Université Grenoble Alpes - INP, Ense^3
Filière d'Automatique et Systèmes Intelligents

Internship and Training for 2nd year

Internship final report

Study of demand prediction using clustering methods: application to medication consumption in hospitals

Author: Douglas MATEUS MACHADO
Supervisors: Prof. Dr. Gulgun ALPAN and Prof. Dr. Zakaria YAHOUNI

Grenoble, September 4, 2023.

Contents

Abstract

Lowering operational expenses within the context of public French hospitals is a pressing requirement. Previous research has established a significant connection between the quantity of medicine consumption and these cost reduction efforts. Various factors influence the consumption patterns for each medicine. These factors include the population served by the hospital, the number of beds, the types of treatments provided by the medical staff, and so on. In Koala et al. (2022), twelve primary factors were selected. By collecting and compiling data from various sources, a database was created, comprising information on 21 different medicines across four hospitals. This database encompasses a total of 21 features.Two main approaches were examined and introduced. One approach focused on time series analysis, while the other took a causal perspective. For the time series approach, a shared time period was chosen, ensuring that valid data overlapped for most medicines and hospitals. In the causal method, the entire period featuring valid data (with non-zero values and excluding outliers) was employed. Once the analysis was conducted and the invalid data were addressed, two additional avenues for causal research emerged. The first approach involved a unified strategy where data from all hospitals were combined to generate clusters for consumption prediction. The second approach entailed the creation of four distinct datasets, each corresponding to a different hospital.

Various data clustering techniques (K-means, mini-batch k-means, agglomerative clustering and DBSCAN) were employed, and the resultant clusters were assessed using the Random Trees regressor model. This forecasting technique was found to be better suited for addressing the current problem compared to other methods explored in prior research. In both approaches, whether utilizing the time series or causal method, the clustering techniques did not yield satisfactory forecasting metrics. This outcome suggests that the limited availability of data is affecting the accuracy of the results. Each medicine exhibits such distinct attributes that it possesses its own distinct characteristics. This underscores the notion that alternative clustering techniques should be explored. For instance, the Dynamic Time Warping (DTW) method could be considered for the time series approach.

As a concluding strategy, a comprehensive assessment was conducted by manually testing combinations of hospitals and medicines. This approach yielded a total of 2310 potential pairings (e.g., Medicine A from Hospital X with Medicine B from Hospital Y). For each pair, trends and quantities were plotted, enabling a visual analysis that facilitated the identification of well-fitted pairs. Among these combinations, 206 pairs demonstrated a strong fit. The forecasting technique was then employed, resulting in improved metrics, particularly in terms of the Mean Absolute Percentage Error (MAPE) for certain pairs. Additionally, clusters formed by intersecting medicines and hospitals were subjected to testing, and the predictions from this analysis also yielded favorable MAPE metric values.

Résumé

Réduire les dépenses opérationnelles dans le contexte des hôpitaux publics français est une exigence pressante. Des recherches antérieures ont établi un lien significatif entre la quantité de consommation de médicaments et ces efforts de réduction des coûts. Plusieurs facteurs influencent les schémas de consommation pour chaque médicament. Ces facteurs comprennent la population desservie par l'hôpital, le nombre de lits, les types de traitements dispensés par le personnel médical, etc. Dans Koala et al. (2022), douze facteurs principaux ont été sélectionnés. En collectant et en compilant des données provenant de différentes sources, une base de données a été créée, comprenant des informations sur 21 médicaments différents dans quatre hôpitaux. Cette base de données englobe un total de 21 caractéristiques. Deux approches principales ont été examinées et présentées. Une approche s'est concentrée sur l'analyse de séries chronologiques, tandis que l'autre adoptait une perspective causale. Pour l'approche des séries chronologiques, une période temporelle commune a été choisie, garantissant que des données valides se chevauchaient pour la plupart des médicaments et des hôpitaux. Dans la méthode causale, la période entière comportant des données valides (avec des valeurs non nulles et excluant les valeurs aberrantes) a été utilisée. Une fois l'analyse effectuée et les données invalides traitées, deux nouvelles voies de recherche causale ont émergé. La première approche impliquait une stratégie unifiée où les données de tous les hôpitaux étaient combinées pour générer des clusters en vue de la prédiction de la consommation. La deuxième approche consistait en la création de quatre ensembles de données distincts, correspondant chacun à un hôpital différent.

Diverses techniques de regroupement de données (K-means, mini-batch k-means, regroupement agglomératif et DBSCAN) ont été utilisées, et les clusters résultants ont été évalués à l'aide du modèle de régression Random Trees. Cette technique de prévision s'est avérée mieux adaptée pour résoudre le problème actuel par rapport aux autres méthodes explorées dans les recherches antérieures. Dans les deux approches, qu'il s'agisse de la méthode des séries chronologiques ou de la méthode causale, les techniques de regroupement n'ont pas donné des mesures de prévision satisfaisantes. Ce résultat suggère que la disponibilité limitée des données affecte la précision des résultats. Chaque médicament présente des attributs si distincts qu'il possède ses propres caractéristiques particulières. Cela souligne l'idée que d'autres techniques de regroupement devraient être explorées. Par exemple, la méthode Dynamic Time Warping (DTW) pourrait être envisagée pour l'approche des séries chronologiques.

En tant que stratégie de conclusion, une évaluation globale a été réalisée en testant manuellement des combinaisons d'hôpitaux et de médicaments. Cette approche a généré un total de 2310 associations potentielles (par exemple, Médicament A de l'Hôpital X avec Médicament B de l'Hôpital Y). Pour chaque paire, des tendances et des quantités ont été tracées, permettant une analyse visuelle qui a facilité l'identification des paires bien adaptées. Parmi ces combinaisons, 206 paires ont démontré une forte adéquation. La technique de prévision a ensuite été utilisée, ce qui a entraîné des améliorations des métriques, en particulier en ce qui concerne l'erreur moyenne absolue en pourcentage (MAPE) pour certaines paires. De plus, les clusters formés par l'intersection des médicaments et des hôpitaux ont été soumis à des tests, et les prédictions de cette analyse ont également donné des valeurs favorables de la métrique MAPE.

1 Introduction

Hospital supplies and products, such as medicines, account for a significant portion of expenditure (Koala et al. (2022), OECD (2013)). Given the limited budget of healthcare facilities, cost reduction in process and logistics can have a profound impact on the lifetime and quality of service provided to the community. Various factors, including socio-demographic, socioeconomic, health-related, facility-related, and staff-related aspects, influence consumption dynamics and can be used to forecast demand Koala et al. (2021). Each country and macro region has its own cultural and specific factors that affect consumption dynamics. Taking this into consideration, French hospitals were chosen for this research due to their continuity with previous studies and data availability Koala et al. (2022).

The healthcare sector in France is constantly evolving in a complex and dynamic environment. As hospital centers see their budgets shrinking over the years, they must optimize their resource management while ensuring the safety and quality of care for the patients they serve. Pharmaceutical logistics, which represents a significant portion of hospitals' budget, is an important lever for optimization. This optimization relies on a better understanding of the hospital's medication demand.

Considering the context of the french hospitals, it is important to assume that the medicine's consumption is more related to the size of the population served by the health unity, the kind of treatments offered and

Previous research in this field, guided by the PhD researcher D. Koala, Prof. Dr. G. Alpan, and Prof. Dr. Z. Yahouni, in collaboration with a private company in the sector, established the foundational databases upon which the current study builds its hypotheses and subsequent solutions. In the initial publication, Koala et al. (2021), an exploration of the diverse factors influencing medicine consumption in French hospitals was conducted through qualitative research and analysis. Subsequently, a correlation analysis was undertaken to investigate principal quantitative relationships. In Koala et al. (2023), various machine learning techniques employed for consumption prediction were introduced. Additionally, Lim et al. (2023) delved into time series methodology, and finally, a Markov Chain-based predictive approach was studied in Vélez et al. (2022).

Among the various factors impacting consumption, a noticeable proportion is attributed to facility infrastructure and staff members, in contrast to the remaining categories Koala et al. (2021). Subsequent investigations entailed quantifying correlations between distinct factors and the volume of medicine consumption. These analyses revealed the preeminent significance of the third category while also underscoring discrepancies in data accountability across different facilities Koala et al. (2022).

In summary, the analysis conducted in previous works led to the selection of twelve main factors to form the consumption forecasting database. These factors include the population size served by the hospital, the number of medical professionals for consultations and hospitalization, the total number of beds, the number of medical visits or total number of patients, the geographical location of the facility, the number of health facilities that comprise the hospital, seasonal factors, the number of departments or medical facilities, the total number of non-medical staff, medical specialists, and the type of medical department Koala et al. (2022).

In this study, our primary aim is to explore whether it's possible to identify similarities among various medications and enhance the predictive capabilities of existing models. The ultimate goal is to reduce computational resources required and minimize the margin of error between predicted and actual consumption. We encounter several challenges in this endeavor. Firstly, there's a notable scarcity of valuable data for conventional clustering algorithms, which affects our ability to effectively group medications. Secondly, we face the issue of unbalanced datasets. Upon analysis, it becomes evident that

different hospitals have contributed varying amounts of data, often during different time periods. These gaps in data inputs pose a significant challenge, particularly in a time series-based approach, and can hinder a comprehensive analysis of the problem when sufficient valid data is lacking.

Indeed, the factors discussed earlier play a significant role in medication consumption patterns. However, it's crucial to ascertain whether these factors alone are sufficient to comprehensively characterize usage patterns across various time periods and regions. Furthermore, it's worth noting that the techniques employed to evaluate clustering in both the causal and time series methods are well-suited for the unique challenges posed by the domain of medications.

Certainly, having well-defined objectives and a structured approach is crucial for conducting successful research and analysis, especially in a multidisciplinary field like this. Leveraging a combination of academic and industry best practices, as well as exploring cutting-edge machine learning techniques, can provide a comprehensive framework for achieving the goals of this internship. This approach allows for flexibility and adaptability, which is essential when dealing with complex data and evolving challenges in the healthcare domain.

The present work is organized in the following way : In **Section 2 - Context**, a concise overview of the problem to be addressed during the internship period is provided, along with an introduction to the relevant stakeholders and the laboratory. **Section 3 - Materials** delves into the dataset used, various clustering and forecasting techniques explored, the selected methodology, and the features either employed or newly created. **Section 4 - Results** entails a discussion of the metrics obtained and the insights gleaned through the course of the work. Furthermore, the implications of these findings in the broader context of the project will be examined. **In Section 5 - Conclusion** a summary of the work conducted is presented, accompanied by a glimpse into potential avenues for future research and development.

2 Context

The internship was done in G-SCOP laboratory, from lab's web site : G-SCOP is a multidisciplinary laboratory which has been created to meet the scientific challenges imposed by the ongoing changes within the industrial world. The scope of the laboratory goes from the products conception to the production systems management and is based on strong skills in optimisation.

This project primarily focuses on employing machine learning techniques to forecast medicine demand. Prior research within this domain sought to identify suitable machine learning models tailored to individual medications. These investigations led to the discovery that the Random Forest Regressors model outperforms others for this specific challenge (Koala et al. (2023)).

In the current internship, the core objective is to analyze patterns in medicine consumption behavior and identify similarities. This analysis is facilitated through the utilization of clustering techniques. The overarching aim is to determine if specific medications exhibit comparable consumption behaviors. If such commonalities are discerned, it becomes advantageous to propose machine learning algorithms for predicting the consumption of groups of medications rather than considering each medication individually.

Approach :

- The first step of this internship involves reviewing the previous work conducted by the doctoral student. These studies will provide essential context and data understanding for the current study.
- The second step includes conducting a brief literature review on clustering methods to identify the most suitable approach for the context.

- In the third step, we will develop and test clustering methods on the dataset.
- The fourth step entails applying prediction methods to the groups obtained from the previous clustering step.
- Finally, the last step involves writing a scientific report that explains the approach and presents the results of the previous steps.

In conclusion, this internship merges academic research with practical application, addressing a pivotal aspect of medicine demand prediction. The collaboration with the G-SCOP laboratory underscores the commitment to bridging the gap between theory and real-world challenges. By analyzing consumption patterns, identifying similarities, and optimizing prediction techniques, this project contributes to the broader landscape of healthcare optimization. Furthermore, the systematic approach outlined in the methodology ensures transparency and reproducibility, crucial components of impactful research.

3 Materials

With the information gathered and data obtained from various sources, a comprehensive database was constructed to enhance our understanding of the contributing factors. A data pipeline was conceptualized and integrated in the study by Koala et al. (2022), a framework that has been embraced in the current research as well, illustrated in Figure 1.

FIGURE 1 : Data pipeline proposed by Koala et al. (2022)

3.1 The data set

The database amalgamated both internal hospital data and external information accessible from public sources. The resulting dataset, which was generated through the data pipeline, encompassed a total of twenty-one features. These features are detailed in Table 1.

Id	Feature name	Description	Kind of variable
01	ID_REF	Drug ID per hospital	Categorical
02	ID_SITE_RATTACHE	Hospital ID	Categorical
03	CODE_ATC	Anatomical Therapeutic Chemical	Categorical
04	HOSPI_CODE_UCD	Drug UCD code	Categorical
05	DATE_MOUV	Date of data input	Date
06	N_UFS	Number of hospital medical units	Numeric
07	QUANTITY	Quantity of consumed drugs	Numeric
08	WEEK	Week of consumption	Categorical
09	MONTH	Month of consumption	Categorical
10	YEAR	Year of consumption	Categorical
11	N_ETB	Number of hospital facilities	Numeric
12	POPULATION	Population in the district	Numeric
13	P_MEDICAL	Number of physician	Numeric
14	PN_MEDICAL	Number of other staff	Numeric
15	LIT_HC	Number of beds for full hospitalization	Numeric
16	LIT_HP	Number of beds for partial hospitalization	Numeric
17	SEJ_MCO	Number of visits in MCO departments	Numeric
18	SEJ_HAD	Number of visits in HAD departments	Numeric
19	SEJ_PSY	Number of visits in PSY departments	Numeric
20	SEJ_SSR	Number of visits in SSR departments	Numeric
21	SEJ_SLD	Number of visits in SLD departments	Numeric

TABLE 1 : Generated dataset by the pipeline with variable types

A sample of the dataset is presented in Fig.2.

