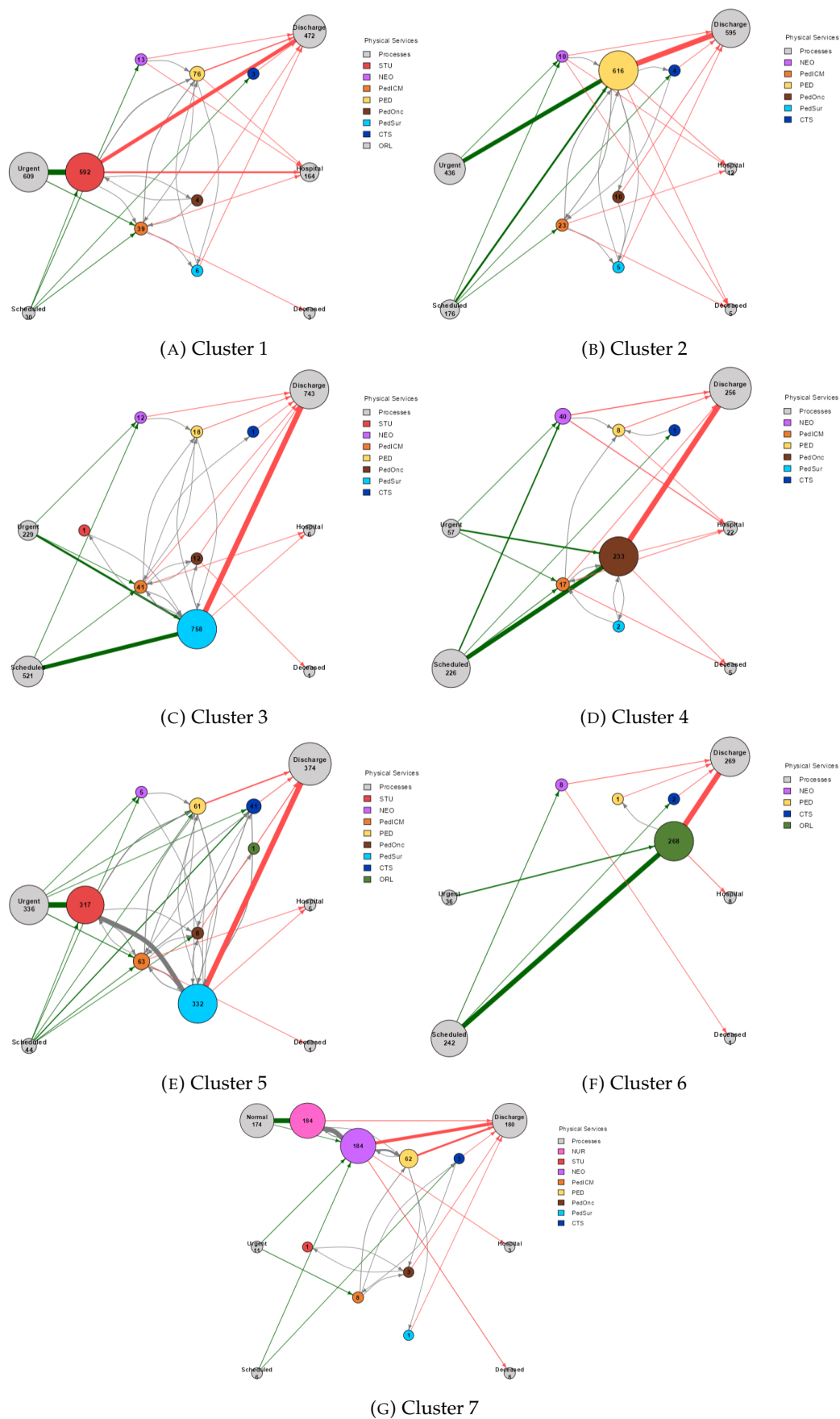


FIGURE 5.8: The 7 clusters of the Additive Dissimilarity represented by directed graphs

