The Islamic University of Gaza
Research and Postgraduate Affairs
Faculty of Information Technology
Master of Information Technology

الجامعـــــــــة الإســــلاميــة بـغــزة
شئون البحث العلمي والدراسات العليا
كـليـــــــــــــة تكنولوجيا المعلومات
ماجستيــــــر تكنــولوجيا المعلــومات

Medium-Term Forecasting for Municipal Water Demand and Revenue

(KhanYounis City as A Case Study)

توقع متوسط المدى لإحتياج و عوائد المياه في البلديات

(مدينة خانيونس كدراسة حالة)

By

Hosam Hasan Abdullah Mukhairez

Supervised by

Dr. Alaa El-Halees

Prof. of Computer Science

**A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Information Technology**

February/2018

إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Medium-Term Forecasting for Municipal Water Demand and Revenue
(KhanYounis City as A Case Study)

توقع متوسط المدى لإحتياج و عوائد المياه في البلديات

(مدينة خانيونس كدراسة حالة)

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الاخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

I understand the nature of plagiarism, and I am aware of the University's policy on this. The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted by others elsewhere for any other degree or qualification.

Student's name:	حسام حسن عبدالله مخيرز Hosam Hasan Mukhairez	اسم الطالب:
Signature:		التوقيع:
Date:		التاريخ:

بسم الله الرحمن الرحيم

الجامعة الإسلامية غزة
The Islamic University of Gaza

عمادة البحث العلمي والدراسات العليا
هاتف داخلي: 1150

الرقم: ج س غ/35/ Ref:
التاريخ: 2018/03/06م Date:

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة عمادة البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **حسام حسن عبد الله مخيرز** لنيل درجة الماجستير في كلية **تكنولوجيا المعلومات/** قسم **تكنولوجيا المعلومات** وموضوعها:

توقع متوسط المدى لإحتياج وعوائد المياه في البلديات - مدينة خانيونس كدراسة حالة

Medium-Term Forecasting for Municipal Water Demand and Revenue KhanYounis City as a Case Study

وبعد المناقشة التي تمت اليوم **الثلاثاء 18 جمادي الثانية 1439**هـ الموافق **2018/03/06**م الساعة **العاشرة صباحاً**، في قاعة مبنى **القدس** اجتمعت لجنة الحكم على الأطروحة والمكونة من:

أ. د. علاء مصطفى الهليس	**مشرفاً ورئيساً**	....
د. توفيق سليمان برهوم	**مناقشاً داخلياً**	....
د. إيهاب صلاح زقوت	**مناقشاً خارجياً**	....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية تكنولوجيا المعلومات/قسم تكنولوجيا المعلومات**.

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله تعالى ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق،،،

عميد البحث العلمي والدراسات العليا

أ. د. مازن إسماعيل هنية

Abstract

The combination of forecasting and planning has increased prominence within few decades and now receives considerable attention from both academics and practitioners. By operating forecast tasks, one can be able to get more accurate results with less usage of resources which means saving available resources for future. Moreover, to properly plan, it is important to be able to forecast resource availability. In the past, forecasting available resources depends on traditional probability techniques.

Palestine municipalities are facing complex circumstances because of the climate change (lack of rainfall) and due to the Israeli occupation. Water resources are decreasing by the passage of time. Therefore, it is necessary to focus on water service projects for best exploit of resources given the growth of population. That is because forecasting water consumption used by citizens is very important in directing decision makers to have a clear perspective.

In this research, we forecasted the consumption used by the existing water subscriptions in the medium-term future and forecasted revenues and profits resulted from water service given the growth of population by the passage of time. Then, we forecasted consumptions and revenues for each sub area in medium-term future. That is important for decision makers to help them focus their investment projects according to accurate results of forecasting water demand and to benefit from the huge data available. In addition, water service suppliers and providers should share how much water they expect to sell in order to ensure that revenues will cover costs. Water rates are typical and only set once every few years, so water agencies must forecast future demand several years in advance to ensure that they collect enough funds.

We conducted the forecasting using the following forecasting algorithms: Auto-Regressive Integrated Moving Average (ARIMA), Hybrid ARIMA, Singular Spectrum Analysis (SSA), and Linear Regression. We applied them on dataset collected from KhanYounis municipality (KHM)-department of customer services. We found that the best algorithm is Hybrid ARIMA which gave the Mean Absolute Percentage Error (MAPE) of 17.38%. Finally, we forecasted three levels, 1st level is the whole city, 2nd level is the sub-areas and the 3rd level is the classes inside KhanYounis city. Generally, we found that the maximum water consumption for the overall city after five years will increase to about 8.4% compared to 2017, but the minimum water revenue will decrease to about 3.8% compared to 2017.

Keywords: Forecasting, water demands, water revenues, Artificial neural networks, ARIMA, Hybrid ARIMA, Linear Regression, Singular Spectrum Analysis.

الملخص

الدمج بين علم التوقع و علم التخطيط زاد الاهتمام به في العقود الاخيرة و الان يتلقى اهتمام الباحثين في هذا المجال. من خلال تنفيذ عمليات التوقع المستقبلي ، يمكن الحصول على نتائج أكثر دقة باستغلال اقل قدر ممكن من المصادر المتاحة مما يؤدي الى توفير هذه المصادر المتنوعة للمستقبل. علاوة على ذلك، انه من المهم جدا الاستفادة من التوقع و استخدام المصادر المتوفرة من اجل التخطيط بصورة صحيحة و مهنية. في الماضي، كان التوقع يعتمد على الطرق و الوسائل التقليدية.

البلديات في دولة فلسطين المحتلة تواجه العديد من الظروف الصعبة في التعامل مع المياه و مصادرها نتيجة تغير المناخ و الاحتلال الاسرائيلي. و بالتالي، انه من الجدير بالاهتمام التركيز على المشاريع التي تخدم قطاع المياه لاستغلال هذه المصادر بالطريقة المثلى خصوصا مع الزيادة السكانية. و ذلك بسبب ان التوقع بمدى حاجة المدينة و السكان للمياه عامل مهم جدا لصناع القرار.

في هذا البحث قمنا بالتوقع متوسط المدى بكميات استهلاك المياه على مستوى المدينة بشكل عام و على مستوى المناطق المقسمة داخل مدينة خان يونس و ايضا التوقع بقيمة العائدات المالية من توفير خدمة المياه خصوصا مع ازدياد السكان مع مرور الزمن. هذه الخطوة مهمة جدا لصناع القرار لمساعدتهم في تركيز المشاريع الاستثمارية اعتمادا على نتائج دقيقة و تحليل علمي و الاستفادة من البيانات الضخمة المتوفرة في البلديات. لذلك، يتوجب على مقدمي خدمات المياه و داعمي المشاريع المائية مشاركة كميات المياه التي يتوقعون تقديمها للمواطنين لضمان أن الإيرادات ستغطي التكاليف.