ID_REF	ID_SITE_RATTACHE	CODE_ATC	HOSPI_CODE_UCD	DATE_MOUV	N_UFS	QUANTITY	WEEK	MONTH	YEAR	N_ETB	POPULATION	P_MEDICAL	PN_MEDICAL	LIT_HC	LIT_HP	SEJ_MCO	SEJ_HAD	SEJ_PSY	SEJ_SSR	SEJ_SLD
502829	HOSPI_3	C01CA03	3400892508566	2014-11-07	3	210.0	45.0	11	2014.0	50	1107398.0	1158	7129	2063.0	521.0	117781	594	2903	1302	97
502829	HOSPI_3	C01CA03	3400892508566	2015-01-22	4	340.0	4.0	1	2015.0	50	1120190.0	1239	7161	2053.0	493.0	118924	650	2878	1334	75
501463	HOSPI_3	J01CR05	3400893022634	2018-02-14	2	200.0	7.0	2	2018.0	50	1159220.0	1322	7439	2008.0	517.0	115376	1093	2481	1183	118
9220364	HOSPI_2	H02AB06	3400892203645	2017-06-30	4	55.0	26.0	6	2017.0	5	539067.0	714	5001	1157.0	187.0	75420	0	1236	261	0
9490	HOSPI_4	B01AC06	3400892065366	2018-03-26	1	630.0	13.0	3	2018.0	39	1859524.0	2627	15723	4477.0	486.0	255490	0	837	8416	209
890900	HOSPI_1	M03BX01	3400892697789	2015-10-06	1	20.0	41.0	10	2015.0	12	571879.0	684	5295	1411.0	94.0	74102	0	0	1140	57
9272958	HOSPI_2	N02AX02	3400892729589	2019-12-30	5	150.0	1.0	12	2019.0	5	542302.0	706	5013	1141.0	141.0	76593	0	1007	206	0
9387549	HOSPI_2	N02BE01	3400893875490	2018-05-21	2	30.0	21.0	5	2018.0	5	541454.0	703	5007	1159.0	139.0	74663	0	1193	237	0
503386	HOSPI_3	N02BE01	3400893875490	2015-03-18	18	410.0	12.0	3	2015.0	50	1120190.0	1239	7161	2053.0	493.0	118924	650	2878	1334	75
503129	HOSPI_3	N02BE01	3400891996128	2015-08-05	35	4350.0	32.0	8	2015.0	50	1120190.0	1239	7161	2053.0	493.0	118924	650	2878	1334	75

FIGURE 2 : Dataset sample

The dataset comprises various features related to hospital supplies and medication consumption. Each row corresponds to a specific daily entry, recording diverse factors. Key features include drug IDs per hospital, hospital IDs, anatomical therapeutic chemical codes, drug UCD codes, and dates of data input.

The UCD code, or "Unité Commune de Dispensation" code, is a standardized identification system commonly employed within the healthcare sector, particularly in French hospitals, for the meticulous traceability and categorization of pharmaceutical products. Developed and overseen by the National Agency for the Safety of Drugs and Health Products, this code is instrumental in uniquely identifying drugs. It consists of a series of alphanumeric characters, which may encode information about the drug, such as its name, dosage, packaging, and other relevant details. The implementation of the UCD code plays a pivotal role in ensuring the accurate dispensation, monitoring, and management of pharmaceuticals, thereby contributing to enhanced patient safety, effective inventory control, and regulatory compliance within healthcare facilities Hada* (2007).

3.2 Features

Moreover, the dataset contains numeric features such as the number of hospital medical units, quantity of consumed drugs, number of hospital facilities, population in the district, number of physicians, and number of other staff members. Categorical features consist of weeks, months, and years of consumption, along with the number of visits in different departments (MCO, HAD, PSY, SSR, SLD).

The dataset appears comprehensive, encompassing vital information related to hospital supplies and medication consumption. It enables the possibility of conducting analyses and applying machine learning techniques for demand forecasting and consumption behavior studies in healthcare facilities.

As demonstrated in the preceding sections, there is valuable information about the context of health facilities in France. As an initial step towards demand forecasting, it is evident that the most robust scenario would involve employing one prediction model per medicine. However, as the number of used drugs increases, so does the computational demand.

3.3 Quantity of data per hospital

The initial dataset comprised $75,692$ data inputs. After eliminating duplicates and addressing missing values, a more refined dataset was obtained, totaling $75,684$ data inputs. Under the division approach, datasets for each hospital contained varying amounts of data :

- **Hospital 1** : $22,725$ data inputs
- **Hospital 2** : $15,439$ data inputs
- **Hospital 3** : $27,591$ data inputs
- **Hospital 4** : $9,929$ data inputs

Each data entry in the dataset corresponds to a day when the hospital unit's operator manually input the data. It's important to note that each hospital may have its unique policies regarding data entry and handling. This variability in data entry practices cannot be precisely determined by this study. For instance, one hospital unit's staff may input data every Friday, while another might do so at the end of each day. To align with the objectives of this study and mitigate the influence of these varying policies across different hospital units, the data is analyzed on a monthly basis, as elaborated upon in subsequent sections. This monthly aggregation clearly highlights a decline in data inputs for each hospital, and in some instances, specific medicines. This reduction in data inputs can be attributed to the disparities in data entry policies and the inherent gaps within the dataset.

3.4 The extended methodology

In this study, an extended methodology is proposed, building upon the one presented in Figure 1. The concept of constructing a data pipeline is fundamental to align with industry best practices, ensuring that this study not only comprehensively investigates the problem at hand but also produces a viable and intelligible solution for future exploration and incorporation into concurrent research efforts. While the complete trajectory may not have been apparent at the outset, the project evolved organically as the problem was subjected to more in-depth examination and potential solutions from both academic literature and open-source endeavors came to light. This iterative process delineated a clear path for the data to traverse, culminating in the final phase of the project : the validation of techniques through

consumption forecasting. The objective is to gain a deeper understanding of the available data and to build upon previous research, as illustrated in Figure 3.

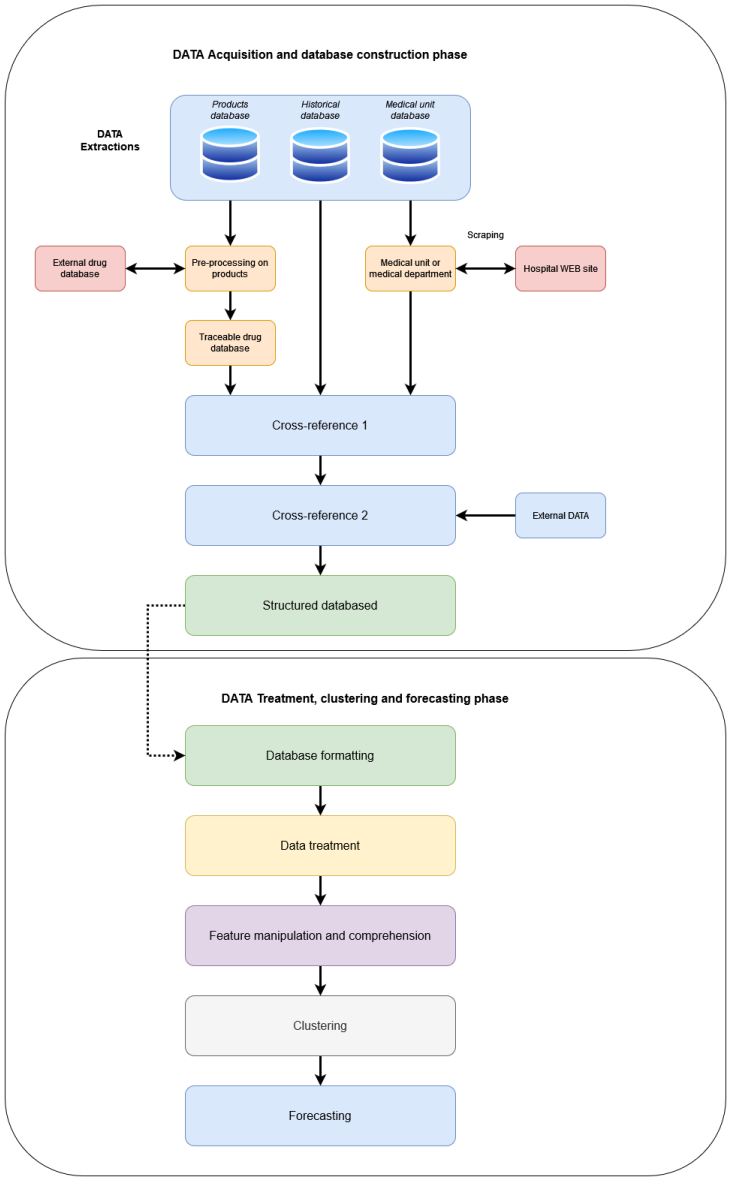

FIGURE 3 : Extended pipeline to implement clustering and forecasting

3.4.1 Data treatment

In the first step of the extended part, **Data treatment, clustering, and forecasting phase**, an exploratory analysis of the data is proposed, where the database is studied and formatted to facilitate subsequent steps. This phase also involves removing missing and incoherent values (**Database formatting**).

During the **Data treatment** step, outliers in the ***"QUANTITY"*** feature are removed, considering the unique characteristics of each hospital and medicine. Two main approaches are considered : a causal method and a time series approach. For the time series approach, a common consumption period spanning from **March 1, 2017**, to **March 1, 2019**, is established. Additionally, the data is aggregated by specific hospital, drug reference, and month and year of consumption.

To handle categorical values, a one-hot encoding is applied to the features : "**ID_SITE_RATTACHE**",

"**MONTH**" and "**YEAR**". Considering the initial ambiguity, it becomes evident that capturing the seasonal patterns from each month or year is essential. To achieve this, the technique of one-hot encoding is employed. It's worth noting that the majority of prediction algorithms exclusively accept numeric features. Consequently, this step holds significant importance in the data preprocessing pipeline.

3.4.2 Feature manipulation and comprehension

At the **Feature manipulation and comprehension** step, the proposed approach involves conducting a correlation analysis, standardizing the data, and applying a dimensionality reduction technique.

In the time series approach, a seasonal decomposition method is proposed to facilitate the comprehension of recurring patterns within temporal data. This method dissects the data into distinct components : a trend component, a seasonal component, and a residual component. It's worth noting that two models can be used for this decomposition, namely the multiplicative model and the additive model. In the context of this specific problem, characterized by its unique traits, the additive model was selected. However, it's important to acknowledge that the data gaps mentioned in previous sections pose a risk to this method, as they can hinder or even render the calculation of temporal information infeasible. To address this challenge, gap-filling techniques could be explored in future research.

Furthermore, a moving average was computed as part of the feature engineering process. This calculation aimed to enhance the dataset by providing additional insights into the available data and how consumption patterns might be approximated by identifying similarities.

3.4.3 Clustering

In the **Clustering** part, various compositions of the K-means, mini-batch K-means and agglomerative clustering algorithms are proposed, tested, and evaluated using the metric of silhouette score.

Subsequently, a manual clustering approach was introduced as a deliberate attempt to identify pairs of medicines across different hospitals or distinct medicines within the same hospital, exhibiting analogous consumption patterns. This approach aimed to tackle the data scarcity observed during the data analysis phase. By establishing connections among medicines, this method aimed to extract valuable insights that might help bridge data gaps and enhance our overall comprehension of consumption dynamics.

One expected outcome from the clustering techniques employed in this study is the possibility of assigning multiple medicines to different clusters, consequently requiring different forecasting models for the same medicine. While this outcome is not ideal, it is understandable given that each medicine can exhibit varying consumption patterns. A potential strategy to mitigate this issue involves restructuring the data set to have only 21 rows, one for each medicine, thus limiting the maximum number of models per hospital to 21. However, due to time constraints, this approach could not be implemented in the current study and is recommended for future research.

The concept of clustering in this study is driven by the practical scenario in hospitals where patients often receive multiple medicines as part of their treatment. The hypothesis to be tested is whether there exists a discernible trend in the consumption of medicines that can be linked to factors studied in previous research. The objective is to develop predictive models for the consumption of different medicines using shared information that may influence the final consumption figures. Additionally, the study seeks to identify potential similarities in the consumption patterns of the same medicines across different hospitals. Clustering offers a way to group medicines based on these shared characteristics, which could ultimately enhance the forecasting process and provide valuable insights into medication consumption behavior.

3.4.4 Forecasting

During this decisive stage of the project, known as the **Forecasting phase**, the data undergoes a process of segmentation based on the outcomes of the preceding clustering phase. This segmentation serves as the foundation for constructing and assessing predictive models within individual clusters. The rigorous evaluation procedure entails the strategic utilization of cross-validation alongside meticulous grid search techniques aimed at optimizing model parameters. An integral element of this method involves the establishment of baseline models for each distinct approach. This includes dedicated baselines for the time series approach, as well as for both the unified and hospital-specific divisions within the causal method. This meticulous approach underscores our commitment to unraveling the underlying dynamics of medicine consumption while aligning with the overarching goal of efficient forecasting.

The evaluation of our forecasting models hinges on several key metrics : Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R2). These metrics collectively provide a comprehensive assessment of our models' predictive capabilities.