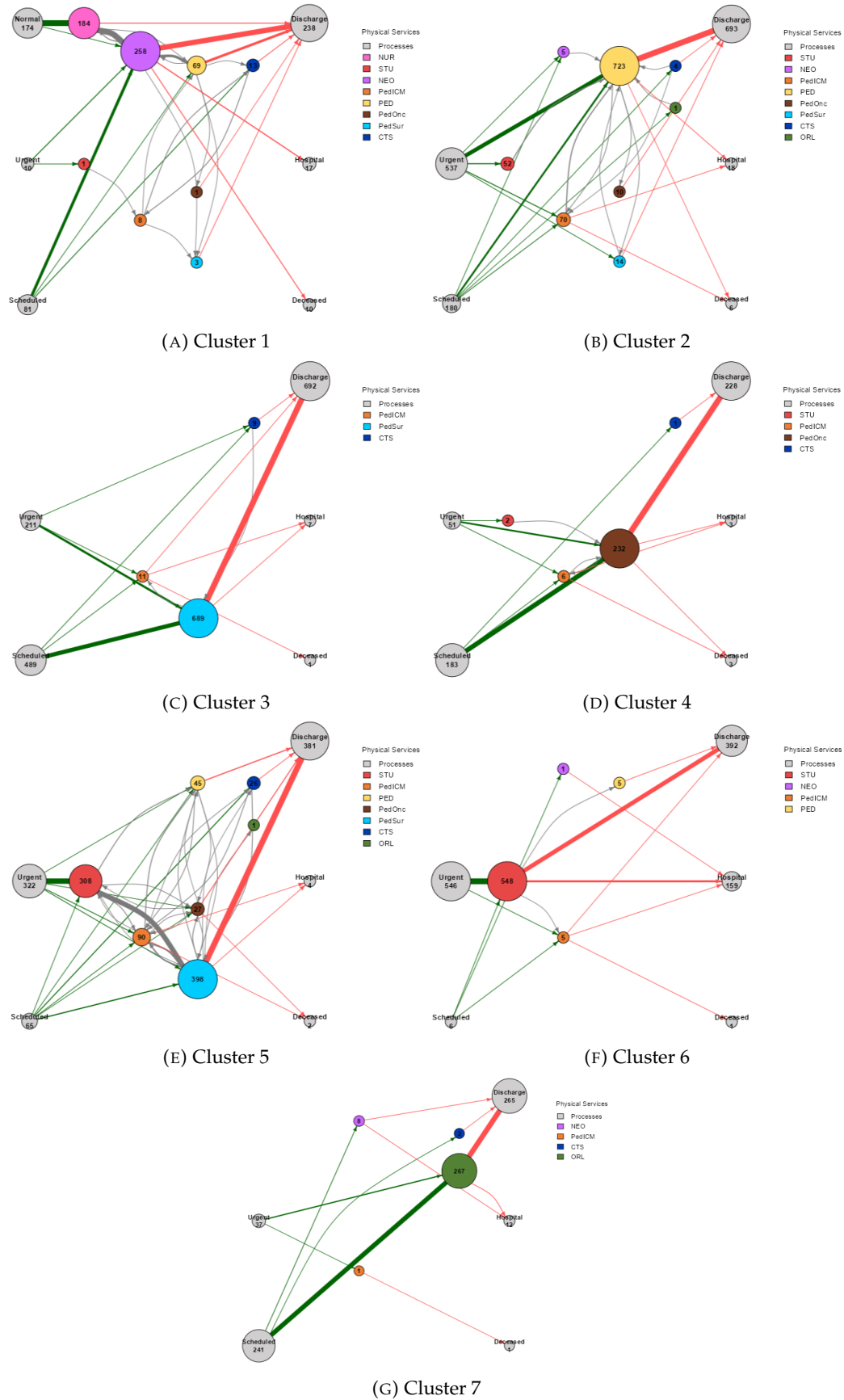


FIGURE 5.9: The 7 clusters of the Jaccard_A Dissimilarity represented by directed graphs