لقد قمنا باجراء عمليات التوقع بمستقبل المياه و العائدات باستخدام اربع خوارزميات و هي نموذج المتوسط المتحرك للانحدار الذاتي (ARIMA) و النموذج الهجين للمتوسط المتحرك للانحدار الذاتي(Hybrid ARIMA) و تحليل الطيف المفرد(SSA) و الانحدار الخطي (Linear Regression). لقد قمنا بتنفيذ هذه الخوارزميات على بيانات تم جمعها من قسم الجباية و خدمات المشتركين في بلدية خان يونس. لقد لاحظنا ان افضل خوارزمية اعطت نتائج اقرب للبيانات الحقيقية هي Hybrid ARIMA و كانت قيمة متوسط النسبة المئوية المطلقة تساوي 17.38%. و في الختام قمنا باستخدام هذه الخوارزمية للتوقع بمستقبل المدينة على ثلاث مستويات: (1) المدينة بشكل عام، (2) على المناطق الفرعية داخل المدينة و عددها 20 ، (3) على الفئات الفرعية في داخل المدينة و عددها 13. بشكل عام لاحظنا ان اقصى احتياج لمدينة خان يونس من المياه خلال الخمس سنوات القادمة سيزداد حوالي بنسبة 8.4% و العائدات المالية ستقل حوالي بنسبة 3.8% مقارنة مع عام 2017.

الكلمات المفتاحية: التوقع، احتياج المياه، عائدات المياه، الشبكات العصبية الاصطناعية، الانحدار الخطي، تحليل الطيف المفرد، نموذج المتوسط المتحرك للانحدار الذاتي، النموذج الهجين للمتوسط المتحرك للانحدار الذاتي.

Dedication

I dedicate this to my mother, father, brothers, sisters, wife, son and daughters. I dedicate this because now you are happy and proud of me, and I am also happy that Great Allah passed me this important stage in my life. I cannot forget your reassurance and questions about me how am I doing in master studies and in thesis progress. Without your prayers this success would not have been completed.

I dedicate this to my great university IUG, which contained me and taught me from bachelor to master's degree. I dedicate this work to my second home KhanYounis Municipality, which provided me with the appropriate environment and data.

Finally, I dedicate this work to everyone is happy to me and prayed Allah to success me.

Acknowledgment

First and foremost, I wish to thank Great Allah for giving me strength, patience and reconciliation to complete my master studies and this thesis, my deep thankful fly to my parents for their financial and moral support, my great thanks to my wife and family for being patient and carrying me up in my master studies, and to those who have assisted and inspired me throughout this research.

There are so many people to whom I am indebted for their assistance during my endeavors to complete my master's candidate in information technology at the Islamic University. First and foremost, I would like to express my gratitude and thankful to my supervisor Prof. Alaa EL-Halees whose invaluable guidance and support was very helpful throughout my research. I don't say that courtesy, but he really helped me in solving problems and gave me useful ideas in analyzing and conducting my experiment work. Gratitude to Information technology lecturer staff, Dr. Rebhi Baraka, Dr. Tawfig Barhoom, Dr. Iyad Al-Agha, Dr. Ashraf AL-Meghary, Dr. Moataz Saad and Dr. Ashraf Al-Atar. Besides that, I would like to thank my family, friends and everyone who supported me, guided me and advised me in this thesis.

My appreciation also goes to KhanYounis municipality my second home for providing me the customer data, overcoming difficulties, and other helpful information about business logic and functionality. I express my appreciation to everyone who has involved directly and indirectly to the success of this research.

Table of Content

List of Abbreviations

ANN	Artificial Neural Network.
ARIMA	Auto Regressive Integrated Moving Average.
ASAR	Adaptive Seasonal Auto Regressive.
BFO	Bacterial Foraging Optimization.
CSO	Cat Swarm Optimization.
DE	Differential Evolution.
ETS	Exponential Smoothing State Space.
FBLMS	Forward Backward Least Mean Square.
FSAR	Fixed Seasonal Auto Regressive.
GIS	Geographic Information System.
GUI	Graphical User Interface.
GP	Genetic Programming.
Hybrid ARIMA	Auto Regressive Integrated Moving Average with ANN.
IOT	Internet of Things.
KHM	KhanYounis Municipality.
MAD	Mean Absolute Deviation.
MAPE	Mean Absolute Percentage Error.
MOH	Ministry of Health.
MSE	Mean Squared Error.
NIS	New Israeli Shekel.
PSO	Particle Swarm Optimization.
ROI	Return of Investment.
SSA	Singular Spectrum Analysis.
SVM	Support Vector Machine.
TSFF	Time Series Forecasting Framework.
UNRWA	United Nations Relief and Works Agency.

List of Figures

List of Tables

Chapter 1

Introduction

Chapter 1
Introduction

1.1. Background and Context

Many countries are facing the real risk in lack of water resources because of the changing in the earth climate with the passage of time. Water resources in Palestine are limited according to low groundwater level due to lack of rainfall, increasing of population, also due to the Israeli occupation which prevents the Palestinian authorities to use many of the ground wells, and the unorganized drilled wells by citizens to bring up the underground water (EWASH, 2016).

Blue water resources, including those shared with Israel, are estimated at approximately 2,989 mcm per annum. These resources include groundwater, representing approximately 1,454 mcm; surface water, especially due to the natural circulation of the Jordan River, estimated at 1,320 mcm; and runoffs, which make-up an estimated 215 mcm. of these resources, approximately 2,570 mcm are used for various purposes, the share that Palestinians utilize represent a mere 271 mcm, (i.e. around 11%), while the remaining 89% is exploited by Israel (Barakat, 2013).

Many water service projects (for example water desalination and drilling of underground wells) were built by local and international organizations in many regions in Gaza strip and West Bank to support water service to citizens, but these projects are not established according to preconceived data belongs to the increasing of population, current per-capita water consumption and other factors may affect the process of distributing water to people.

To design a sufficient water supply system, engineers should evaluate the consumption water amount that is required, known as the ‘water demand’. So, water demand can be defined as the water consumed by beneficiaries to satisfy their needs. There are generally several classes of water consumption that can be widely grouped into following categories (Ripley, 2015) :

1. Domestic water demand.
2. Public water use.
3. Commercial water use.
4. Industrial water use.
5. Fire demand.
6. Irrigation water demand.
7. Losses and wastes.

Data mining (or knowledge discovery KDD) is the function of analyzing data from several viewpoints and summarizing it into useful information, information that can be used in supporting decision making and to increase earning, cuts costs, or both, also it could increase the value of social and economic norms. It allows researches and customers to analyze data from different viewpoints dimensions or angles, classify it, and summarize the discovered relationships. Scientifically, data mining is the procedure of finding correlations or patterns among dozens of fields in large relational databases (Oracle, 2016).

Forecasting is a subfield of data mining technology which its primary function to predict the future trends or values using (time series related) data we have in hand. Forecasting task is analyzing the past trends and behaviors of a variable to estimate or predict its future behavior or trend. It is one of the oldest known predictive analytics techniques. The idea is to use historical data to make forecast and prediction about future data. Forecasting is one of the most common techniques of time series data analysis. It's used to predict future trends in water demands, retail sales, economic indicators, weather and climate forecasting, stock prices, and other application (Aggarwal, 2015).

The combination of forecasting and planning has increased to prominence within a few decades and now receives considerable attention from both academics and practitioners. By operating forecasting tasks, we will be able to get more accurate results with less usage of resource which means save available resource for future by supporting effective decision making and for water researchers seeking to extend the current knowledge in the field. Moreover, to properly plan, it is important to be able to forecast your resource availability. In the past, resource forecasting relied on probability techniques. While metrics can help contribute to the process, prioritizing projects through simple Return of Investment (ROI) scoring can be a gamble in agency life (Ripley, 2015).