MAPE measures the average percentage difference between predicted and actual values, thereby quantifying the overall accuracy of our predictions in terms of percentage error. MAE computes the average magnitude of the prediction errors, shedding light on the absolute deviation between forecasts and actual outcomes. RMSE quantifies the square root of the average squared differences between predictions and actual values, accentuating the importance of larger errors. R2, also known as the coefficient of determination, gauges the proportion of the variance in the dependent variable that our model explains, elucidating how well the model fits the observed data.

Of these metrics, MAPE holds particular prominence, given its ability to showcase the relative error in terms of percentage. This metric's emphasis on proportionate error makes it a valuable indicator of our models' performance.

The evaluation of model performance in the present research encounters certain constraints and challenges. As we move towards a unified approach, where the entire dataset from all four hospitals is analyzed together, the metrics tend to improve compared to when we divide the dataset per hospital or cluster. However, this unified approach also results in a limited amount of data available for training and testing. The issue of data scarcity has been discussed in previous works and within the earlier sections of this research.

Another challenge pertains to the interpretation of metrics within a cluster. Consider a scenario where a cluster contains 10 different medicines. These medicines may have varying data compositions in terms of characteristics that could be deeply analyzed by the model or remain hidden within a black-box approach, known only to the model itself. Within this cluster, some medicines might have only a few data points, such as three, while others may have more extensive data. Metrics for medicines with limited data exposure may not perform well, as there are not enough data points for accurate evaluation. For instance, the R2 metric requires at least two points for calculation. This issue can potentially be addressed through the acquisition of more data or by employing different data distribution techniques within each cluster.

Another critical aspect to consider in the analysis is the composition of the prediction model. Given the constraints of computational resources, there are limitations on the range of model parameters that can be tested. It is not feasible, both in terms of time and resources, to explore an exhaustive set of parameter combinations. Therefore, practical constraints necessitate setting boundaries on the range of parameters that can be evaluated and measured. These limitations are crucial to keep in mind when interpreting the results and understanding the model's performance in the context of the present research.

In summary, each line of research pursued in this work should adhere to a structured pipeline with

the ultimate goal of predicting medicine consumption. This aligns with the overarching problem of reducing hospital costs by optimizing medicine inventory management. This structured approach provides a clear path to addressing the research objectives and contributing to the broader healthcare management challenge at hand.

A more detailed diagram for each step is presented in Fig.4.

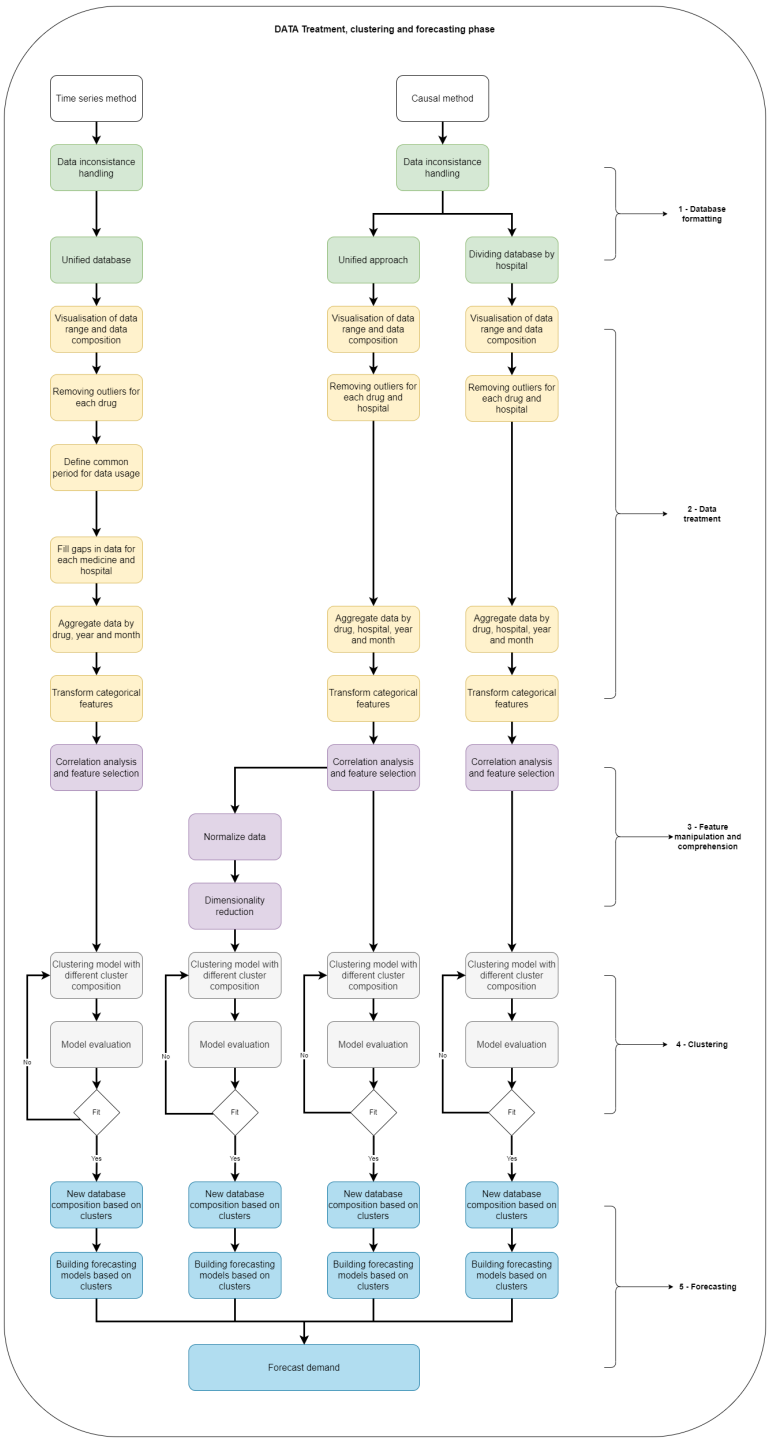

FIGURE 4 : Detailed extended pipeline to implement clustering and forecasting

Each segment of the diagram will be comprehensively elucidated in the subsequent sections, offering an in-depth exploration of the methodology and strategy applied at each stage. Furthermore, a hyperlink to the developed work, housed within a dedicated GitHub repository, will be furnished at the conclusion of each subsection. This link will grant access to the underlying codebase and supplementary resources

associated with the corresponding facet of the project. Such an approach ensures transparency and ease of access, facilitating future reference and fostering collaborative engagement.

4 Results and discussion

4.1 Database formatting

The initial stage involved addressing missing data. Given the low occurrence of lines with empty spaces, the chosen approach was to remove the rows containing invalid values. Consequently, merely eight lines were excluded, accounting for a mere 0.01% of the entire dataset.

To enhance organization and clarity, a structured approach led to the creation of five distinct datasets. The initial dataset involves a unified compilation of data from all four hospitals, consolidated into a single dataframe. The subsequent four datasets, on the other hand, are partitioned according to individual hospitals, containing data unique to each respective medical institution.

This division allows for a more focused analysis of each hospital's data while also providing the opportunity to explore the unified dataset to identify common patterns and trends across all hospitals. By separating the data in this manner, the analysis and modeling process can be conducted more efficiently, leading to better insights and forecasting results.

This discrepancy in the data volume for each hospital might have implications for the subsequent analysis and modeling process. It is important to take this variation into account while developing the forecasting models and interpreting the results for each individual hospital.

Notebook(s) in : Git hub - 1_Database_Formatting.

4.2 Data treatment

In this step, both the unified and division approaches share common tasks, such as data aggregation by hospital and drug, as well as handling outliers.

In the initial step depicted in Figure 4, the focus is on data visualization and understanding the dataset's composition. This analysis is performed individually for each hospital. For example, the distributions of hospital one's data are illustrated in Figure 5. The vertical axis represents the density or concentration of occurrences of values, while the numerical values are displayed on the horizontal axis. This visualization aids in gaining insights into the data's distribution and characteristics for further analysis and decision-making.

It is evident from the "QUANTITY" plot that there are higher values on the right side, with a low density. This observation suggests the presence of a potential outlier. It is important to note that this does not necessarily imply an error on the part of the personnel responsible for inputting the data. However, these values do not align with the expected dynamics of certain medications. In order to gain a better understanding of this situation, a technique was implemented and applied separately for each hospital and medication. The objective was to preserve the patterns of individual consumption.

Another interesting analysis pertains to the variables "N_ETB," "SEJ_HAD," and "SEJ_PSY." It is worth noting that these variables maintain constant values. Consequently, it may be appropriate to exclude these constant numerical features as they may not contribute significantly to the clustering method or the forecasting step.

The plots of the other hospitals can be found in appendix A.

The percentile technique was employed to address outliers. This method involves calculating the

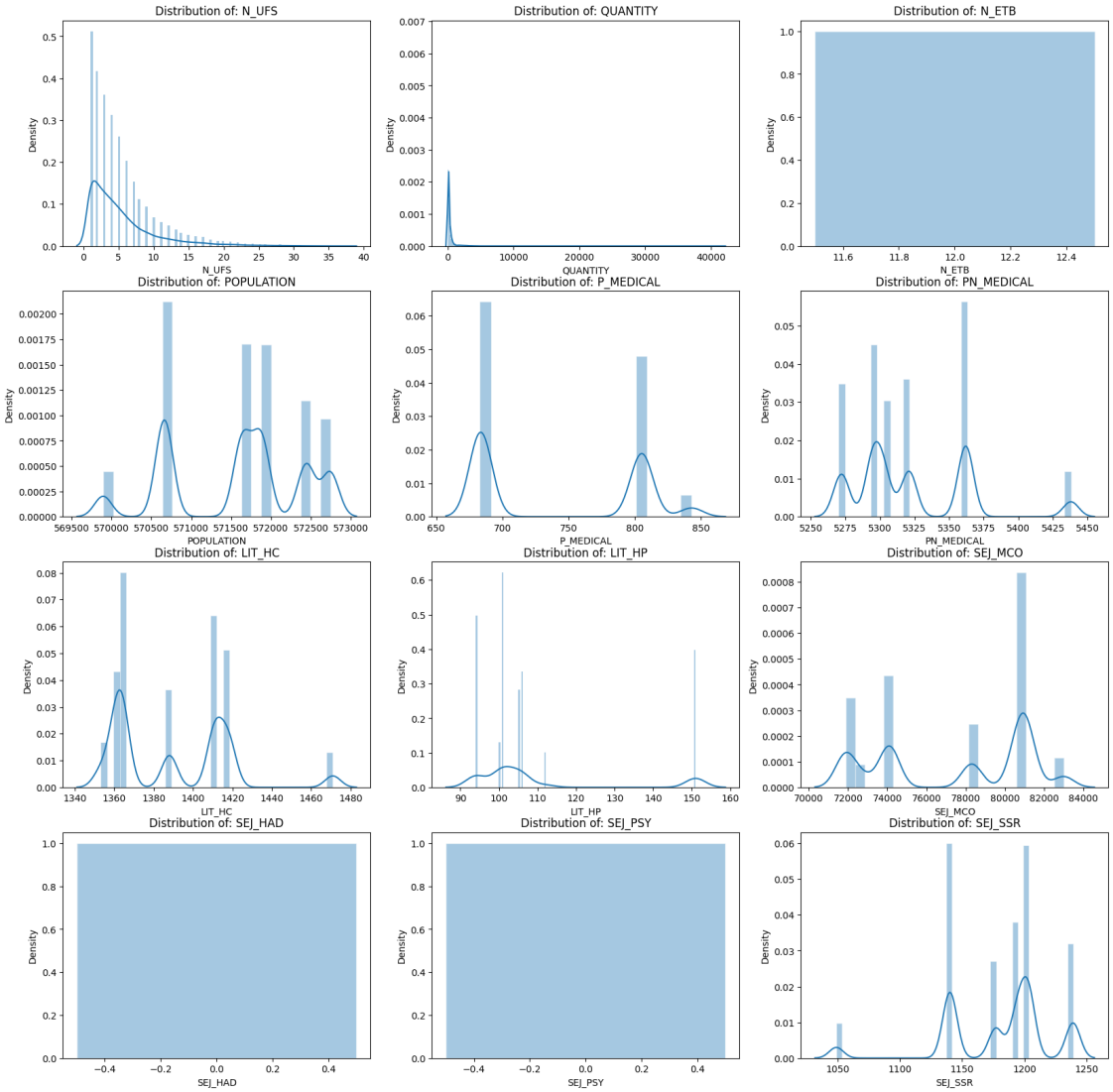

FIGURE 5 : Numerical variables distribution before removing the outliers - Hospital 1

percentiles of the variable "QUANTITY" within a specific data subset. Subsequently, cutoff values are determined based on these percentiles. The interquartile range (IQR), defined as the difference between the 75th and 25th percentiles, is utilized to establish a range wherein the majority of data points are expected to fall. Outliers are identified as values below the lower cutoff ($Q_1 - 1.5 \times IQR$) or above the upper cutoff ($Q_3 + 1.5 \times IQR$). By removing these outliers, the dataset can be cleansed, enabling further analysis that focuses on the central distribution of the data. This technique offers a straightforward and objective approach to detecting extreme values that have the potential to significantly impact the statistical properties of the dataset.

After applying the technique, the distribution of the variables was plot once again and is shown in Figure 6.

The difference in the distribution of "QUANTITY" is substantial when compared to the values depicted in Figure 5. The variation of values is more pronounced, ranging from zero to one thousand, which corresponds to the highest density of data inputs.

The plots of the other hospitals can be found in appendix B.