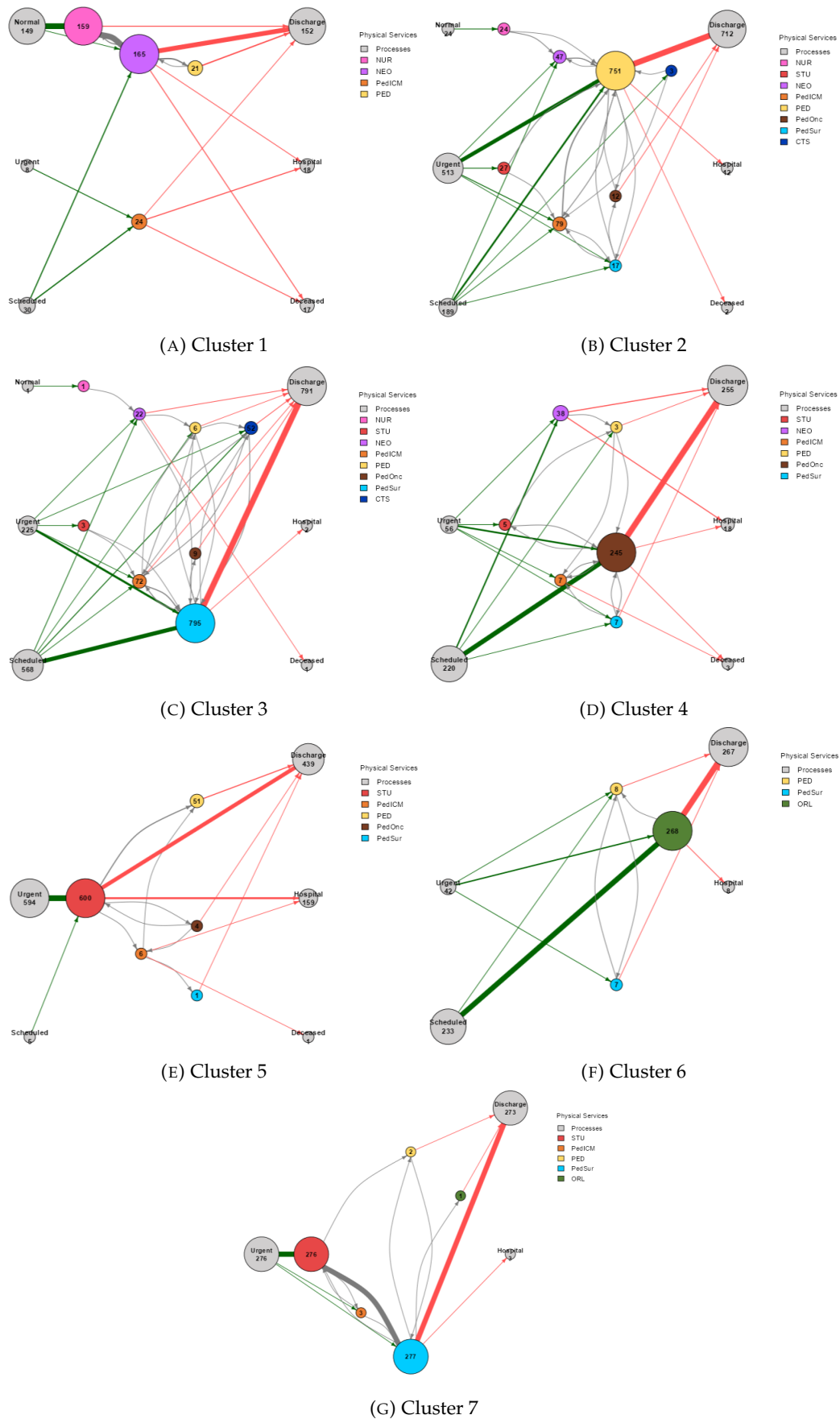


FIGURE 5.10: The 7 clusters of the Complex Dissimilarity represented by directed graphs

Chapter 6

Future Work

6.1 Time Series

One of the intriguing questions that arises in hospital operations is the predictability of service demand over time. Understanding this could provide crucial insights into resource allocation and service optimization. Seasonal Autoregressive Integrated Moving Average (SARIMA) models offer a robust method for time-series forecasting and could be instrumental in predicting the demand for various responsible services within the hospital. By analyzing past data, SARIMA could help forecast trends and seasonal patterns, aiding in proactive decision-making. Future work could involve the extensive use of SARIMA models to evaluate the demand of specific services, thus enabling better resource planning and potentially improving patient care.

6.2 Modelling of the Total Length of Stay

One of the primary aims of this study was to model the Length of Stay (LoS) at individual services. A logical extension of this work involves investigating the Total Length of Stay (TLoS) for path. The variables employed for such a model would differ somewhat from those used in the LoS per service model. A particularly compelling question to address is the impact of discrepancies between the physical service and the responsible service on TLoS. In our LoS model, we observed a counter intuitive effect: when the responsible service differed from the physical service, the LoS was generally shorter. This phenomenon may be attributed to the rapid relocation of the patient to the appropriate

physical service, however, this might have the opposite effect in the TLoS. Understanding the nuanced interplay between these variables in the context of TLoS could provide invaluable insights into hospital operations and identify avenues for potential optimization.

6.3 Markov Models

The dynamism inherent in patient pathways through various services in a hospital environment poses an interesting question: can we predict the probability of a patient transitioning to a specific service at time $t + 1$ given their current service at time t along with other socio-demographic variables and admission type? Utilizing Markov models could be a promising approach to tackle this question. By constructing a Markov chain where each state represents a service, and transitions are influenced by a variety of factors such as age, sex, type of admission, etc., we could potentially predict the next service a patient will visit.

Another intriguing avenue to explore would be the application of Continuous-Time Markov Chains (CTMC). Given that we also have data on the length of stay (LoS) at each service, CTMC can incorporate this temporal dimension, thus enriching the predictive model. By associating rates of transitioning with the LoS, the model could yield more nuanced insights into how quickly patients are likely to need to move from one service to another, providing a valuable layer of information for hospital management in terms of resource planning and allocation.