Water demands amount guides to build proper water distribution systems, and reliable demand forecasting helps in simulating and managing such systems. Depending on the forecasting time horizon term, water demand forecasts can be classified, as suggested by (Billings & Jones, 2011), as shown in ***Table (1-1***).

Table (1-1) Types of Water Demand Forecasts and Major Applications

Forecast Type	Forecast Horizon	Applications
Long-Term	Decades	Sizing system capacity, raw water supply
Medium-Term	Years to a decade	Sizing, staging treatment and distribution system improvements
Short-Term	Years	Setting water rates, revenue forecasting, program tracking and evaluation
Very Short-Term	Hours, days, weeks (up to two weeks)	Optimizing, managing system operations, pumping

Water service managers forecast water service data for a variety of purposes. These analyses can help managers understand spatial and timely patterns of future water use to optimize system operations, plan for future water purchases or system expansion, or for future revenue and expenditures.

In 2008, a study by Palestinian Central Bureau of Statistics (PCBS) shows the population of KhanYounis city and the increase rate for the following years until year 2025 will reach 22.5% with comparison in 2018.

From previous introduction population in Palestine in general and Gaza Strip in specific is growing up by time and water resources are limited due to low rainfall annual amounts or by Israeli politics against Palestinian Authority about water underground resources. So, handling these factors in future water service projects is very important issue for mayors of municipalities and decision makers.

1.2. Statement of the problem

Municipalities have many data belongs to water services. Unfortunately, according to our experience in local municipalities, these data are not used by all managers in municipalities in planning and developing their strategic plans and decisions depending on accurate assessment of data.

'The major problem is Gaza Strip municipalities are not using forecasting science, which means decision makers and funders of water service projects deprived of having a clear predictive view of the future status of water demand and revenue'. So, by using forecasting algorithms and techniques they will be able to have a clear perspective while developing underground water distributing network depending on the most areas that consumes water amounts and the increasing of new established units. And because our country is economically unstable, they can predict the revenue they will gain in future and present the relationship between water consumption and related water revenue.

1.3. Objectives

1.3.1. Main objective

Our objective is to propose a method depending on best forecasting algorithm that estimate medium-term future water demand and revenue for a given area (e.g. Al-Amal, Al-Mahata. etc.) over a given time (1-5 years), also for a given class (Banks, Houses, Mosques. etc.) over a given time. We take KhanYounis city as a case study.

1.3.2. Specific objective

The specific objectives of the project are:

1. Collect data from normal database and applications which belongs to water services.
2. Apply some preprocessing tasks like cleaning, integrating, filling missing data, abandonment of strange or abnormal data …etc. and arranging data as a time series dataset to be ready and suitable for forecasting algorithms
3. Select the best forecasting algorithm from (ARIMA, Hybrid ARIMA, Linear Regression and Singular Spectrum Analysis) by applying the methods on our data and measure MAPE.
4. Using the best algorithm in forecasting water demand and revenue for the city of KhanYounis.
5. To estimate the future of water demand for each sub-area in KhanYounis city (20 sub-area exists).
6. To forecast the water demand and water revenue for some classes.

1.4. Importance of the project

Forecasting time data especially for water demands and revenues are very important in designing of water distribution networks and predicting the capacity of the water pumps to support an accurate decision making. The importance of our proposed model stems from:

1. Estimating current per-capita water consumption
2. Support decision makers to help them focus their investment projects according to results of forecasting water demand.
3. To benefit from the huge data available in KhanYounis municipality.
4. The approach can be used by other cities to estimate future demand of water.
5. The approach could be used for other utilities such as electricity and Gas.

6. To improve Gaza Strip municipality's capabilities and decision-making process by making them depends on forecasting methods.
7. Taking care of water consumption and revenue data and conduct some technical research and operate certain data mining algorithms will helps decision makers and managers to have a clear perspective specially while holding water service projects and development task on the underground distributing network.

1.5. Scope and limitations of the project

1.5.1. Main Scope

The main scope of this study is analyzing and forecasting the water demand and revenue data at KhanYounis municipality KHM in advance 5 years and because our country is economically unstable, we study the relationship between water consumption and related water revenue, is it positive relationship or not? The idea of this research assumption is limited to any financial billing system that's responsible for managing and calculating water consumption and water revenue for customers. This research will focus on KhanYounis Municipality historical transactions data.

1.5.2. Main limitations

1. Data collected from 01-2007 to 10-2017 of water consumption and water revenue.
2. Our work focuses only on forecasting water consumption and water revenue.
3. Our work focuses only on forecasting the whole city, sub-regions and classes.
4. Forecasting horizon is 5 years only.
5. An existing forecasting method are planned to be used in this study.

1.6. Methodology

Our methodology depends on preparing data collected from KhanYounis Municipality databases for time series forecasting tasks and operate some preprocessing tasks (merging redundant records, remove noise data) over three sub-datasets (overall city dataset, regions dataset and classes dataset).

Then evaluate the selected algorithms (ARIMA, Hybrid ARIMA, SSA and Linear Regression) on overall city training set (from 01-2007 to 12-2015) using testing set (from 01-2016 to 10-2017) depending on MPE measure to detect the smallest MPE percentage error. After selecting the best algorithms, we then forecast water consumption and water revenue in advance 62-month (~5 years) for three levels, (1) the overall city, (2) sub-areas separately (20 sub-area), (3) classes and categories in KhanYounis city (13 class). Finally, we forecasted some special cases in both regions and class category in advance ~5 years.

1.7. Thesis Outlines

The following chapters including Chapter 2 Literature Review, descripting the used algorithms used in our forecasting tasks and presenting their theoretical and scientific ideas. Chapter 3 Related Work, summarizing some recent related works belong to our research field and techniques. Chapter 4 Methodology and Model Development, description our methodology and methods we use in evaluating and processing our forecasting process. Chapter 5 Experimental Results and Discussion, presenting all our results, graphs and tables belongs to evaluation process and forecasting next 5 years. Chapter 6 Conclusions, presenting the conclusion of our research and showing up some recommendations for future work.

Chapter 2
Literature Review

Chapter 2

Literature Review

Over the last few decades researches and scientific papers have shown an interest in forecasting field and attempted to quantify and justify the wide variety of techniques used. In this chapter we talk about the main of data mining technologies, then we address time series and forecasting methods and algorithms we used in the thesis which are ARIMA, Hybrid ARIMA, SSA and Linear Regression.

2.1. KhanYounis City

Gaza Strip municipalities are working on establishing and collecting spatial data belong to their services on each city and analysis the water data to benefit from in planning water system projects and decision-making process. For example, KhanYounis municipality has central database and large datasets belong to its service on the city such as crafts information, rentals information and water service data and information. These available resources give us the advantage to advance our work in applying forecasting techniques and algorithms to predict the future.

KhanYounis municipality as a case study of our research, is well prepared municipality technically and graphically and has a large dataset belongs to all its services on the city. Daily and weekly there are thousands of records and data are collected from all distributed stations (for example water distribution pumps) belongs to KhanYounis municipality especially to water service. So, these data scientifically should be used in planning for water demand in the future.