Regarding the time series approach, after handling the outliers, a unified time period was established for all four hospitals. This choice was made to facilitate the analysis of different medications within the

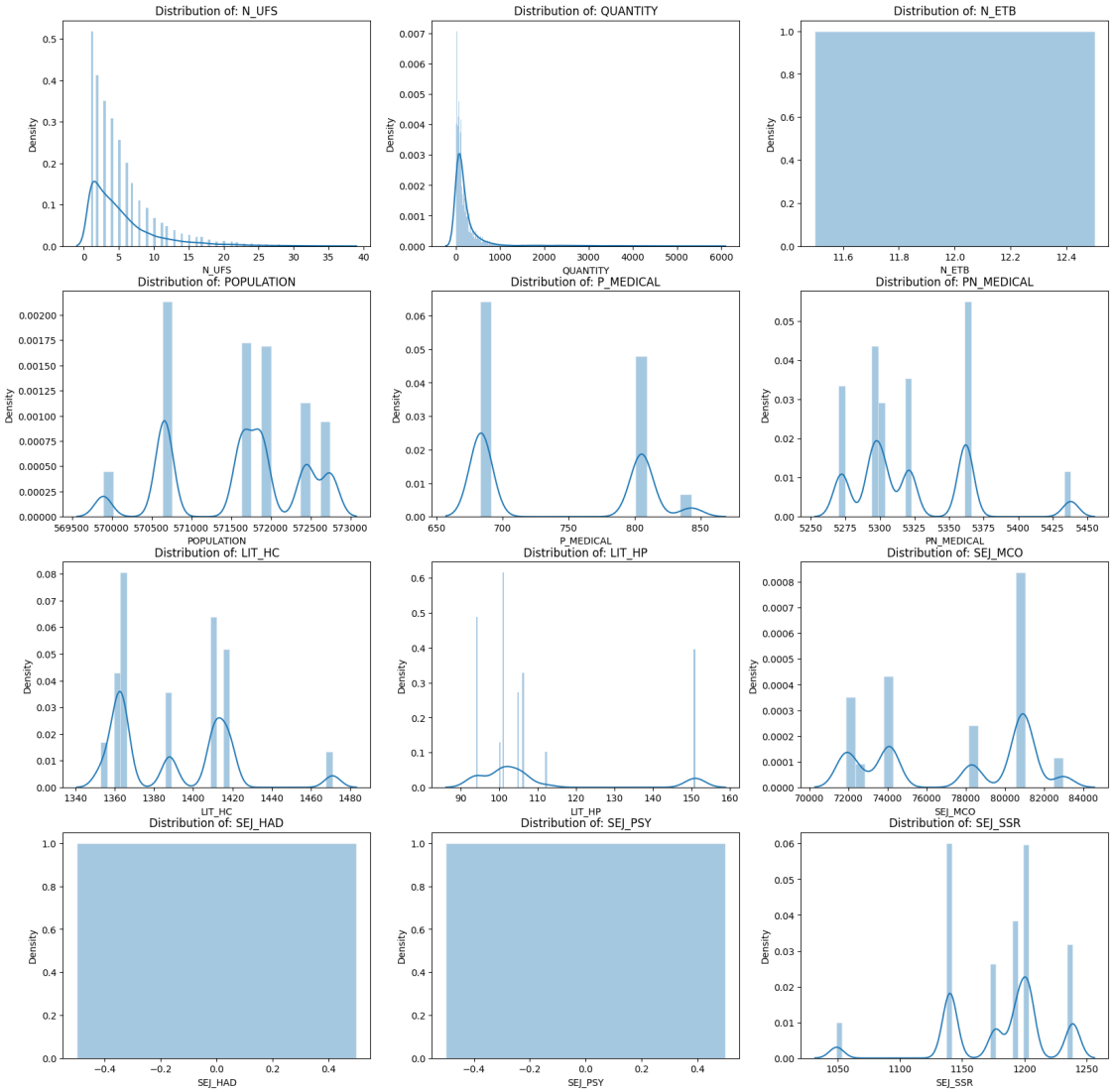

FIGURE 6 : Numerical variables distribution after removing the outliers - Hospital 1

same timeframe, reducing the influence of distinct temporal conditions on the clustering and forecasting processes. A visual representation of this unified time period is depicted in Figure 7. This ensures a consistent and synchronized analysis, enabling more accurate and reliable comparisons across different medications and hospitals.

As depicted in Figure 7, a temporal constraint arises with an endpoint in December 2019 attributed to the impact of *SARS-CoV-2*. However, the intersection of dates across all four hospitals commences in March 2017 (as evidenced by the first data input for hospital 2) and persists until March 2019 (coinciding with the final data input for hospital 4). Subsequent analysis unveiled a concentration of valid data between October 2017 and February 2019.

Following the establishment of the shared time period, it became evident that certain medications featured missing data inputs across various months. To bridge these gaps within the corresponding time series, a data filling strategy was implemented, leveraging forward fill and backward fill methods. Notably, a conscious choice was made to abstain from replacing missing values for the "QUANTITY" attribute.

In the context of this study, a missing data point may indicate no consumption of the medicine due to various reasons. By preserving these missing values, the analysis ensures data integrity and prevents potentially misleading interpretations of the consumption patterns for certain drugs. Subsequently, each

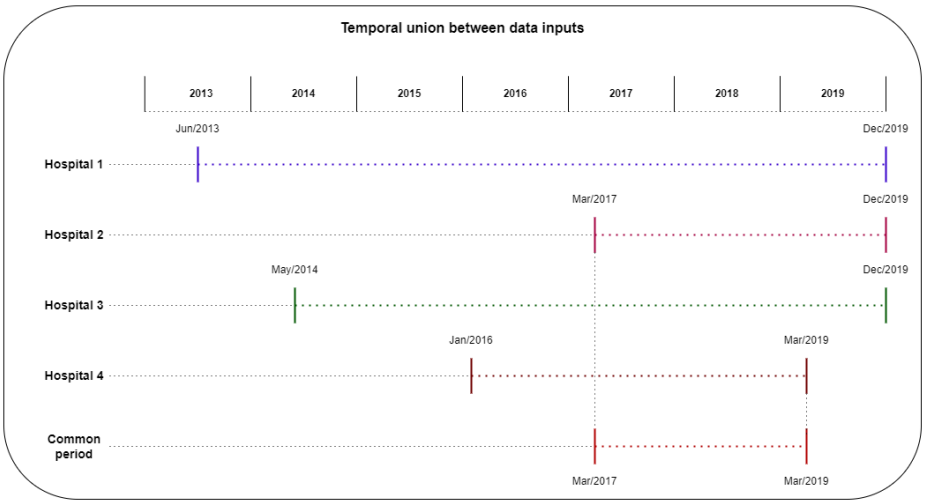

FIGURE 7 : Temporal union by a common period for the 4 hospitals

dataset's inputs were aggregated based on the UCD code, year, and month. In the unified approach, the aggregation also considered the hospital's information. This aggregation process facilitates further analysis and modeling, providing a more consolidated representation of the data for forecasting purposes.

A selection of figures illustrating the numerical data before and after the aggregation process, including the handling of outliers, is presented in Figure 8. These figures depict the values of "N_UFS," "HOSPI_CODE_UCD" : 3400893875490, "HOSPI_CODE_UCD" : 3400891996128, and "POPULATION". These specific features were chosen to exemplify the treatment applied to the numerical variables. In the original dataset, data inputs were recorded on a daily basis, while after aggregation, they were summed by month to reflect the consumption patterns.

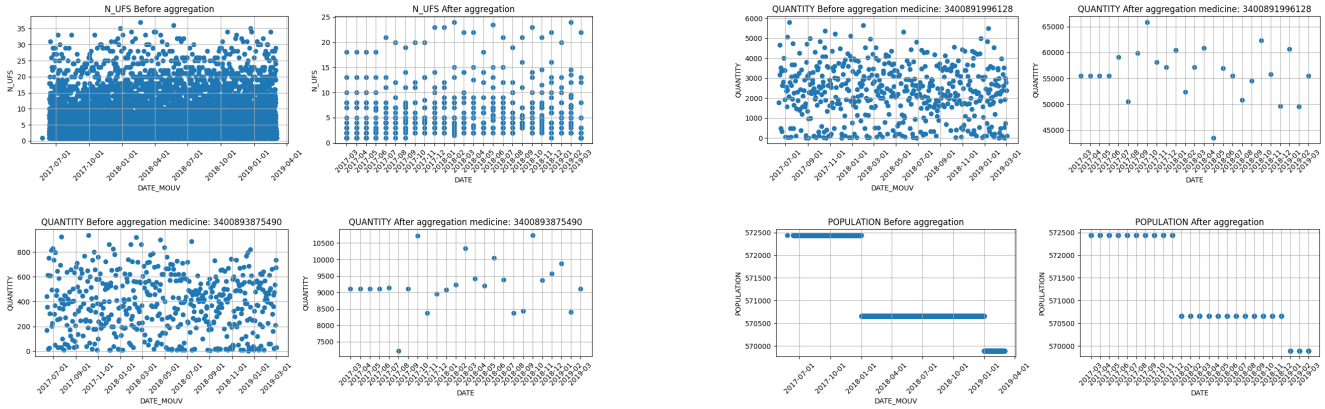

FIGURE 8 : Numerical features before and after aggregation and outliers handling

In the data plot for "N_UFS", it can be observed that there are 21 data inputs for each month. These inputs correspond to the 21 different medicines and indicate the number of medical units in which each medicine was consumed. In the remaining plots, it is evident that the data has been uniformly sampled over a one-month period. Additionally, regarding the consumption quantity, there is a noticeable change in the magnitude of values. This change is a result of aggregating the daily inputs into a single value representing the total consumption per month.

After the aggregation, the categorical features were transformed into numerical ones. The other figures can be seen in the shared repository.

To transform the information of the categorical features into numerical ones, the columns of "MON-

TH" and "ID_SITE_RATTACHE" were replaced with 16 other columns using one-hot encoding. One-hot encoding is a common technique used to convert categorical data into a numerical format, where each category is represented by a binary column. This transformation allows the categorical information to be incorporated into the analysis and modeling processes, as most machine learning algorithms work with numerical data. By applying one-hot encoding, the dataset becomes more suitable for further analysis and provides a comprehensive representation of the categorical variables in a numerical format.

Notebook(s) in :

Git hub - 2_Data_Treatment - Division approach.

Git hub - 2_Data_Treatment - Unified approach

Git hub - 2_Data_Treatment - Time series

4.3 Feature manipulation and comprehension

In the feature manipulation step, it was done some feature engineering in order to have more valuable information about the consumption patterns and also to be easier to identify in the clustering step the similarities between the different medicines or group of medicines. The moving average for different time periods was calculated and after testing the chosen period was of three months, an example with four different medicine is given in Figure 9.

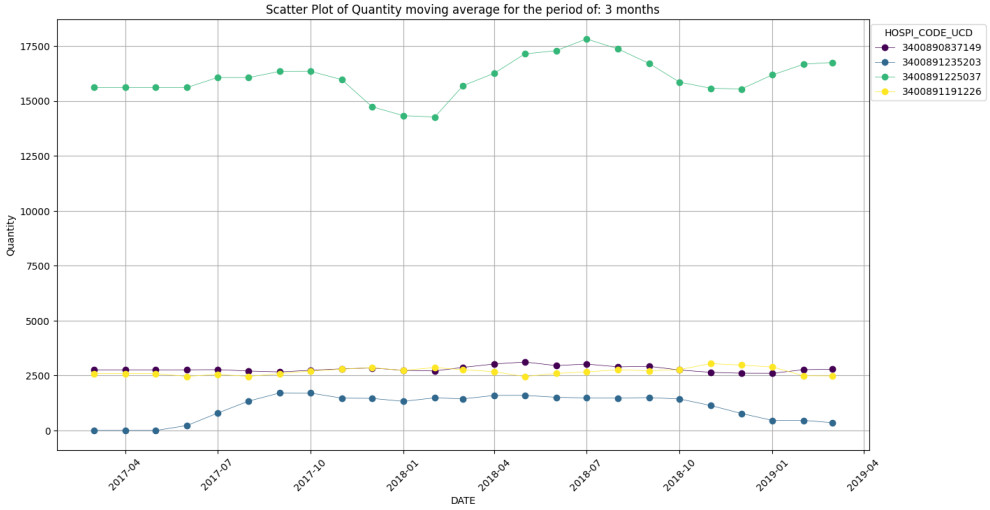

FIGURE 9 : Moving average calculated for four medicines in hospital 1

From the figure presented above, it is evident that the moving averages of medicines 3400890837149 and 3400891191226 exhibit a striking similarity. Additionally, the remaining two medicines also demonstrate comparable seasonal variations, with patterns starting in July 2017 and continuing until January 2018, followed by similar behavior between the period of 2018 and 2019. However, this information alone is insufficient to form definitive groupings. Therefore, to gain more insights and facilitate clustering, a correlation analysis was conducted, and clustering techniques were subsequently applied. These additional steps help identify more precise groupings and provide a comprehensive understanding of the consumption behavior of different medicines.

For the correlation analysis, Hospital 1 was selected as an example, and the other plots and results can be found in the shared repository. To conduct a meaningful correlation analysis, the features "N_ETB," "SEJ_HAD," and "SEJ_PSY" were excluded from the analysis as they remained constant for this hospital, offering no independent variation from the values of other features. The correlation plot for

Hospital 1 is presented in Figure 10, generated using the method described in The pandas development team (2020). This plot provides valuable insights into the relationships between different features and serves as a basis for the subsequent clustering techniques.