Appendix A

Algorithms

Algorithms for Sets Q and P

Algorithm 1 Algorithm for Q

```
1: Start  $Q = \emptyset$ 
2:  $J = (1, 2, \dots, j, \dots, n)$ 
3:  $M = (1, 2, \dots, m, \dots, n)$ 
4: for  $j \in J$  do
5:   for  $m \in M$  do
6:     if  $S_{ij} = S_{km}$  then
7:        $Q \leftarrow Q \cup \{(j, m)\}$ 
8:        $J \leftarrow J \setminus \{j\}$ 
9:        $M \leftarrow M \setminus \{m\}$ 
10:      break
11:    end if
12:  end for
13: end for
```

Algorithm 2 Algorithm for P

```

1: Initialize  $P = \emptyset$ 
2: while  $J \neq \emptyset$  and  $M \neq \emptyset$  do
3:    $\text{min\_d} = 10$  ▷ Initial Value
4:    $\text{best\_pair} = (\text{None}, \text{None})$ 
5:   for each pair of indices  $(j, m)$  where  $j \in J$  and  $m \in M$  do
6:      $\text{current\_d} = d(S_{ij}, S_{km})$ 
7:     if  $\text{current\_d} < \text{min\_d}$  then
8:        $\text{min\_d} = \text{current\_d}$ 
9:        $\text{best\_pair} = (j, m)$ 
10:    else if  $\text{current\_d} = \text{min\_d}$  then ▷ Handle tie-breaker conditions
11:      if  $(j + m) < (\text{best\_pair}[0] + \text{best\_pair}[1])$  then
12:         $\text{best\_pair} = (j, m)$  ▷ Choose smallest sum
13:      else if  $(j + m) = (\text{best\_pair}[0] + \text{best\_pair}[1])$  then
14:        if  $j = \text{best\_pair}[1]$  and  $m = \text{best\_pair}[0]$  then
15:          if  $j < m$  then
16:             $\text{best\_pair} = (j, m)$ 
17:          else
18:             $\text{best\_pair} = (m, j)$  ▷ Choose smallest j if simetric
19:          end if
20:        else if  $j \text{ or } m < \min(\text{best\_pair})$  then
21:           $\text{best\_pair} = (j, m)$  ▷ Choose smallest j or m
22:        end if
23:      end if
24:    end if
25:  end for
26:   $J \leftarrow J \setminus \{\text{best\_pair}[0]\}$  ▷ Remove the selected indices from sets J and M
27:   $M \leftarrow M \setminus \{\text{best\_pair}[1]\}$ 
28:   $P \leftarrow P \cup \{\text{best\_pair}\}$  ▷ Add the best pair to set P
29: end while

```

Algorithm 3 Algorithm for BUILD

```

1: Initialize Medoids  $\leftarrow$  empty list
2:  $l = 0$ 
3:  $D = \{\}$  ▷ For the first medoid  $m_1$ 
4: for each data point  $x_i$  do
5:    $l = l + 1$ 
6:    $D[l] \leftarrow \sum_{j \neq i} \text{Dist}(x_i, x_j)$ 
7: end for
8:  $m_1 \leftarrow \arg \min_{x_i}(D)$ 
9: Medoids = Medoids  $\cup m_1$ 
▷ For additional medoids  $m_2, m_3, \dots, m_K$ 
10: for  $k = 2$  to  $K$  do
11:    $l = 0$ 
12:    $C = \{\}$ 
13:   for each point  $x_i$  that is not a medoid and every  $x_j$  that is closer to  $x_i$  than to any
       medoid do
14:      $l = l + 1$ 
15:      $C[l] \leftarrow \sum_j \text{Dist}(x_i, x_j)$ 
16:   end for
17:    $m_k \leftarrow \arg \max x_i(C)$ 
18:   Medoids = Medoids  $\cup m_k$ 
19: end for
20: return Medoids