In our research, we collected some related data to water demands, consumptions and revenues used in KhanYounis municipality KHM applications. These data were prepared for the forecasting algorithms to present useful results and show it to managers and water service specialists depending on the estimation of water consumption and new expected water revenues.

KhanYounis city is the second biggest city in Gaza Strip and one of the big cities in Palestine. It's area size equals 54.5 Km^2. It considered the biggest city in area in Gaza strip and the second city in population. It consists from 20 sub-area or zone. The Coastal Aquifer and groundwater are the main sources of water for the Palestinians in the Gaza Strip, providing more than 90% of fresh water supply for various purposes. While 4% of purchased water in the Gaza Strip imported from the Mekorot systems. In 2008, a study by Palestinian Central Bureau of Statistics (PCBS) shows the population of KhanYounis city and the increase rate for the following years until year 2025 as show in ***Table (2-1***.

Table (2-1) Population increase rate in KhanYounis city from 2007 to 2025 (PCBS, 2016)

Year	Population	Population increase rate (%)
2007	180,342	0
2008	187,195	3.8
2010	201,692	7.7
2016	241,870	19.9
2018	266,614	10.2
2020	284,501	6.7
2025	326,625	14.8

The results of increase rate in population from the year 2018 to 2025 according to the studies from engineering and management consulting center and center for engineering and planning 'CEP & EMCC', which depended on the average population increase which announced by (PCBS, 2016).

2.2. Data Mining

Data Mining is the science of discovering useful hidden relationships from large datasets or databases. It is a discipline lying at the intersection of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other areas(Shen, Tong, & Deng, 2007). The modern technologies and high capabilities of computers and electronic devices have made gathering data an almost effortless task. However, the captured data need to be converted into actionable information from recorded data to become useful. Traditionally, analysts have performed the task of extracting useful information from the recorded data, but the increasing volume of data in modern business and science calls for computer-based approaches. As datasets have grown and complexity, there has been an inevitable shift away from direct hands-on data analysis toward indirect, automatic data analysis using more complex and sophisticated tools. Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery, from data (Humaid, 2012).

Data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Shen et al., 2007). Data mining techniques play a big role in analyzing large databases with millions of records to achieve new unknown patterns. These patterns help with study the customer transactions behavior in many scientific researches as in mine. When we get the suspicious customer, patterns using data mining techniques, then we can narrow the circle of investigation to get fraudulent issues very fast according to the manual way that could be impossible.

2.2.1. Data mining application

One of data mining goals is to find actionable information that can be utilized in a concrete way to improve profitability. Some of the earliest applications were in retailing, especially in the form of market basket analysis. Figure 2-1 shows the general application areas in representative way rather than comprehensive (Delen & Olson, 2008).

Application area	Applications	Specifics
Retailing	Affinity positioning, Cross-selling	Position products effectively Find more products for customers
Banking	Customer relationship management	Identify customer value, Develop programs to maximize revenue
Credit Card Management	Lift Churn	Identify effective market segments Identify likely customer turnover
Insurance	Fraud detection	Identify claims meriting investigation
Telecommunications	Churn	Identify likely customer turnover
Telemarketing	On-line information	Aid telemarketers with easy data access
Human Resource Management	Churn	Identify potential employee turnover

Figure (2-1) Data mining business areas (Delen & Olson, 2008).

2.2.2. Data Mining Methodology

Knowledge discovery as a process is depicted in **Figure 2.2** and consists of an iterative sequence of the following steps (Han, Pei, & Kamber, 2011):

1. Data cleaning (to remove noise and inconsistent data).
2. Data integration (where multiple data sources may be combined).
3. Data selection (where data relevant to the analysis task are retrieved from the database).
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance). Data mining (an essential process where intelligent methods are applied to extract data patterns).
5. Pattern evaluation (to identify the truly interesting patterns representing knowledge Based on some interesting measures).
6. Knowledge presentation by visualizing mined knowledge to the user.

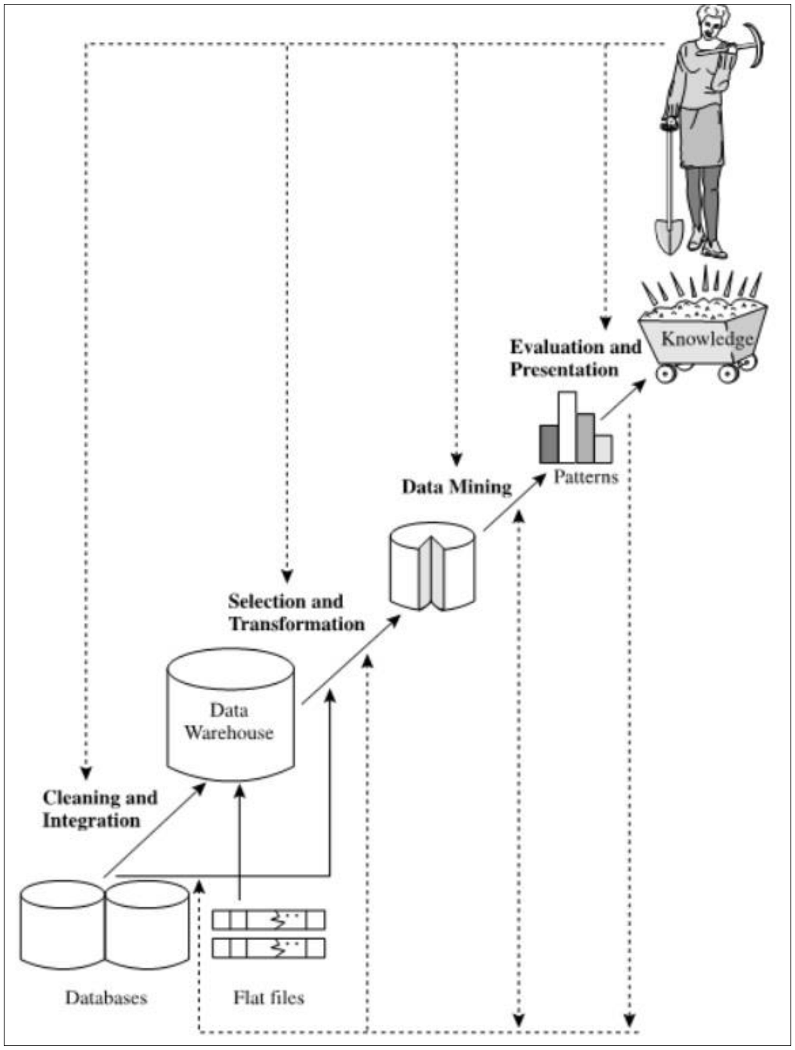

Figure (2-2) Data mining process step (Han et al., 2011)

2.3. Time Series Analysis and Forecasting

A time series is a series of discovered observations on an attribute measured at successive points in time or over successive periods of time. The measurements may be taken periodically (hour, day, week, month, or year) or at any other regular interval. The trends or patterns of the data are important factor in understanding the behavior of time series data in the past. If such behavior can be expected to continue in the future, we can use the past pattern to guide us in selecting an appropriate forecasting method. To identify the underlying pattern in the data, a useful first step is to construct a time series plot. A time series plot is a graphical presentation of the discovered relationship between time and the time series attributes; time presents the horizontal axis and the time series values presents the vertical axis. There are many forecasting methods available, these methods can generally be divided into three groups (Nosedal, 2011):

1. Judgmental methods.

Expert knowledge observations depends at the core of judgmental forecasting, it depends on the power of integration between knowledge and its forecasting (Alvarado-Valencia, Barrero, Önkal, & Dennerlein, 2017).