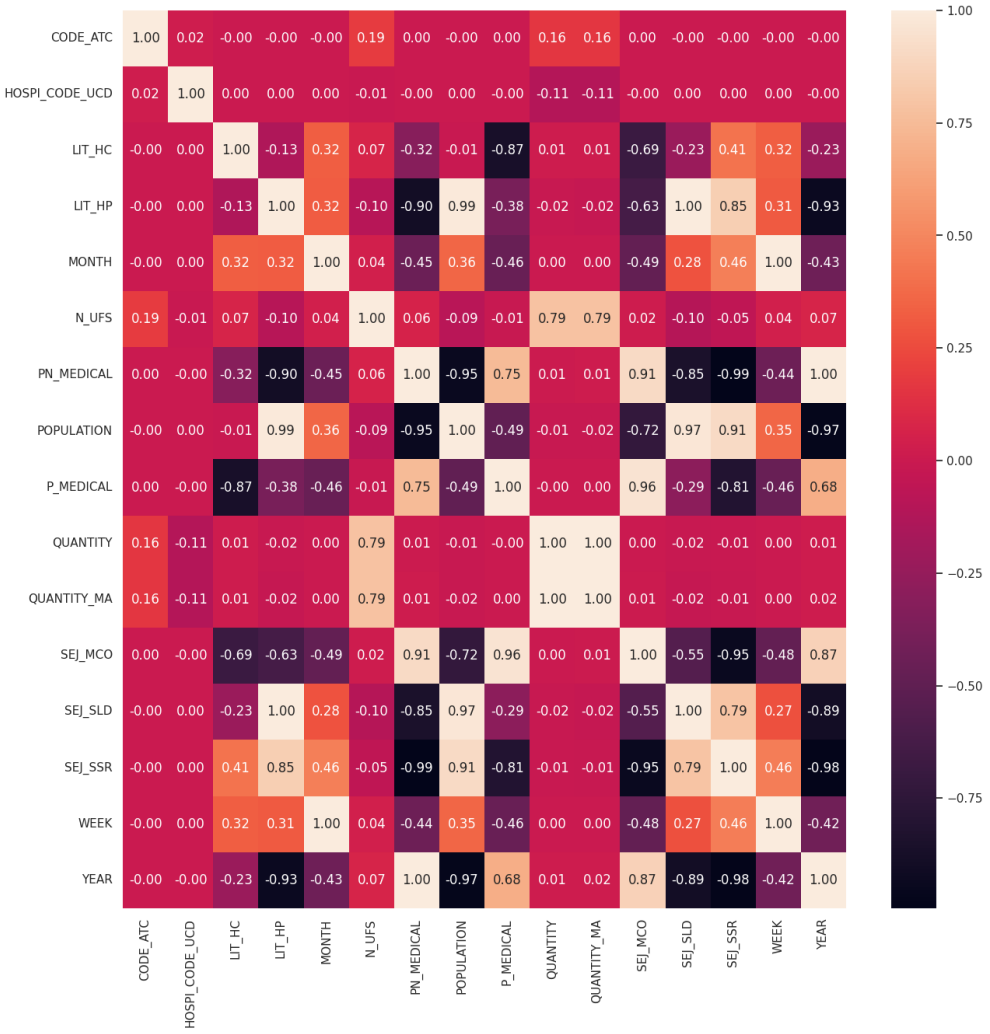

FIGURE 10 : Correlation analysis for the hospital 1 numerical features

As it is possible to see, there is a strong correlation between the features ”N_UFS” and ”QUANTITY”, the focus of this work, the second most expressive correlation is given by the ”CODE_ATC” feature followed by ”HOSPI_CODE_UCD”, from this, one can conclude that the consumption pattern has a strong link in the number of medical units that use the medicine, also, which medicine is being used and it’s chemical compound.

Notebook(s) in :

Git hub - 3_Feature_Manipulation_and_Comprehension - Division approach.

Git hub - 3_Feature_Manipulation_and_Comprehension - Unified approach.

Git hub - 3_Feature_Manipulation_and_Comprehension - Time series.

4.3.1 Time series seasonal decompose

In the pursuit of generating features that more accurately capture the consumption behavior of each medication, a tool from the *StatsModels* Python library (Seabold and Perktold (2010), McKinney et al. (2011)) was employed. This tool facilitated the examination of three fundamental components within the time series : seasonality, trend, and residuals.

Seasonality captures the time-dependent fluctuations that illustrate how specific time intervals influence the fluctuations in the variable of interest, in this context, "QUANTITY". Meanwhile, trend delineates the overarching direction in which variations unfold and their cumulative impact, augmenting the seasonal dynamics. In contrast, the residual component encompasses the unaccounted variation that cannot be attributed to either seasonal decomposition or trend. Employing the summing model, the decomposition methodology leverages the amalgamation of these three distinct variables—trend, seasonality, and residual—to depict the time series comprehensively Seabold and Perktold (2010).

In order to ilustrate what was explained, the decomposition of the medicine Doliprane of the fourth hospital, show in Figure 11.

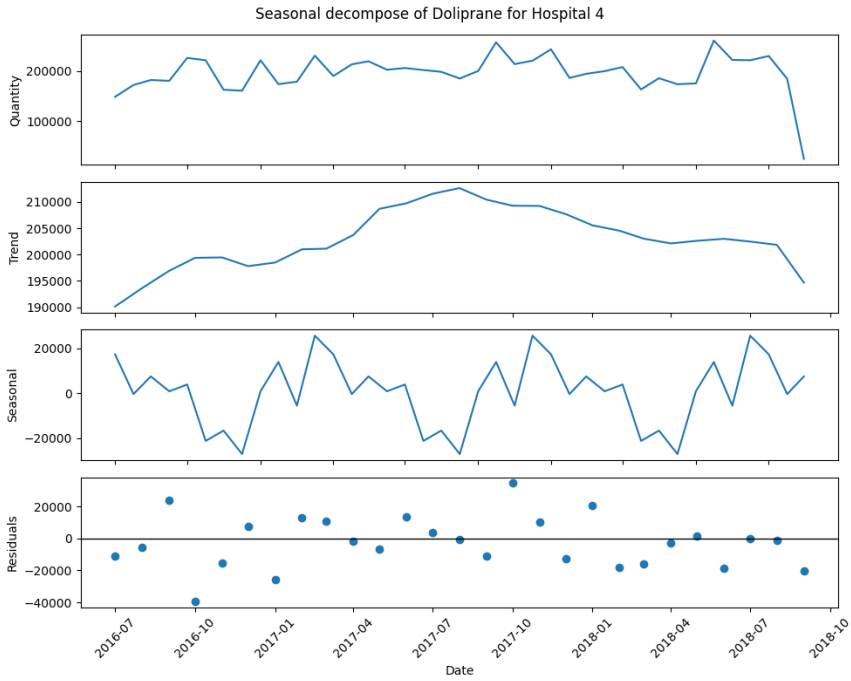

FIGURE 11 : Seasonal decomposition of Doliprane for the hospital 4

The initial plot in the figure depicts the "QUANTITY" variable's values. The subsequent plot illustrates the trend, revealing a discernible pattern of increasing medicine consumption over a certain period followed by a subsequent decrease. The seasonal plot showcases time-dependent variations, displaying approximately three recurring consumption peaks. As for the residual element, its magnitude is in the order of 10^4, aligning with the scale of the "QUANTITY" variable, which itself ranges around 10^5.

4.3.2 Dimensionality reduction

Principal Component Analysis (PCA) was chosen as the dimensionality reduction methodology for its simplicity and extensive application in the literature. PCA aims to find the linear and orthogonal projection of a high-dimensional dataset into a lower-dimensional subspace. It is a well-established technique widely used in various fields for data compression and feature extraction purposes (Murphy (2022)).

By reducing the number of features to 15 using PCA, the dataset's variability is preserved, providing a more concise representation while maintaining essential information. This enhanced dataset facilitates improved manageability and efficiency in subsequent clustering and forecasting steps.

The dataset underwent Principal Component Analysis (PCA) using the scikit-learn library to analyze the cumulative explained variance ratio based on the number of features. Out of the 30 generated features,

it was observed that 15 of them accounted for 95% of the dataset's variance. The results of this analysis are visually presented in Figure 12.

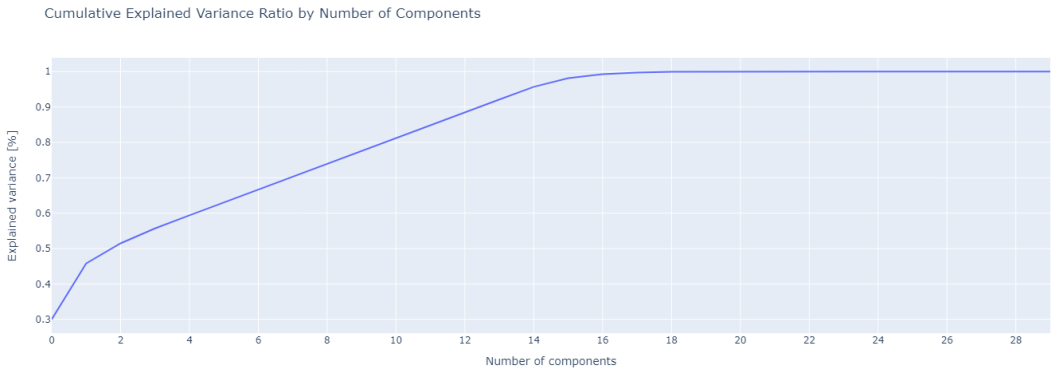

FIGURE 12 : Principal component analysis for the unified approach

Following the application of PCA, a new dataset was created with the reduced number of features, which will be used in the next section for implementing various clustering techniques.

4.4 Clustering

The clustering methodologies applied in this study encompassed the 'k-means', 'mini-batch k-means', 'agglomerative clustering', and 'DBSCAN' algorithms, all of which are part of the scikit-learn Python library Pedregosa et al. (2011). Different combinations of hyperparameters were explored to find appropriate clusters for the medicines. The number of clusters was varied within a range of 2 to 21, representing the total number of medicines in the dataset.

For the k-means algorithm, a number of iterations equal to 300 was chosen, considering the algorithm's complexity of $O(knT)$, where k is the number of clusters, n is the number of samples, and T is the number of iterations. By fine-tuning these hyperparameters, the clustering algorithms were optimized for the specific dataset and problem at hand.

Determining the optimal number of clusters is crucial as it affects the number of forecasting models that need to be developed. For instance, if a single model is used per medicine and per hospital, considering the 21 medicines in the dataset, it would require 84 distinct models. However, by utilizing the division approach with 4 clusters per hospital or the unified approach with just 4 clusters, the number of necessary models reduces to 16 and 4, respectively.

Reducing the number of models, provided that it yields satisfactory forecasting results, can significantly reduce the time required for demand prediction. Furthermore, it supports the hypothesis of this study that grouping similar medicines in the same cluster leads to more precise forecasting.

Each clustering technique possesses unique principles and characteristics that govern their performance and outcomes. In the subsequent subsections, more in-depth information will be provided regarding the specific configurations and approaches employed for each model. This detailed analysis will shed light on the effectiveness and suitability of each clustering technique for the given dataset and research objective.

Notebook(s) in :

Git hub - 4_Clustering - Division approach, Git hub - 4_Clustering - Unified approach, Git hub - 4_Clustering - Time series.

4.4.1 Metrics

The metrics used to evaluate the models were : Davies Bouldin score and Silhouette score from module metrics of scikit-learn. From the developper page we can define the metrics as :

- Davies Bouldin score : The score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Thus, clusters which are farther apart and less dispersed will result in a better score. The minimum score is zero, with lower values indicating better clustering Pedregosa et al. (2011).
- Silhouette score : The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is $\frac{(b-a)}{max(a,b)}$. To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of. Note that Silhouette Coefficient is only defined if number of labels is $2 \leq n_{labels} \leq n_{samples} - 1$ Pedregosa et al. (2011).

4.4.2 'K-means' and 'mini-batch K-means'

Utilizing the previously outlined parameters, the k-means and mini-batch k-means algorithms (a distinct variant of the conventional algorithm) were implemented. To determine the optimal number of clusters, an elbow plot was generated. This visual aid, depicted in Figure 13 for both the unified approach and the causal method, offers insights into the appropriate cluster count to be employed.

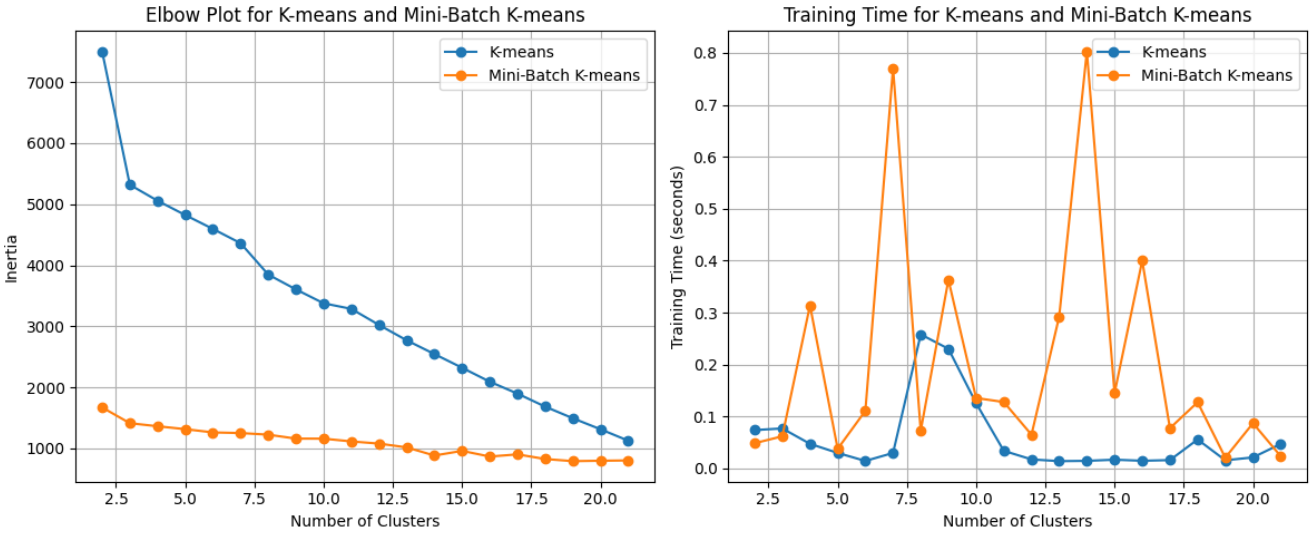

FIGURE 13 : Elbow plot for unified approach and causal method

Observing the left plot, it becomes evident that the inertia value decreases consistently from 3 clusters onwards. A similar trend is observed in the right plot, indicating that the fluctuation in training time is not substantially impactful for our specific application. Consequently, concerns regarding this aspect need not be a primary focus.