```

Appendix B

Dissimilarities Values

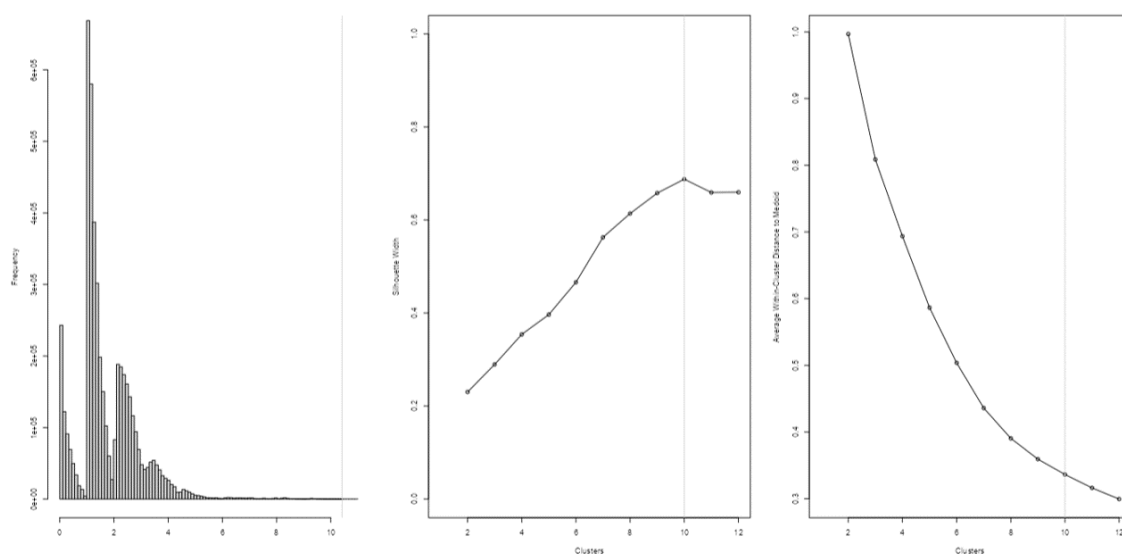


FIGURE B.1: Additive: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Right: Within Cluster dissimilarity to medoid per number of cluster

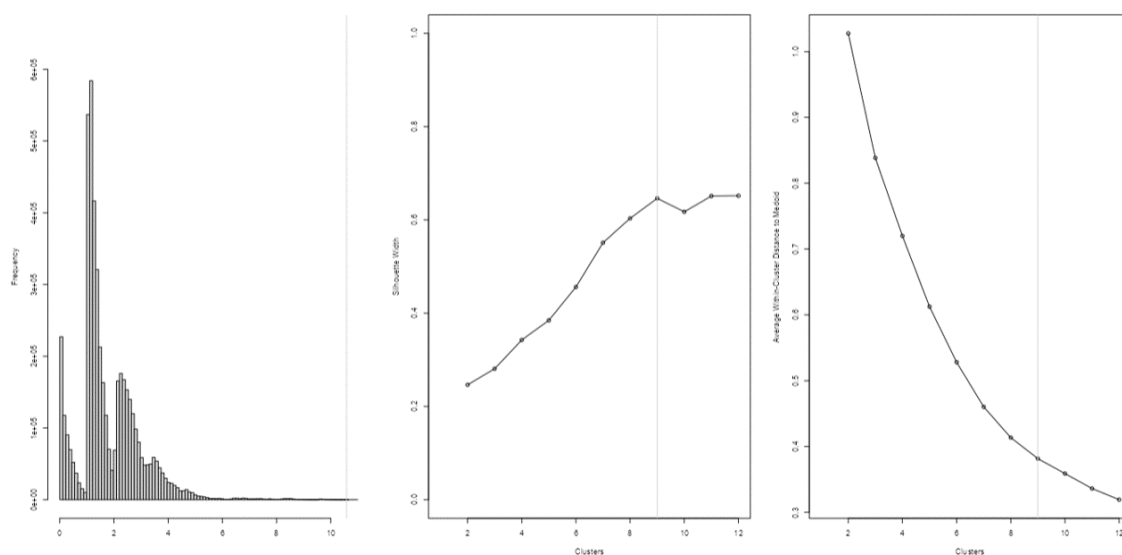


FIGURE B.2: Additive': Left: histogram of dissimilarity values; Center: ASW per number of cluster; Right: Within Cluster dissimilarity to medoid per number of cluster