2. Extrapolation (or Time Series) methods.

Extrapolation methods are quantitative methods that use past data of a time series variable-and nothing else, except possibly time itself-to forecast future values of the variable. The idea is that we can use past movements of a variable, such as some company sales to forecast its future values(Nosedal, 2011).

3. Econometric (or causal) methods.

Econometric models discover the relationship between such attributes and influencing factors, and econometric models need a large amount of observations to achieve a higher forecasting accuracy. In addition, Some studies have proposed a hybrid forecasting by combining econometric and data mining techniques (Li, Pan, Law, & Huang, 2017).

Allan Steel (Steel, 2014) described the main components of a time series as:

1. Trend - the general direction of the series, is it upward or downward by time as shown in **Figure (*2-3*)**.

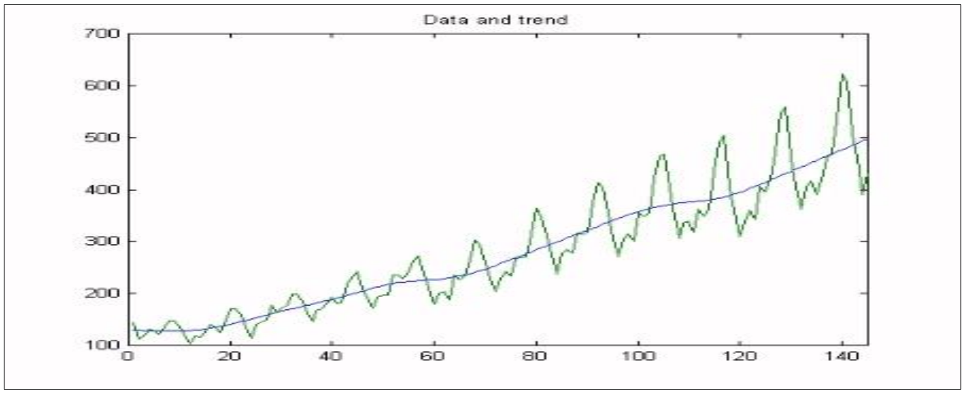

Figure (2-3) Trend time series presentation.

2. Seasonality - regular repeated patterns in the time series that is caused by repeated events as shown in **Figure** ***(2-4***, for example a spike in sales during the Christmas period (So & Chung, 2014).

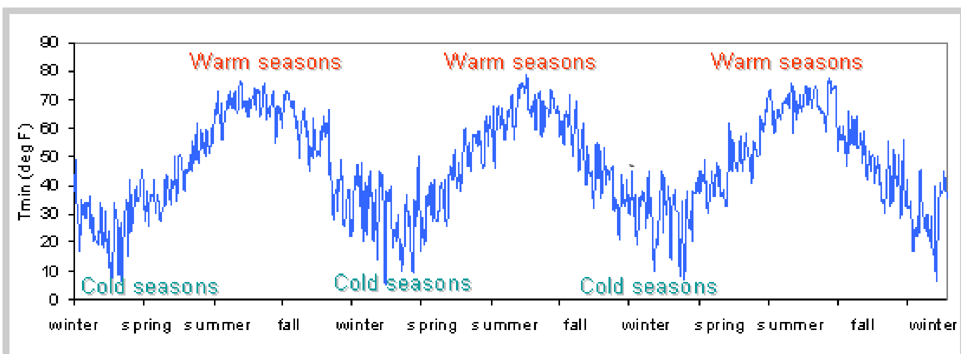

Figure (2-4) Seasonality time series presentation.

3. Random component - additional fluctuations in the series that may be attributed to noise or other random events as shown in **Figure (*2-5***.

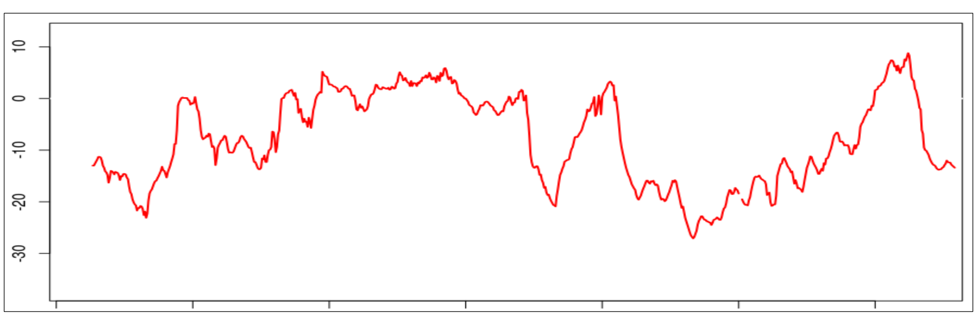

Figure (2-5) Random time series presentation.

Also, Steel (2014) explained the three general types of time series which are stationary, additive and multiplicative. Usually, stationary time series are repetitive, it shows constant auto-correlation and are considered the easiest type to model. Stationary series have constant amplitude without a trend element:

stationary time series = seasonality (and/or) noise **(Equation 2-1)**

The second type of time series is the additive type. In this type all three components of the series are present, trend, seasonality and noise. The distinguishing feature here is the amplitude of the seasonal component in that it is quite regular being static over time. This time series is trending upwards overall but there is a clear repetitive pattern of peaks and troughs caused by the seasonality, with the heights of the peaks all being similar. We can consider an additive time series as:

additive time series = trend + seasonality + noise **(Equation 2-2)**

The third type of time series is multiplicative. This is like the additive version except the amplitude of the seasonality increases over time. It can be considered as:

*multiplicative time series = trend * seasonality * noise* **(Equation 2-3)**

Financial time series can be considered as containing all three elements of a time series. They can show properties of a stationary time series when they are range bound and only move between two values. At other times, markets trend strongly consistently, making new highs or lows and exhibit properties of an additive and occasionally a multiplicative series (Steel, 2014).

2.4. Auto-Regressive Integrated Moving Average (ARIMA)

In reality of course, many time series data sets have trend, and in the world of financial data this is also true. To account for pattern or trend in a time series it is usually first transformed into a stationary data, modeling is then performed on this manipulated data after which it is returned to its original state. In effect the trend aspect is removed, modeling is done, then the trend component is added back into the data.

One such method for removing trend is differencing (Mills, 2011). Differencing is the technique of replacing the actual values of the observations with the values of the differences between them. This is represented as:

$$Diff_{1t} = R_t - R_{(t-1)} \quad \textbf{(Equation 2-4)}$$

Differencing is the same as calculating the derivative of the series, thus a time series that has under gone differencing is considered 'integrated'. If taking this so-called first difference doesn't remove the trend one can go further and use the second difference:

$$Diff_{2t} = (R_t - R_{t-1}) - (R_{t-1} - R_{t-2}) \quad \textbf{(Equation 2-5)}$$

Addition of an integration step to the ARMA model results in an auto regressive integrated moving average (ARIMA) model, with the general formula:

$$\begin{aligned}
&r(t) = c + \\
&b_1 * R_{t-1} + b2 * R_{t-2} \ldots b_p * R_{t-p} + \\
&a_1 * ma_{t-1} + a2 * ma_{t-2} \ldots a_q * ma_{t-q} \\
&d_1 * diff_{t-1} + d2 * diff_{t-2} \ldots d_d * diff_{t-d} \\
&+ err \quad \textbf{(Equation 2-6)}
\end{aligned}$$

where:

c is the intercept, which is often zero and the mean of the time series.