The observation that the turning point for better results occurs at 3 clusters, despite the results not being satisfactory, suggests that the clustering might be influenced more by the quantity of consumption than other factors. To better understand this phenomenon, further investigation into the clustering process and the features used for clustering is needed. It's essential to identify whether the current clustering

approach effectively captures the underlying patterns in the data or if adjustments to the clustering methodology or feature selection are required to achieve better results. This could involve exploring different clustering algorithms, feature engineering techniques, or even redefining the clustering criteria based on the specific characteristics of the data.

The metrics discussed before are presented in Figure 14

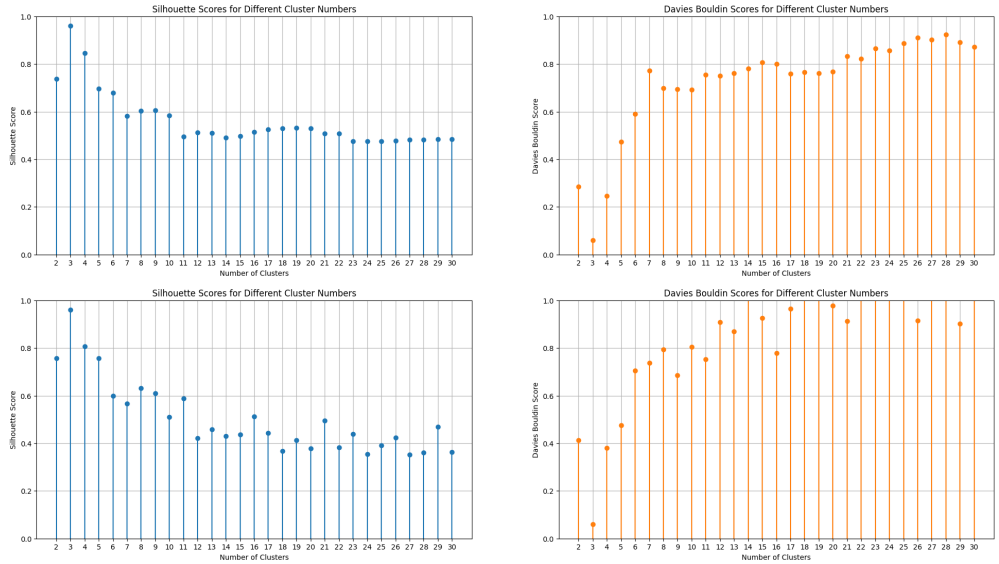

FIGURE 14 : Metrics of k-means (top) and mini-batch k-means (bottom) clustering

Both plots indicate a discernible trend in terms of the Silhouette score. Notably, an acceptable threshold emerges within the range of 10 to 11 clusters. However, interestingly, the peak value for this metric is consistently attained at the 3-cluster mark for both approaches.

4.4.3 'Agglomerative clustering'

Agglomerative clustering is a hierarchical technique used for partitioning data into groups based on similarity. It starts by treating each data point as an individual cluster and iteratively merges the closest clusters, progressively building a hierarchical structure. The process continues until all data points are part of a single cluster or a predefined number of clusters is achieved. Agglomerative clustering measures the distance between clusters using methods like single linkage, complete linkage, or average linkage, and allows for the visualization of data relationships through dendrogram representations. This approach is effective for revealing hierarchical patterns and relationships within complex datasets Muller and Guido (2018).

For the agglomerative clustering, there was a setting of possible combinations for the linkage and metric to be used.

- Linkage : Complete, average and single.
- Metric : L1, L2, cosine and Manhattan.

Comparing the different possibilities, it was the combination of metric Manhattan and linkage average that yields the best results for silhouette scores and number of clusters, Figure 15 illustrate the scores obtained.

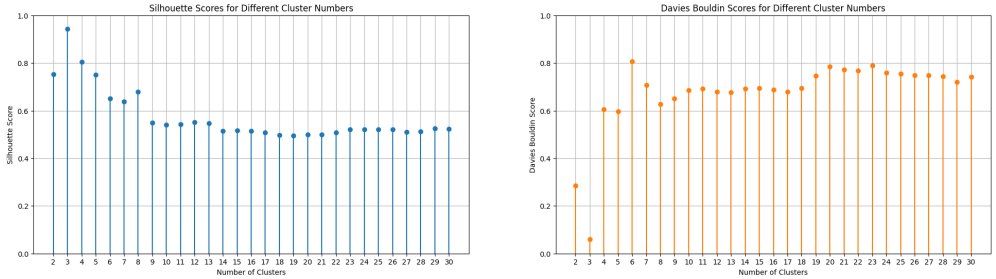

FIGURE 15 : Agglomerative clustering metrics in unified approach and causal method

Based on the observations from the left plot, it can be inferred that cluster counts of 4, 6, and 8 present as favorable candidates for testing in the subsequent forecasting phase. To elucidate an issue encountered in the clustering process, Figure 16 visually represents the diverse medicines segmented into their respective clusters.

FIGURE 16 : Medicines grouped by clusters for 4 clusters (top), 6 clusters (middle) and 8 clusters (bottom)

The plots reveal a noteworthy pattern : each medicine is associated with multiple clusters. Given varying quantity levels for a given medicine, different cluster assignments emerge. Consequently, it becomes apparent that employing multiple forecasting models for the same medication is essential. Additionally, the data scarcity exacerbates another challenge : certain clusters comprise an insufficient number of data points. In some cases, this scarcity may hinder metric computation and model training, underscoring the impact of data availability on clustering outcomes and subsequent forecasting endeavors.

4.4.4 'Manual clustering'

An alternative avenue explored was the manual clustering approach, specifically within the context of the time series methodology. This involved exhaustively testing all conceivable combinations of medicines and hospitals, generating a population of 2310 pairs. From this exhaustive search, 206 pairs were identified for subsequent analysis. Interestingly, certain medications, like Doliprane, exhibited a lack of alignment with other medicines, signifying that their distinct characteristics and available data were insufficient to facilitate meaningful clustering alongside other medications.

To elucidate the manual clustering methodology, Figure 17 provides a visual representation of the tested pairs involving Hospital 1 and the other hospitals.

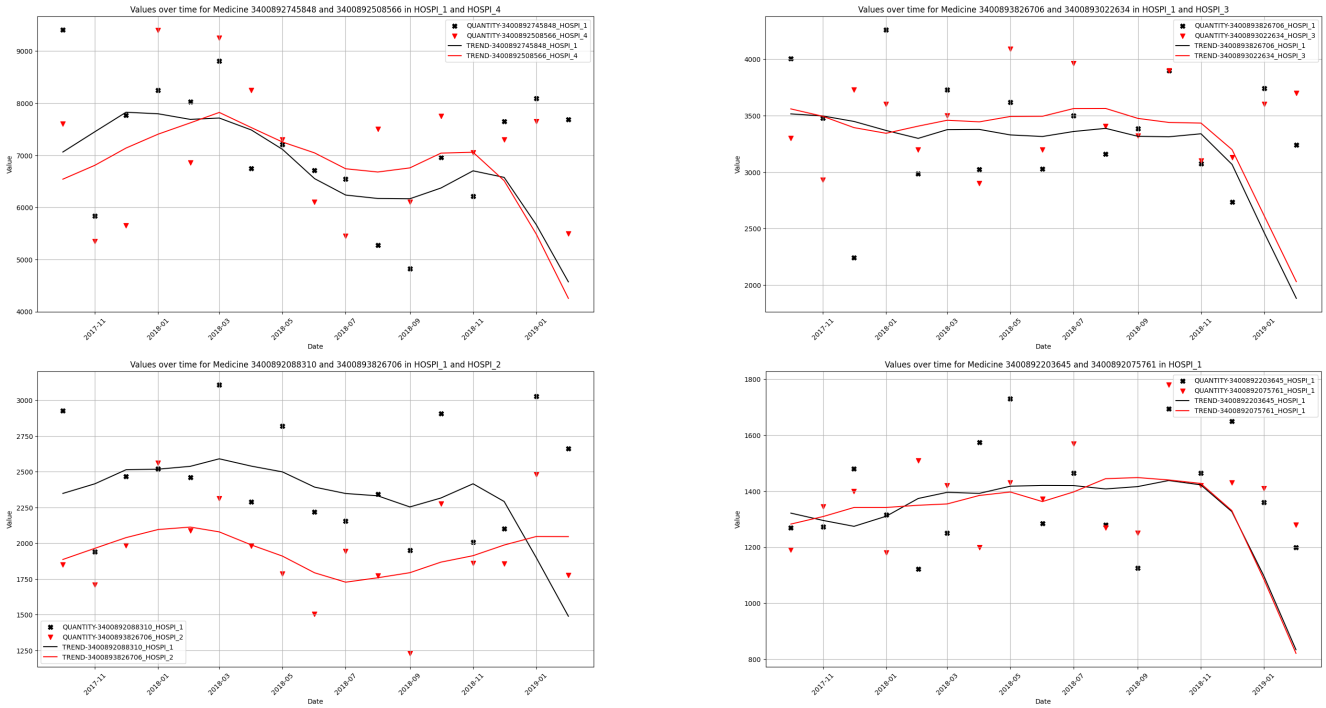

FIGURE 17 : Manual clustering of different medicines from different hospitals

Starting from the top left, the figure showcases the pairing of medicine 3400892745848 from hospital 1 and 3400892508566 from hospital 4. These two time series are synchronized within the same temporal span, facilitating a more nuanced examination of their resemblances. For instance, both series exhibit a comparable trend that commences its upward trajectory towards the close of 2017, peaking in March 2018, followed by a joint decline in both quantity and trend. The top-right portion illustrates the series of medicine 3400893826706 from hospital 1 and medicine 3400893022634 from hospital 3. Notably, despite the distinct hospital and medication sources, the similarity in trends is striking. This alignment raises the possibility that the data from these two combinations is so congruent that they could potentially be integrated into a shared forecasting model. In the bottom-left quadrant, a different form of similarity is observed. Here, although the consumption magnitudes differ, the shared trend pattern is apparent. Specifically, for medicine 3400892088310 from hospital 1 and medicine 3400893826706 from hospital 2, the trend displays consistent alignment despite the varying consumption levels. Lastly, the bottom-right section features medicines 3400892203645 and 3400892075761, both originating from hospital 1. Despite being distinct medications, the striking similarity in trends leads to an almost superimposed alignment of the two lines. This observation suggests the potential for improved forecasting outcomes when contrasted with the results obtained through the division approach.

Another noteworthy observation in this phase was the occurrence of intersections among distinct medicines within a single combination. This inadvertent clustering of components presents an intriguing avenue for investigation, warranting consideration in the subsequent forecasting phase.

Notebook(s) in : Git hub - 6_Manual_Clustering.

4.5 Forecasting

In the context of the forecasting phase, several key assumptions have been made. Primarily, it is crucial to note that after assessing the baseline performance for each method, any subsequent results following the clustering of medicines must demonstrate an improvement over or at least equivalence with the baseline metrics. Any opposite outcome is deemed unacceptable. Additionally, an essential facet of the prediction model's composition is its construction methodology. As discussed in earlier sections, prior research indicates that Random Forest Regressors models are well-suited for the problem at hand. Given this established premise, the grid search technique was chosen for parameter optimization to mitigate the risk of overfitting. The parameter range encompassed maximal depths from 2 to 8, n estimators ranging from 2 to 10% of each training set, and finally, the max features for tree splits were explored using options of square root, one feature, and two features. The train-test split was implemented for both the causal method and the time series method, with a division of 20% of the available data. In the context of the time series approach, it is important to note that the split was conducted without shuffling, ensuring the preservation of time-dependent characteristics inherent in the data. The baseline for the causal method with the division approach is presented inf Figure 18.

Id	CODE UCD	Hospital 1	Hospital 2	Hospital 3	Hospital 4
		MAPE	MAPE	MAPE	MAPE
1	CODE_UCD_3400892088310	5.00	0.06	0.14	0.80
2	CODE_UCD_3400892075761	0.21	0.09	0.09	0.08
3	CODE_UCD_3400892203645	0.44	0.12	0.09	0.10
4	CODE_UCD_3400892065366	0.22	0.18	0.08	0.16
5	CODE_UCD_3400892052120	0.16	0.11	0.18	0.12
6	CODE_UCD_3400891996128	0.13	0.07	0.05	0.09
7	CODE_UCD_3400893826706	6.80	0.10	0.11	0.14
8	CODE_UCD_3400893736135	0.90	0.22	0.17	0.18
9	CODE_UCD_3400893875490	1.05	0.05	0.02	0.05
10	CODE_UCD_3400890837149	2.42	0.10	0.07	0.19
11	CODE_UCD_3400891235203	0.39	0.21	0.43	0.37
12	CODE_UCD_3400891225037	21.04	0.09	0.05	0.21
13	CODE_UCD_3400891191226	0.10	0.11	0.10	0.09
14	CODE_UCD_3400892729589	0.20	0.06	0.03	0.31
15	CODE_UCD_3400892745848	0.13	0.22	0.28	0.08
16	CODE_UCD_3400892697789	5.08	0.15	0.21	0.11
17	CODE_UCD_3400892761527	5.31	0.11	0.06	0.32
18	CODE_UCD_3400893022634	0.29	0.14	0.08	0.33
19	CODE_UCD_3400892761695	0.14	0.08	0.04	0.22
20	CODE_UCD_3400892669236	0.24	0.04	0.03	0.15
21	CODE_UCD_3400892508566	0.56	0.27	0.13	0.19

FIGURE 18 : Division approach baseline with prediction scores for each hospital

In a comprehensive evaluation, it is apparent that the attained results for the MAPE metric fall short of the targeted performance level. The predefined threshold parameter, set at a maximum of 0.3 for the MAPE score, has not been met in a satisfactory manner. This inconsistency can be linked to the limited data availability stemming from the division of data by hospital. Consequently, the resulting diminished dataset hampers the quantity of data accessible for both training and testing, thus influencing the forecasting performance.