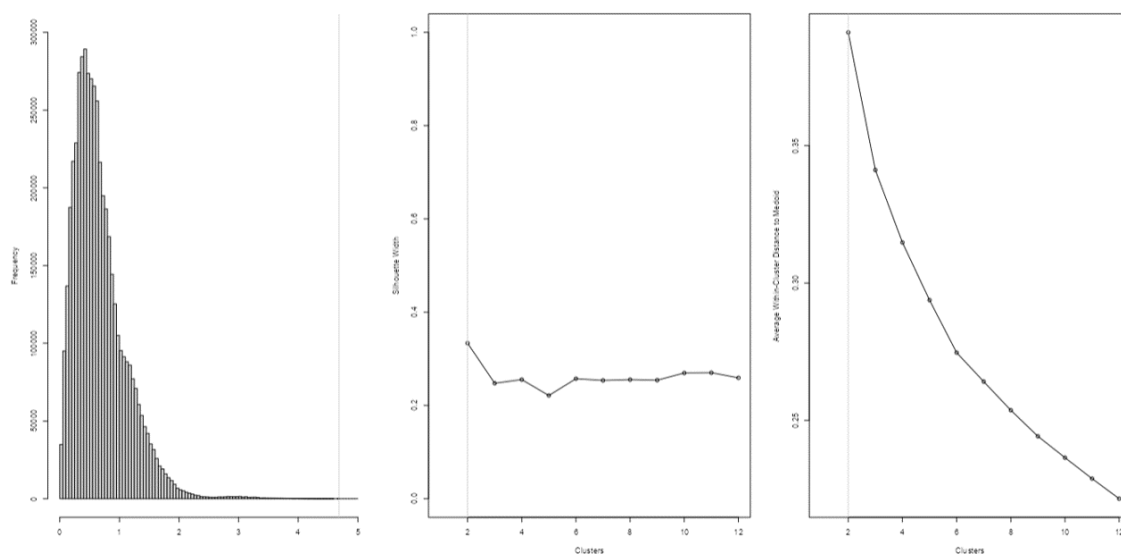


FIGURE B.3: Product: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Right: Within Cluster dissimilarity to medoid per number of cluster

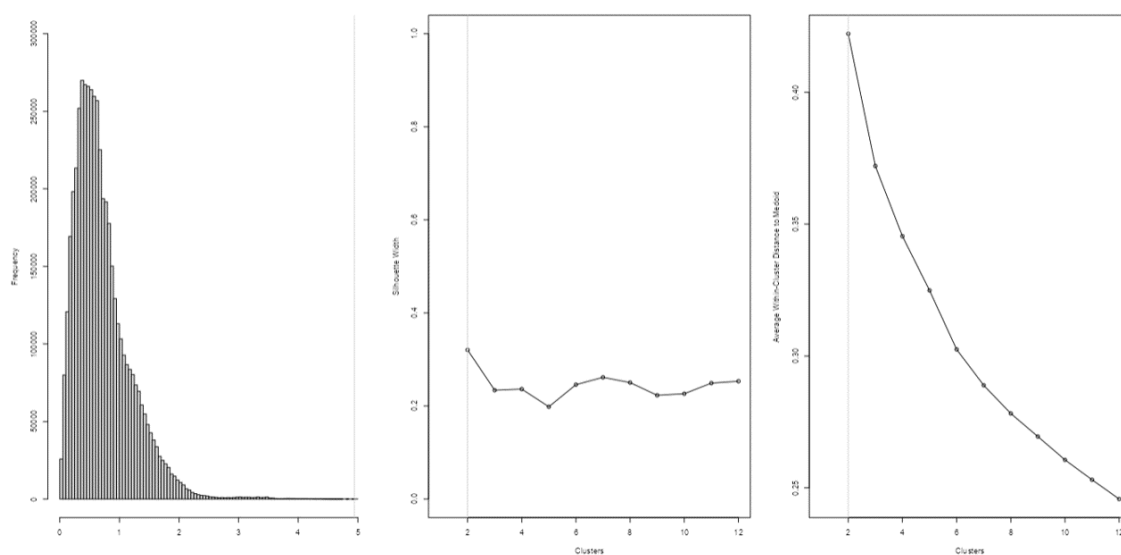


FIGURE B.4: Product': Left: histogram of dissimilarity values; Center: ASW per number of cluster; Right: Within Cluster dissimilarity to medoid per number of cluster

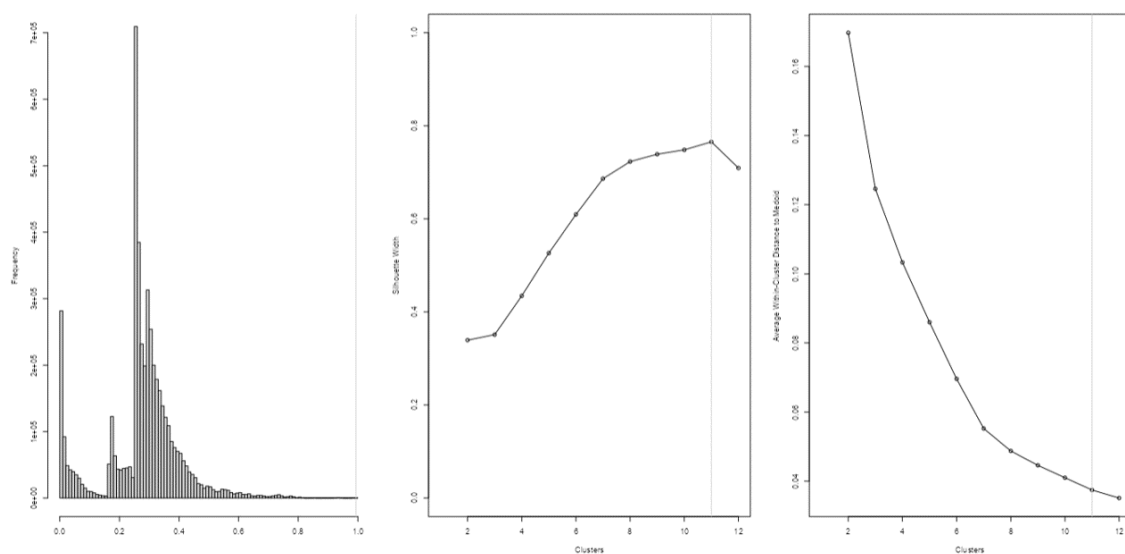


FIGURE B.5: Jaccard_A: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Right: Within Cluster dissimilarity to medoid per number of cluster