$b_1 - b_p$ are the independent variables, the previous values in the auto-regression term.

$a_1 - a_p$ are parameters of the moving average model.

d_{1-p} are the parameters of the differencing term.

An ARIMA model typically consists of three parts (1) auto regression AR (order 'p' the number of terms used in the auto-regression), (2) moving average MA (order 'q' the number of terms used in the moving average) and (3) differencing to strip off the integration of the series (order 'd' the number of differencing terms) and then form ARIMA (p, d, q). Some of the limitation of ARIMA is, it handles time series data in linear form and must be stationary and the accuracy of ARIMA forecasting model is significantly affected by the noise of time-series dataset.

2.5. Neural Networks Model (NNs)

NN models defined as multivariate, nonparametric and nonlinear models which can well reveal the correlation of nonlinear time series in delay state space. Neural networks models applied widely to fields such as finance, data mining, physics and medicine. In finance, neural networks have been used for stock market prediction, credit rating, bankruptcy prediction and economic indicator forecasts. In medicine, neural networks used in medical diagnosis, detection and evaluation of medical conditions and cost estimation. Furthermore, neural networks purposes are to predict, classify, discovery knowledge and time series analysis. **Figure 2.6** shows four main classes of NN models, Feed forward NN, Recurrent NN, Stochastic NN and Modular NN.

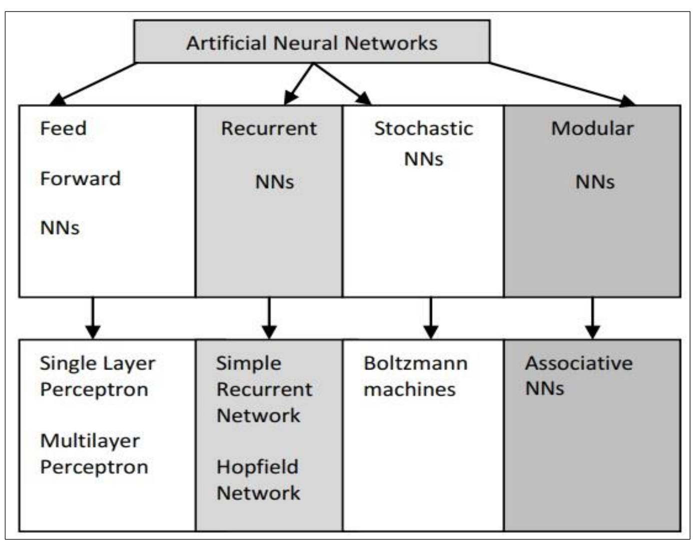

Figure (2-6) A taxonomy of neural network architectures (Sibanda & Pretorius, 2012).

2.5.1. Feed-forward Neural Networks

Feed-forward is the simplest type of ANNs. In this type of NNs the data or information goes in single direction, starting from input nodes, going through hidden nodes and finally to output nodes and there are no cycles or loops in the network. Examples of this category are single layer perceptron (SLP) and multi-layer perceptron (MLP).

2.5.2. Recurrent Networks

Recurrent networks are bi-directional processes. It spreads data linearly from input to output and spreads data from later processing steps to earlier steps (Bitzer & Kiebel, 2012). Examples of this category are Simple Recurrent Network (SRN) and Hopfield Network.

2.5.3. Stochastic Neural Networks

A stochastic neural network introduces random variations into the network (Su, Li, Wang, & Ding, 2011). Boltzmann machines is an example of a stochastic neural network.

2.5.4. Modular Neural Networks

A modular neural network is a neural network characterized by a series of independent neural networks moderated by some intermediary (Pandey, Jain, Kothari, & Grover, 2012). Each independent neural network serves as a module and operates on separate inputs to accomplish some subtask of the task the network intends to perform (Azam, 2000). The intermediary takes the outputs of each module and processes them to produce the output of the network. The intermediary accepts the modules outputs but not respond to, nor otherwise signal, the modules. The modules do not interact with each other. One of the benefits of a modular neural network is the ability to reduce a large neural network to smaller, more manageable components(Azam, 2000). Examples of modular neural networks include committee of machines (CoM) and associative neural networks (ASNN).

2.6. Hybrid ARIMA

Auto-regressive integrated moving average models are important modeling algorithms for time series data. However, these techniques have limitations that have detracted from their popularity, which assumes a linear relationship, stationary data and it needs a lot of data to produce accurate results. To pass these issues a suggested hybrid solution has been proposed in which ARIMA models are combined with other techniques, often non-linear prediction algorithms.

Hybrid ARIMA is one of the best and famous hybrid models in operating forecasting tasks. The motivation of the hybrid ARIMA comes from the following perspectives (Khashei & Bijari, 2011). (1), it is often difficult in practice to determine whether water consumption series is linear or nonlinear nature or whether such method is more accurate than the other in forecasting. That is why it is difficult to choose the most appropriate technique for a certain problem. Also, many potential factors such as sampling variation, model uncertainty, and structure change may affect in choosing the best model for future use. By combining different methods, selecting the best algorithm is eased with some efforts. (2), real-world time series contains both nonlinear and linear patterns are rarely found linear or nonlinear. If this is the case, then neither ARIMA nor ANNs can be generalized the best method in modeling and forecasting time series since ARIMA model cannot deal with nonlinear patterns while the NN model alone is not able to handle both linear and nonlinear patterns equally well. Hence, by combining ARIMA with ANN models, complex autocorrelation structures in the data can be modeled more accurately. (3), it is almost universally known that no such algorithm is best in every situation. For example, in the literature of time series forecasting with neural networks, most studies use the ARIMA models as the benchmark to test the effectiveness of the ANN model with mixed results.

Many studies including several large-scale forecasting competitions suggest that forecasting performance can be enhanced by combining several different models. Therefore, combining different models can increase the chance to discover more hidden patterns and relationships in the data and improve forecasting performance.

2.7. Linear Regression

Linear regression is a statistical method to model the relationship between two variables by fitting a linear equation to observed data. One variable is an explanatory variable, and the other is a dependent variable. Linear regression method can be used for forecasting under the assumption of continuing the correlation between the variables in the future. A linear regression model is as follows:

$$Yi = \alpha + \beta xi$$ **(Equation 2-7)**

where α and β are coefficients.

There are several types of linear regression analyses available to researchers (Solutions, 2013):

- Simple linear regression: 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous).
- Multiple linear regression: 1 dependent variable (interval or ratio), 2+ independent variables (interval or ratio or dichotomous).
- Logistic regression: 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous).
- Ordinal regression: 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous).
- Multinomial regression: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous).