The baseline for the unified (the four hospitals togheter in the same data set) approach considering the causal method and time series is presented in Figure 19.

Id	CODE UCD	Unified dataset - Time series	Unified dataset - Causal
		MAPE	MAPE
1	CODE_UCD_3400892088310	0.10	2.59
2	CODE_UCD_3400892075761	0.11	0.64
3	CODE_UCD_3400892203645	0.24	1.35
4	CODE_UCD_3400892065366	0.15	0.29
5	CODE_UCD_3400892052120	0.16	5.87
6	CODE_UCD_3400891996128	28.48	28.48
7	CODE_UCD_3400893826706	0.22	74.65
8	CODE_UCD_3400893736135	0.14	0.48
9	CODE_UCD_3400893875490	0.07	11.84
10	CODE_UCD_3400890837149	55.67	0.58
11	CODE_UCD_3400891235203	0.18	2.41
12	CODE_UCD_3400891225037	6.75	3.34
13	CODE_UCD_3400891191226	0.24	0.39
14	CODE_UCD_3400892729589	15.88	11.19
15	CODE_UCD_3400892745848	0.20	6.99
16	CODE_UCD_3400892697789	0.26	1.54
17	CODE_UCD_3400892761527	0.25	7.02
18	CODE_UCD_3400893022634	12.41	26.69
19	CODE_UCD_3400892761695	0.39	1.82
20	CODE_UCD_3400892669236	0.12	0.11
21	CODE_UCD_3400892508566	3.20	0.39

FIGURE 19 : Unified approach baseline for causal method and time series

In a broader assessment, it is evident that the time series approach outperforms the causal method. This discrepancy can potentially be attributed to the exclusion of certain periods during the time series approach. These periods might have contained abnormal data points that did not accurately represent the predominant consumption behavior but had a significant impact on the overall forecasting results.

Despite the presence of gaps in the time series dataset, the model performance exhibited significant improvement compared to its counterpart. Further enhancements can be achieved by addressing these gaps, either through techniques like extrapolation or by leveraging machine learning capabilities available in libraries such as scikit-learn.

Calculating the financial cost of errors and classifying medicines based on their potential risk to human life is a valuable consideration. Assigning different weights to medicines according to their criticality can enhance model training and provide insights into improved cluster options. Medicines with higher potential risks indeed warrant more attention and precise forecasting due to their impact on patient

health and safety. This approach could lead to more targeted and accurate predictions, ensuring that critical medications are readily available when needed.

To test the clustering classification forecasting capacity for the clusters originated withe the agglomerative method that was the most positive compared with the others, each cluster was tested and the scores are presented in Figure 20 for the causal method and unified approach.

Unified dataset - Time series

ID	CODE UCD	CLUSTER	MAPE	ID	CODE UCD	CLUSTER	MAPE
	CODE_UCD_3400890837149	0	0.38		CODE_UCD_3400892669236	0	0.77
1	CODE_UCD_3400890837149	1	0.19	11	CODE_UCD_3400892669236	1	0.51
	CODE_UCD_3400890837149	2	1.81		CODE_UCD_3400892669236	2	54.27
	CODE_UCD_3400891191226	0	0.66		CODE_UCD_3400892697789	0	0.49
2	CODE_UCD_3400891191226	1	0.68	12	CODE_UCD_3400892697789	1	0.27
	CODE_UCD_3400891191226	2	0.22		CODE_UCD_3400892697789	2	13.61
	CODE_UCD_3400891225037	0	0.74		CODE_UCD_3400892729589	0	0.65
3	CODE_UCD_3400891225037	1	0.48	13	CODE_UCD_3400892729589	1	0.30
	CODE_UCD_3400891225037	2	11.33		CODE_UCD_3400892729589	2	0.47
	CODE_UCD_3400891235203	0	1.01		CODE_UCD_3400892745848	0	5.26
4	CODE_UCD_3400891235203	1	1.02	14	CODE_UCD_3400892745848	1	0.22
	CODE_UCD_3400891235203	2	161.15		CODE_UCD_3400892745848	2	0.51
	CODE_UCD_3400892052120	0	0.71		CODE_UCD_3400892761527	0	0.92
5	CODE_UCD_3400892052120	1	0.18	15	CODE_UCD_3400892761527	1	1.04
	CODE_UCD_3400892052120	2	9.63		CODE_UCD_3400892761527	2	1.17
	CODE_UCD_3400892065366	0	0.67		CODE_UCD_3400892761527	3	0.10
6	CODE_UCD_3400892065366	1	0.47		CODE_UCD_3400892761695	0	0.91
	CODE_UCD_3400892065366	2	0.39	17	CODE_UCD_3400892761695	1	0.56
	CODE_UCD_3400892075761	0	0.22		CODE_UCD_3400892761695	2	4.63
7	CODE_UCD_3400892075761	1	0.70		CODE_UCD_3400892761695	3	0.06
	CODE_UCD_3400892075761	2	8.30		CODE_UCD_3400893022634	0	13.12
	CODE_UCD_3400892088310	0	0.58	18	CODE_UCD_3400893022634	1	8.73
8	CODE_UCD_3400892088310	1	0.58		CODE_UCD_3400893022634	2	0.44
	CODE_UCD_3400892088310	2	30.55		CODE_UCD_3400893736135	0	0.59
	CODE_UCD_3400892203645	0	0.58	19	CODE_UCD_3400893736135	1	0.67
9	CODE_UCD_3400892203645	1	0.46		CODE_UCD_3400893736135	2	3.24
	CODE_UCD_3400892203645	2	1.51		CODE_UCD_3400893826706	0	0.64
	CODE_UCD_3400892508566	0	0.57	20	CODE_UCD_3400893826706	1	1.69
10	CODE_UCD_3400892508566	1	0.59		CODE_UCD_3400893826706	2	0.80
	CODE_UCD_3400892508566	2	23.18		CODE_UCD_3400893875490	0	0.84
				21	CODE_UCD_3400893875490	1	0.50
					CODE_UCD_3400893875490	2	0.68

FIGURE 20 : Prediction scores for the clustering technique in causal method and unified approach

On the whole, the scores attained through the clustering technique fall short of meeting the predefined limits set by the baseline. This recurring outcome further bolsters the contention that each medication possesses unique characteristics that impede effective grouping through clustering. This reaffirms the notion that the individuality of medicines renders them less amenable to clustering-based grouping. The unified approach is the only one presented here, as the division approach yielded unreliable results owing to the insufficiency of data.

As anticipated and discussed in earlier sections, the division of data among different clusters resulted in a reduction of data inputs for each cluster, compounding the issue of already limited data. This reduction had a noticeable impact on the overall performance of the final models and the prediction scores obtained. While certain medicines exhibited better scores than others, the average performance remained subpar. It's worth noting that some medicines were categorized into three clusters, while others were placed into four. This observation suggests that some medicines possess highly unique characteristics that warrant their placement in a distinct cluster.

In this section, the results for the four-cluster test were presented as it yielded the most promising outcomes among the various experiments conducted.

To comprehend why fewer clusters produced better results, it's essential to revisit the challenges outlined earlier. Fragmenting a small dataset into numerous clusters can detrimentally affect model training and, subsequently, performance. Ultimately, this issue is intertwined with the volume of accessible data, the number of clusters, and model performance. Considering that we're dealing with medicines and implying that an erroneous prediction leading to the absence of a necessary treatment could potentially impact patient well-being, the cost of model errors is prohibitively high. Consequently, this aspect cannot be overlooked when choosing the number of clusters.

The prediction scores for the time series approach are depicted in Figure 21.

Unified dataset - Time series - 4 clusters

ID	HOSPI_CODE_UCD	CLUSTER	MAPE	ID	HOSPI_CODE_UCD	CLUSTER	MAPE
1	CODE_UCD_3400890837149	0	0.20	11	CODE_UCD_3400892669236	0	0.62
	CODE_UCD_3400890837149	1	0.32		CODE_UCD_3400892669236	1	0.49
	CODE_UCD_3400890837149	2	28.41		CODE_UCD_3400892669236	2	8.17
	CODE_UCD_3400890837149	3	0.74		CODE_UCD_3400892669236	3	0.86
2	CODE_UCD_3400891191226	0	0.28	12	CODE_UCD_3400892697789	0	1.32
	CODE_UCD_3400891191226	1	0.52		CODE_UCD_3400892697789	1	0.26
	CODE_UCD_3400891191226	2	0.30		CODE_UCD_3400892697789	2	1.63
	CODE_UCD_3400891191226	3	0.36		CODE_UCD_3400892697789	3	0.91
3	CODE_UCD_3400891225037	0	0.61	13	CODE_UCD_3400892729589	0	0.36
	CODE_UCD_3400891225037	1	0.51		CODE_UCD_3400892729589	1	0.28
	CODE_UCD_3400891225037	2	0.88		CODE_UCD_3400892729589	2	0.63
	CODE_UCD_3400891225037	3	0.87		CODE_UCD_3400892729589	3	22.21
4	CODE_UCD_3400891235203	0	13.37	14	CODE_UCD_3400892745848	0	8.06
	CODE_UCD_3400891235203	1	0.92		CODE_UCD_3400892745848	1	0.16
	CODE_UCD_3400891235203	2	0.31		CODE_UCD_3400892745848	2	17.43
	CODE_UCD_3400891235203	3	1.09		CODE_UCD_3400892745848	3	0.76
5	CODE_UCD_3400892052120	0	1.63	15	CODE_UCD_3400892761527	0	0.86
	CODE_UCD_3400892052120	1	0.52		CODE_UCD_3400892761527	1	0.93
	CODE_UCD_3400892052120	2	0.13		CODE_UCD_3400892761527	2	0.85
	CODE_UCD_3400892052120	3	0.32		CODE_UCD_3400892761527	3	0.60
6	CODE_UCD_3400892065366	0	0.47	16	CODE_UCD_3400892761695	0	0.84
	CODE_UCD_3400892065366	1	0.37		CODE_UCD_3400892761695	1	0.17
	CODE_UCD_3400892065366	2	0.68		CODE_UCD_3400892761695	2	0.90
	CODE_UCD_3400892065366	3	0.63		CODE_UCD_3400892761695	3	0.13
7	CODE_UCD_3400892075761	0	0.58	17	CODE_UCD_3400893022634	0	0.19
	CODE_UCD_3400892075761	1	0.20		CODE_UCD_3400893022634	1	0.79
	CODE_UCD_3400892075761	2	0.20		CODE_UCD_3400893022634	2	0.68
	CODE_UCD_3400892075761	3	0.49		CODE_UCD_3400893022634	3	0.48
8	CODE_UCD_3400892088310	0	0.19	18	CODE_UCD_3400893736135	0	0.20
	CODE_UCD_3400892088310	1	0.43		CODE_UCD_3400893736135	1	0.51
	CODE_UCD_3400892088310	2	0.43		CODE_UCD_3400893736135	2	0.59
	CODE_UCD_3400892088310	3	0.39		CODE_UCD_3400893736135	3	0.15
9	CODE_UCD_3400892203645	0	0.26	19	CODE_UCD_3400893826706	0	0.46
	CODE_UCD_3400892203645	1	0.17		CODE_UCD_3400893826706	1	1.14
	CODE_UCD_3400892203645	2	0.48		CODE_UCD_3400893826706	2	7.80
	CODE_UCD_3400892203645	3	0.20		CODE_UCD_3400893826706	3	0.44
10	CODE_UCD_3400892508566	0	0.09	20	CODE_UCD_3400893875490	0	0.71
	CODE_UCD_3400892508566	1	0.35		CODE_UCD_3400893875490	1	0.51
	CODE_UCD_3400892508566	2	40.88		CODE_UCD_3400893875490	2	0.79
	CODE_UCD_3400892508566	3	0.48		CODE_UCD_3400893875490	3	0.83

FIGURE 21 : Prediction scores for the clustering technique in the time series methodology

Despite employing features derived from seasonal decomposition, the prediction scores did not align with the anticipated outcomes. Notably, the MAPE metrics for each medicine did not meet the baseline expectations for the majority of values. This outcome reaffirms that the time series approach was not a suitable fit for the problem at hand. Notably, experiments involving 8 clusters, 12 clusters, and 18 clusters resulted in even poorer metrics. As a result, only the results with 4 clusters are presented here.

Within this section, each pair is associated with a score, indicating the predictive performance of each medicine within the pair when compared to other clusters or methods. Interestingly, the prediction scores resulting from this technique exhibit superior performance. This could be attributed to the fact that the current dataset comprises combinations of medicines and hospitals that do possess certain similarities. However, it's important to note that these shared characteristics represent a relatively small portion of the total dataset. This raises the possibility that the employed clustering techniques in this study are

insufficient to effectively group the data due to the limited presence of substantial similarities within the dataset.