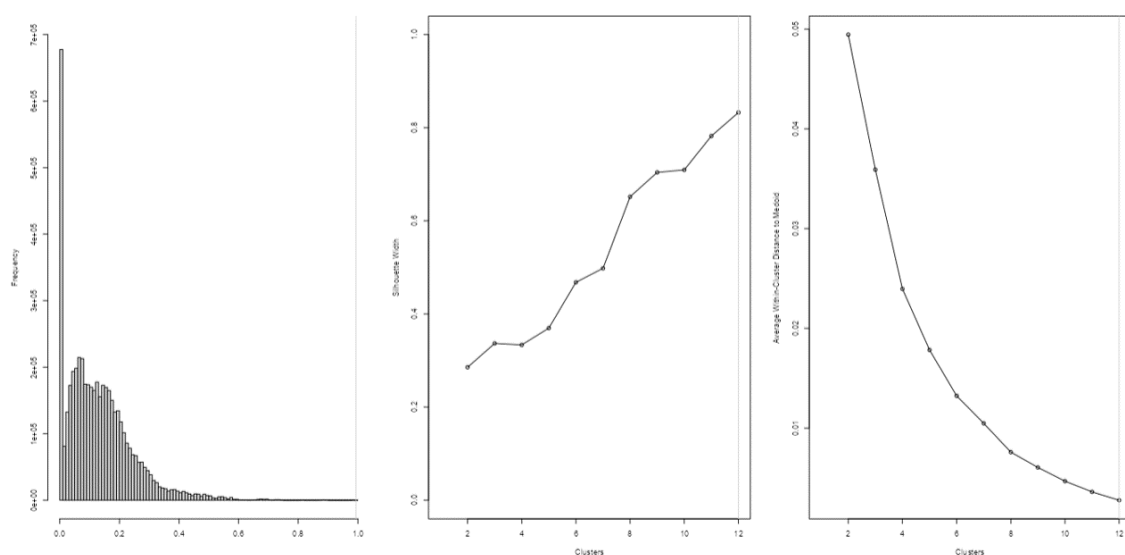


FIGURE B.6: Jaccard_P: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Right: Within Cluster dissimilarity to medoid per number of cluster

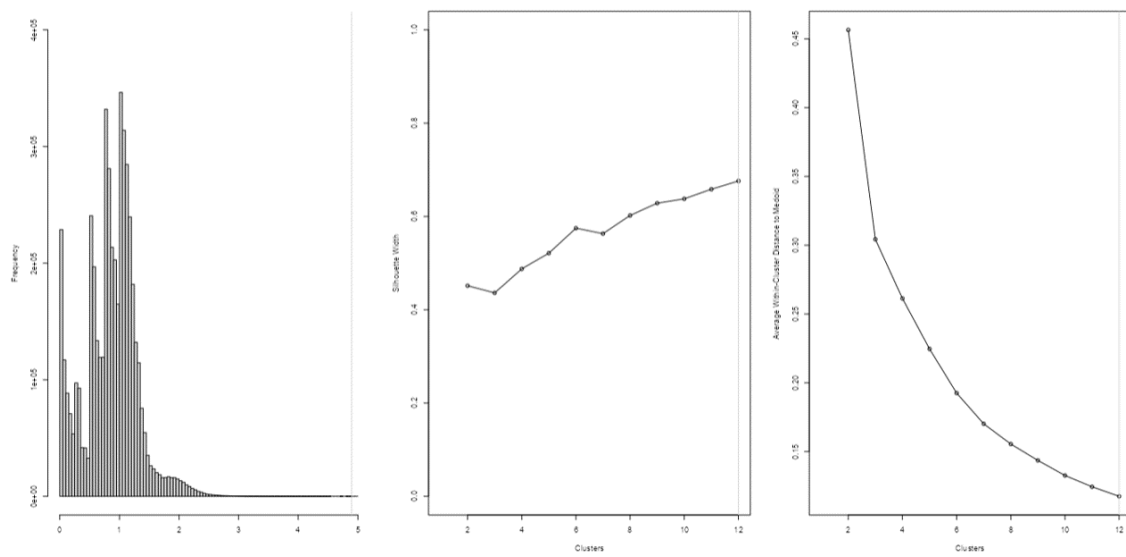


FIGURE B.7: Complex: Left: histogram of dissimilarity values; Center: ASW per number of cluster; Right: Within Cluster dissimilarity to medoid per number of cluster