Multiple linear regression analysis model uses the interactions between the variables to produce more accurate forecasts. In linear regression the deviations between the real values and the estimated values are caused by the observation errors (Heshmaty & Kandel, 1985). These deviations depend on the fuzziness of the system parameters (Tanaka & Asai, 1984). In decision-making tasks, it is not preferred to depend on a Boolean decision, where the answer is "yes" or "no", because if the answer is wrong, the decision will be completely error or if it is true the answer is 100% true, and this issue has a major effect in the decision-making model (Week, 1982). Therefore, using fuzzy techniques reduces the heavy impact of errors and as a result yields better forecasts for decision-making.

More data is needed to discover accurate relationships between the dependent and the independent variables, because the more available data the more training and learning is gained by algorithms. Thus, there is a trade-off between the accuracy and the cost in regression (Bails & Peppers, 1993). There are some assumptions about the relationships, the model, and the data being used in models. (1) first assumption is that the relationship between the variables effective in the model does not change over the range of values of the independent variables; thus, longer time periods make better relationships. (2) second assumption is that the relationship found from discrete observations among the variables does exist between the discrete data points as well, i.e. the relationship is continuous over the data range. In linear regression, it is also assumed that the deviations of historical data from the measured relationships of variables are randomly distributed. The final assumption is that the dependent variable is normally distributed about the regression line.

The most common use of regression in business is to predict events that have yet to occur. Demand analysis, for example, predicts how many unit's consumers will purchase. Many other key parameters other than demand are dependent variables in regression models, however. Predicting the number of shoppers who will pass in front of a billboard or the number of viewers who will watch the Super Bowl may help management assess what to pay for an advertisement. Insurance companies heavily rely on regression analysis to estimate how many policy holders will be involved in accidents or be victims of burglaries (Ozyasar, 2011).

Another key use of regression models is the optimization of business processes. A factory manager might, for example, build a model to understand the relationship between oven temperature and the shelf life of the cookies baked in those ovens. A company operating a call center may wish to know the relationship between wait times of callers and number of complaints. A fundamental driver of enhanced productivity in business and rapid economic advancement around the globe during the 20th century was the frequent use of statistical tools in manufacturing as well as service industries. Today, managers considers regression an indispensable tool (Ozyasar, 2011).

2.8. Singular Spectrum Analysis

Singular Spectrum Analysis (SSA) is a relatively new and powerful nonparametric tool for analyzing and forecasting economic data. SSA can decompose the main time series into independent components like trends, oscillatory manner and noise.

The birth of SSA is usually associated with the publication of papers by Broomhead (Broomhead & King, 1986) while the ideas of SSA were independently developed in Russia (St. Petersburg, Moscow) and in several groups in the UK and USA. At present, the papers dealing with the methodological aspects and the applications of SSA number several hundred.

The fact that the original time series must satisfy a linear recurrent formula (LRFs) is an important property of the SSA decomposition. Generally, the SSA method should be applied to time series governed by linear recurrent formulae to forecast the new data points. There are two methods to build confidence intervals based on the SSA technique: the empirical method and the bootstrap method. The empirical confidence intervals are constructed for the entire series, which is assumed to have the same structure in the future. Bootstrap confidence intervals are built for the continuation of the signal which are the main components of the entire series (Golyandina, Nekrutkin, & Solntsev, 2001).

The possible business fields of SSA are distinguish over areas of mathematics, physics, economics, meteorology, oceanology, social science and market research. Any seemingly complex series with a potential structure could provide another example of a successful application of SSA (Hassani, 2007).

SSA is a very useful tool which can be used for solving the following problems(Hassani, 2007): (1) finding trends of different resolution. (2) smoothing. (3) extraction of seasonality components. (4) simultaneous extraction of cycles with small and large periods. (5) extraction of periodicities with varying amplitudes. (6) simultaneous extraction of complex trends and periodicities. (7) finding structure in short time series. (8) change-point detection.

Five main steps in SSA-forecasting: (I) Embedding the sampled time series in a vector space of dimension M; (II) Computing the M x M lag-covariance matrix CD of the data; (III) Diagonalizing CD; (IV) Recovering the time series; and (V) forecasting. Singular spectrum analysis (SSA) is a data adaptive technique (Elsner & Tsonis, 2013) and therefore it has potential to capture such variations and can prove extremely promising for short-term forecasts.

2.9. Summary

From previous presented information for ARIMA, Hybrid ARIMA, Linear regression and SSA, we noticed that these algorithms are used widely in time series forecasting tasks and many researches and papers enriched their performance and business usage in real life applications. Next chapter lists the most related papers and thesis belongs to our topic and how researches enhanced and improved forecasting algorithms performance.

Chapter 3
Related Work

Chapter 3
Related Work

Forecasting take an important role in decision making. Many researches and scientists are conducting, building tools and algorithms to improve the quality of forecasting tasks. Also, many papers and thesis were published in this field. We display related literature reviews of three level of forecasting time series data. First level is short term, which forecast data up to one year only. Second level is medium term which forecast data from one year to 10 years and this is our concentration field. Third level is long term which forecast data more than 10 years. Then, we display related works and researches in other domains and fields using similar forecasting methods. Then we listed some researches in other domains for example electricity consumption and crude palm oil (CPO) price. etc.

3.1. Short-term forecasting

One of the most recent researches talking about short-term water demand forecasting model using a moving window on previously observed data presented by (Pacchin, Alvisi, & Franchini, 2017). They used 24-h time window using a pair of coefficients whose value is updated at every forecasting stage. The first coefficients represent the ratio between the average of water consumption after 24 h that follow the time the forecast is made and the average water consumption before the 24 h that precede it. While the second coefficient represents the relationship between the average water consumption in a falling hour within the 24 h forecasting period and the average water consumption over that period. Also, authors take in account the holidays and special occasions in building their model and they applied the proposed model to the real-life case of Castelfranco Emilia (Italy). Finally, they concluded that the forecasting accuracy of the proposed model may not be as good when evaluated in relation to certain public holidays falling in the year; they assume, if data was available during the same holidays in several previous years about water demand, the application of a variant of the model will improve the forecasting accuracy.

Chen & Boccelli (2014) developed an integrated Time Series Forecasting Framework (TSFF) to predict hourly/quarter-hourly demands in real-world, real-time scenarios. The proposed framework consists of two models, the first is a fixed seasonal auto-regressive (FSAR) model, the second is an adaptive seasonal auto-regressive (ASAR) model. TSFF approach can be integrated with SCADA water systems. They evaluated the TSFF model structure using 24 hours and 7-day periodicities and applied to additional demand time series to evaluate the performance of the forecasting results by comparing with water consumption from different distribution systems.

Wen et. al. (2013) proposed a GUI software application based on Java programming language, which designed to connect to specific database, then execute and run data mining algorithms, after that the application displays visual output data and charts. They analyzed the different prediction models and designed the water consumption system that has two functions, the first is the analysis of possible correlations between the water consumption and nature of the industry and the second is predicting future water consumption. **Figure 3.1** presents their experimental results showing the curve of observed and predicted water consumption which are almost having the same trend. They intend to develop their software to enhance dynamic data services and real-time Internet-of-Things (IoT).