Indeed, it's essential to provide a thorough examination of the manual clustering approach's superior performance and offer a detailed analysis of these results. This will facilitate a deeper understanding of the factors driving the success of manual clustering and highlight its potential for future applications. The results can be visually represented in Figure 22 for clarity and reference.

FIGURE 22 : Scores of each pair in manual clustering

The positive results observed in this study may suggest that manual clustering was a suitable choice, and the performance of the tested models is satisfactory. However, it's crucial to consider that out of the 2,310 possible combinations between medicines and hospitals, only 206 were selected for testing, representing approximately 10% of the total combinations. This implies that some combinations were not explored in this study. Addressing this issue in future research by testing and evaluating every possible combination with respect to the MAPE metric would provide a more comprehensive understanding of the model's performance and its applicability in a broader context.

Exploring the practical application of this manual clustering approach is an important avenue for

future research. It raises questions about how this clustered information can be effectively utilized in a commercial solution and the accuracy of the predictions it would yield. For instance, consider a scenario where a new hospital seeks to predict its medicine consumption for the next six months. How can the parameters derived from the two medicines that formed a pair in the manual cluster provide valuable insights for this new hospital ? Addressing these questions and further analyzing this issue will be essential for translating this research into practical solutions that benefit healthcare facilities.

To exemplify the performance of the manual clustering technique, Figure 23 portrays a specific pair comprising medicine 3400892745848 from hospital 1 and medicine 3400891191226 from hospital 3, along with their corresponding prediction scores.

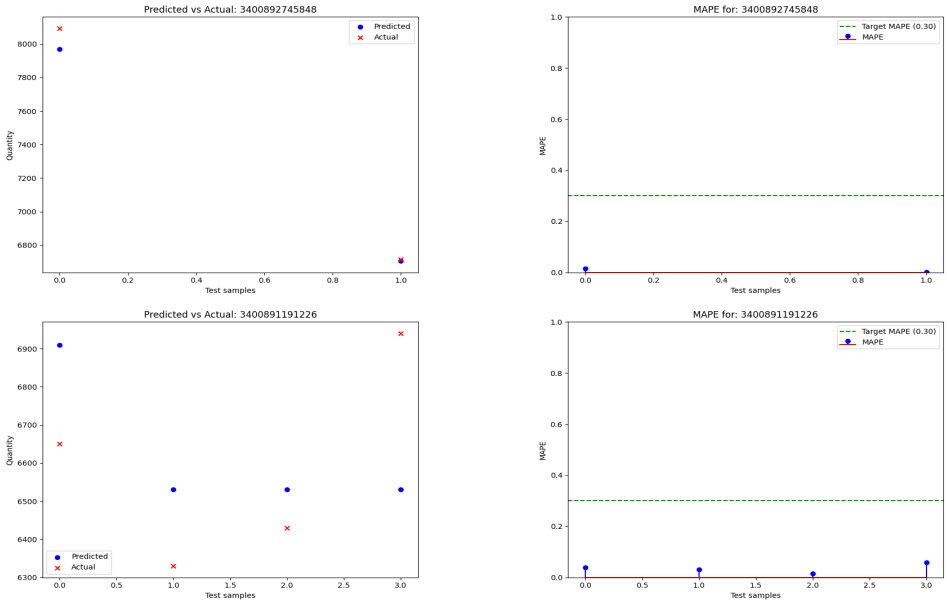

FIGURE 23 : Prediction scores for medicine 3400892745848 (top) and medicine 3400891191226 (bottom)

It is evident the good performance of the model predicting the consumption for the medicines even tho there is not a significant amount of data to be trained and tested for each pair, this could be used in further studies and be explored in deep.

In summary, the forecasting validation phase has provided valuable insights into the tested clustering techniques and potential directions for future research. Various tests can be employed to evaluate the quality of the generated models. For instance, assessing how well models trained on a subset of hospitals in the main dataset perform when applied to a simulated new hospital. Similarly, testing the models with some medicines and then evaluating their performance on new medicines. These tests were proposed during the internship but were not implemented due to time constraints. Exploring these avenues in future work could further enhance the applicability and robustness of the developed models and clustering techniques.

Indeed, the choice of regression models for clustering methods and time series techniques can significantly impact the forecasting performance. It's worth exploring whether Random Forest Regressors remain the best option for clustering methods and if dynamic time warping could lead to improved results in the time series approach. This question presents an interesting avenue for future research, where various regression models and techniques can be thoroughly tested and compared to identify the most suitable ones for each specific application scenario.

Notebook(s) in : Git hub - 5_Forecasting.

5 Conclusion

Traversing through the various stages of this study, a concise summary can be provided. The initial exploration of prior research in the field facilitated a deeper comprehension of the underlying problem, the structure of the data set, and the potential influences of individual features on the anticipated outcomes. Upon scrutinizing the entries within the data frame, several points necessitating formatting and data handling were discerned. This prompted the exploration of distinct approaches, namely the causal method and the time series method, leading to the consideration of both individual hospital-based and unified approaches. A variety of algorithms and clustering techniques were experimented with, ultimately revealing that agglomerative clustering yielded the most noteworthy silhouette metric scores. An issue identified that warrants addressing in future endeavors is the potential for each medicine to reside in multiple clusters within this technique. Consequently, the need arises to employ multiple models to predict the consumption of a single medicine, which introduces a complexity surpassing that of the initial problem. The application of forecasting techniques furnished valuable insights regarding the preceding stages. Notably, among the options explored, the manual clustering technique emerged as the most aligned with the project's objectives.

The complexity of the problem at hand was anticipated from the outset and has indeed validated those initial suspicions. Each medicine exhibits distinct characteristics and consumption patterns so inherently unique that aggregating multiple medicines into the same model cluster proves challenging. Reflecting on the outcomes of the testing and considering the more promising results, the manual clustering approach has demonstrated potential to surpass baseline expectations. However, in practical terms, its usability could be limited. Consider, for instance, a scenario involving a new hospital seeking to predict the consumption of medicines over the next six months to a year. In such cases, the dearth of data or means to effectively organize information for input into the models can hinder the application of manual clustering.

In assessing the outcomes of the conducted research across various techniques, it becomes apparent that, given the current context of available data and hospital information, the unified approach with each medicine having its distinct forecasting model, as developed by PhD student D. KOALA, emerges as the preferred choice.

As we look ahead to future endeavors and seek to address the gaps left by the constraints of time in the current study, an intriguing avenue for exploration lies in dynamic time warping (DTW). DTW presents a time series approach that holds promise in detecting and amalgamating distinct series based on their individual components. Although this subject is intricate and lacks extensive information and studies in the context of the current problem, delving into DTW could offer valuable insights and open new dimensions for enhancing our understanding of medication consumption forecasting. (Müller (2007), Oates et al. (1999)).

The completion of this project has imparted invaluable insights and practical application of the theoretical concepts learned. It has bridged the gap between academic knowledge and real-world problem-solving, aligning with the needs of a practical scenario within a company. I extend my sincere gratitude to Prof. Dr. Zakaria YAHOUNI, Prof. Dr. Gulgun ALPAN, and PhD student Denis KOALA for their unwavering support, guidance, and knowledge-sharing throughout my tenure at G-SCOP laboratory. Their mentorship has been instrumental in shaping my understanding and skills in this field.

References

François Hada*. Les sources d'informations et de données sur le médicament. *Revue française des affaires sociales*, 1(3) :087–098, 2007.

Denis Koala, Zakaria Yahouni, Gülgün Alpan, and Yannick Frein. Factors influencing drug consumption and prediction methods, 2021.

Denis Koala, Zakaria Yahouni, Gülgün Alpan, and Djamal Si Mohand. Correlation analysis of factors impacting health product consumption in French hospitals. *IFAC-PapersOnLine*, 55(10) :895–900, 2022. ISSN 24058963. doi : 10.1016/j.ifacol.2022.09.415.

Denis Koala, Zakaria Yahouni, Gülgün Alpan, and Djamal S I Mohand. Machine learning versus consumption-based models for predicting medicine demand in french hospitals : a case study. -, 2023.

Stanley Suan You Lim, Siao-Leu Phouratsamay, Zakaria Yahouni, and Eric Gascard. Medicine consumption demand forecasting in french hospitals using SARIMA model. Master's thesis, Your University Name, 2023.

Wes McKinney, Josef Perktold, and Skipper Seabold. Time series analysis in python with statsmodels. *Jarrodmillman Com*, pages 96–102, 2011.

A.C. Muller and S. Guido. *Introduction to Machine Learning with Python : A Guide for Data Scientists*. O'Reilly Media, Incorporated, 2018. ISBN 9789352134571. URL `https://books.google.fr/books?id=jGdXswEACAAJ`.

Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

Kevin P. Murphy. *Probabilistic Machine Learning : An Introduction*. MIT Press, 2022. URL `probml.ai`.

Tim Oates, Laura Firoiu, and Paul R Cohen. Clustering time series with hidden markov models and dynamic time warping. In *Proceedings of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning*, pages 17–21. Citeseer, 1999.

OECD. *Health at a Glance 2013 : OECD Indicators*. OECD Publishing, 2013. URL `http://dx.doi.org/10.1787/health_glance-2013-en`. DOI : 10.1787/health_glance-2013-en.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 : 2825–2830, 2011.

Skipper Seabold and Josef Perktold. statsmodels : Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

The pandas development team. pandas-dev/pandas : Pandas. Zenodo, Feb 2020. URL `https://doi.org/10.5281/zenodo.3509134`.

Daniel Vélez, Siao-Leu Phouratsamay, Zakaria Yahouni, and Gülgün Alpan. Predicting medicine demand fluctuations through markov chain. In *12th International Workshop on Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future*, pages 329–340, Bucharest, Romania, September 2022. doi : 10.1007/978-3-031-24291-5_26. URL `https://hal.archives-ouvertes.fr/hal-03951534`.

Appendix

A Distribution of different hospitals before removing outliers

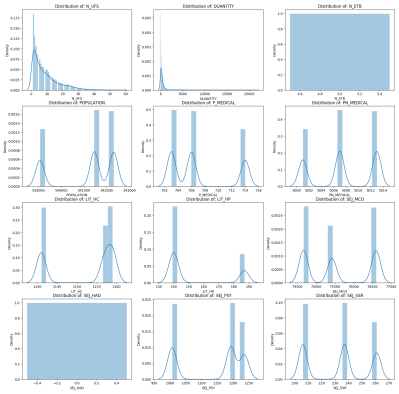

FIGURE 24 : Numerical variables distribution before removing the outliers - Hospital 2

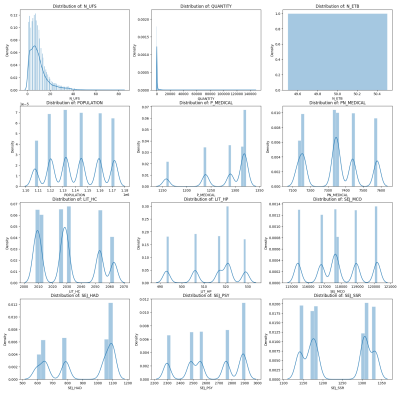

FIGURE 25 : Numerical variables distribution before removing the outliers - Hospital 3

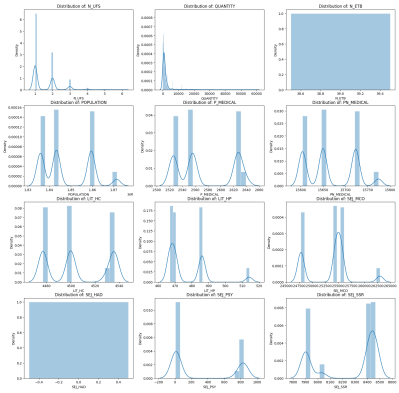

FIGURE 26 : Numerical variables distribution before removing the outliers - Hospital 4

B Distribution of different hospitals after removing outliers

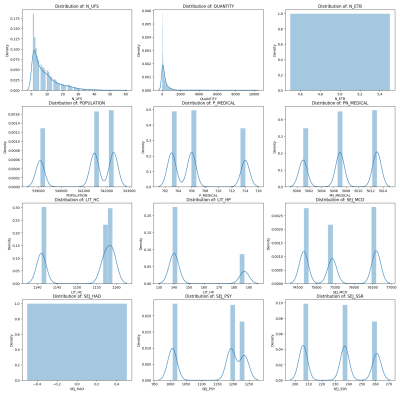

FIGURE 27 : Numerical variables distribution after removing the outliers - Hospital 2

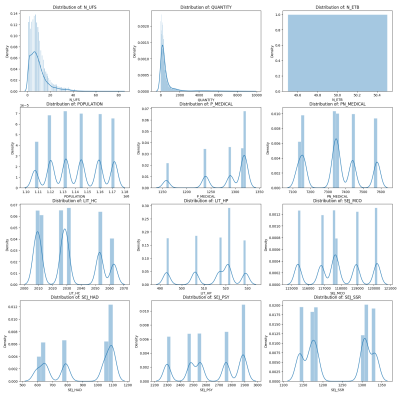

FIGURE 28 : Numerical variables distribution after removing the outliers - Hospital 3

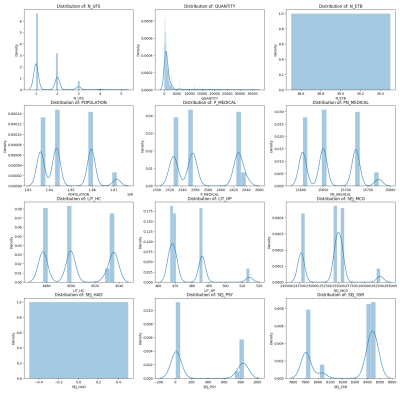

FIGURE 29 : Numerical variables distribution after removing the outliers - Hospital 4