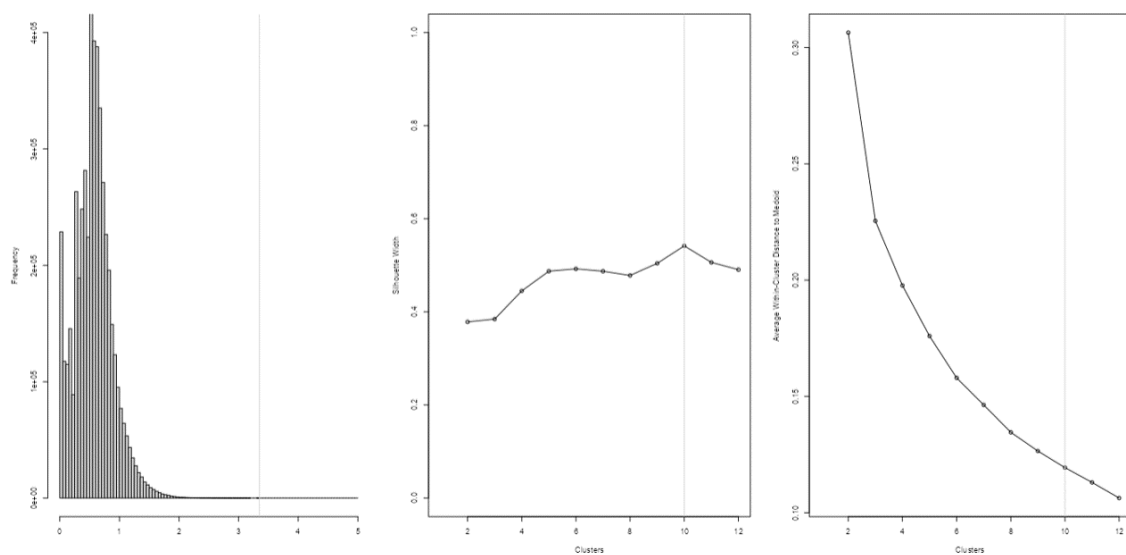


FIGURE B.8: Complex': Left: histogram of dissimilarity values; Center: ASW per number of cluster; Right: Within Cluster dissimilarity to medoid per number of cluster

Bibliography

- [1] T. Amemiya, “Generalized least squares with an estimated autocovariance matrix,” *Econometrica: Journal of the Econometric Society*, pp. 723–732, 1973. [Cited on page [17](#).]
- [2] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied longitudinal analysis*. John Wiley & Sons, 2012. [Cited on page [17](#).]
- [3] J. Pinheiro and D. Bates, *Mixed-effects models in S and S-PLUS*. Springer science & business media, 2006. [Cited on page [17](#).]
- [4] J. P. Shaffer, “Multiple hypothesis testing,” *Annual review of psychology*, vol. 46, no. 1, pp. 561–584, 1995. [Cited on page [26](#).]
- [5] J. C. Gower and P. Legendre, “Metric and euclidean properties of dissimilarity coefficients,” *Journal of classification*, vol. 3, pp. 5–48, 1986. [Cited on page [31](#).]
- [6] J. C. Gower, “A general coefficient of similarity and some of its properties,” *Biometrics*, pp. 857–871, 1971. [Cited on pages [32](#), [34](#), and [39](#).]
- [7] F. de Bello, Z. Botta-Dukát, J. Lepš, and P. Fibich, “Towards a more balanced combination of multiple traits when computing functional differences between species,” *Methods in Ecology and Evolution*, vol. 12, no. 3, pp. 443–448, 2021. [Cited on page [32](#).]
- [8] S. Pavoine, J. Vallet, A.-B. Dufour, S. Gachet, and H. Daniel, “On the challenge of treating various types of variables: application for improving the measurement of functional diversity,” *Oikos*, vol. 118, no. 3, pp. 391–402, 2009. [Cited on page [32](#).]
- [9] K. Agustin and H. Soewandi, “Capturing social value using ordinal data, gower distance & quality point,” Ph.D. dissertation, Petra Christian University, 2022. [Cited on page [32](#).]

- [10] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009. [Cited on pages [32](#) and [35](#).]
- [11] F. Murtagh and P. Legendre, "Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm," *arXiv preprint arXiv:1111.6285*, 2011. [Cited on page [39](#).]
- [12] B. Everitt, *Finite mixture distributions*. Springer Science & Business Media, 2013. [Cited on page [39](#).]
- [13] F. Leisch, "Flexmix: A general framework for finite mixture models and latent glass regression in r," 2004. [Cited on page [39](#).]
- [14] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970. [Cited on pages [47](#) and [57](#).]
- [15] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987. [Cited on page [59](#).]