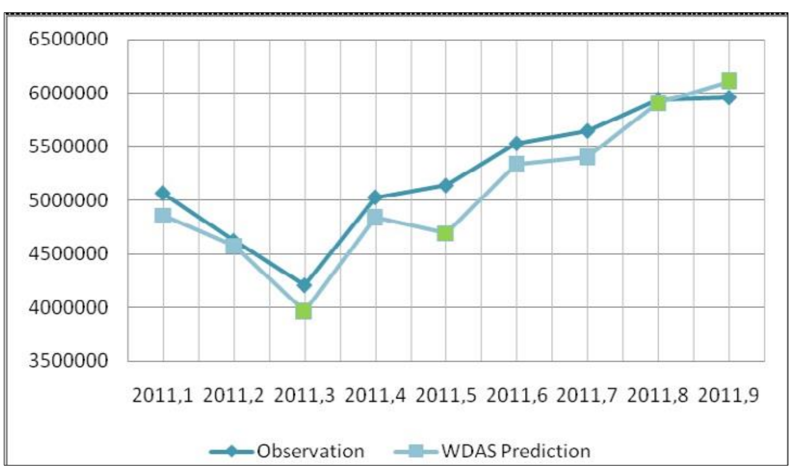

Figure (3-1) WDAS Predictions compare with Observations (Wen et al., 2013).

Shabani et. al (2017) used Support Vector Machines model ‘SVM’ using polynomial kernel function to forecast short term ‘monthly’ water demand of City of Kelowna (CKD), Canada. Their main goal is to assess the use of phase space reconstruction prior to design model’s input variables combinations. They claim that results of their study as shown in Figure (***3-2*** proved optimum lag time of the input variables can significantly enhance SVM models.

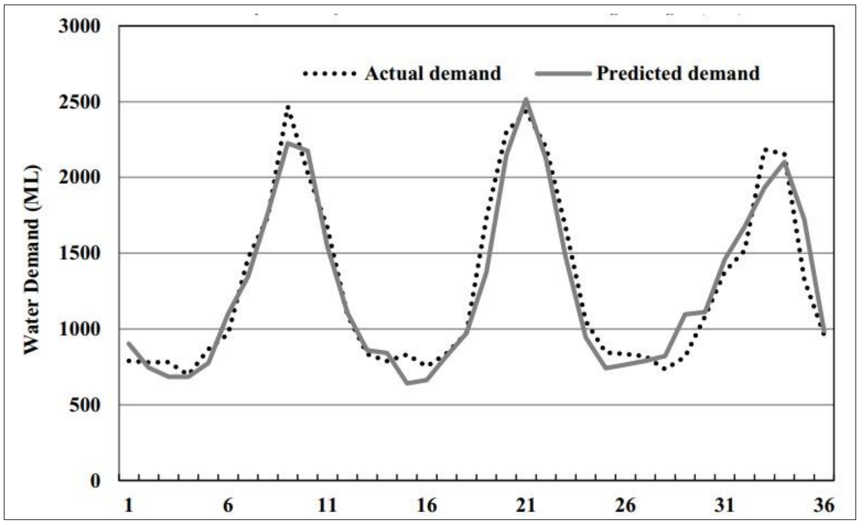

Figure (3-2) Actual demand Vs Predicted demand (Shabani et al., 2017)

Ceballos and Walke (2016) developed linear transfer function model that generates monthly frequency out-of-sample simulations of water consumption and demands for periods when actual consumption is known. They observed that changes in economic conditions and climatic may affect consumption per customer more rapidly than changes in water rates. They found that the LTF parameter estimates indicate that a 10% rate increase will lead to a 3.2% decline in water demand after a lag of three months. They used U-statistics and a directional accuracy test to evaluate LTF forecasts against three alternative benchmark forecasts.

Bougadis et. al (2005) performed analysis to quantify the relationship of climatological variables with peak demand. They investigated the relative performance of regression, artificial neural network (ANN) models and time series analysis for short-term forecasting over data of the city Ottawa, Ontario, Canada. Linear and multiple linear regression models were hypothesized using weekly maximum temperature, weekly rainfall amounts, the occurrence of rainfall in a week, and the peak water demand. They observed that there is no predominant single periodicity component in the demand series. They observed that ANN models outperformed the regression. Finally, they found that water demand is more correlated with the rainfall amount than the occurrence of rainfall.

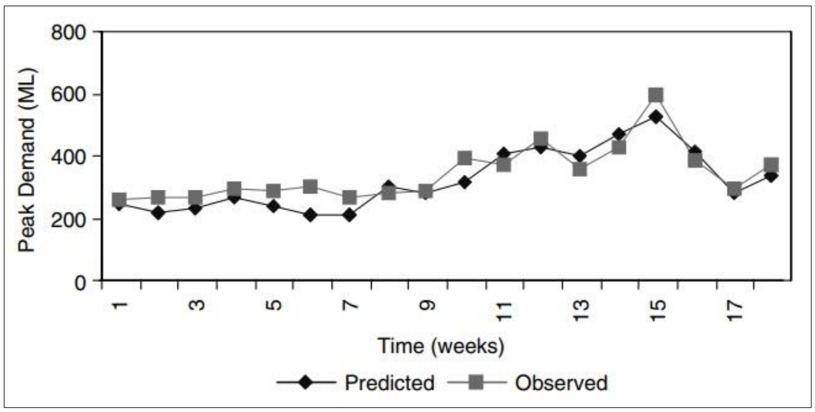

Figure (3-3) Observed and predicted water demand from the ANN (Bougadis et al., 2005)

Bakker et. al (2014) have studied the weather attribute on the performance of water demand forecasting models. The performance of the models was analyzed with and without using weather attribute, to assess the possible performance improvement due to using weather attribute. They observed that when using weather attribute on models the largest forecasting errors can be reduced by 11% and the average errors by 7%. They claim that this reduction is important for the application of the forecasting model for the control of water supply systems.

3.2. Medium-term water demand forecasting

Many authors researched on water demand forecasting and its effects on urban water systems design, Donkor et. al (2012) reviewed most of water demand forecasting papers published between 2000 and 2010 to identify the useful methods and model for specific water utility decision making process.

This general literature review listed general topics such as basis of water demand forecasting, forecast variables and determinants and measures of forecast error. Then, authors listed some methods and models were discovered from published researches at the interval mentioned above for example Forecasting with Qualitative Methods, and Unit Water Demand Analysis, Forecasting via Univariate Time Series Analysis, Moving Average and Exponential Smoothing Model, Stochastic Process Models, forecasting with time series regression models, forecasting with Scenario-based Approaches and Decision Support Systems, Forecasting with Artificial Neural Networks and Forecasting with Composite Models.

Then authors mentioned some problems and recommendations may face water demand forecasting. These points talk about the nature of data collected and the variable and determinants used in forecasting process. Also choosing an appropriate forecasting method according to data and purpose of research can effect on the results and the uncertainty in input parameters affects demand estimates. They studied the relationship among planning level, water utility decision problems and forecast attributes.

Ajbar & Ali (2015) developed a neural network model for forecasting the monthly and annual water demand for Mecca city, Saudi Arabia. They used historical data of water production and estimated visitors' distribution to calibrate a neural network model for water demand forecast. They observed that NN prediction outperforms that of a regular econometric model. They developed two models for the forecast of monthly and annual water demand in the city. They designed the model to combine the impact of monthly-varying and yearly-varying explanatory variables. The neural network model was found to be superior to the econometric model in the sense of capturing the water consumption dynamics both in the short and long terms. Also, the NN model was useful in investigating the influence of the changes occur in the number of visitors specially in AL Haj season on the overall water demand of Mecca city.

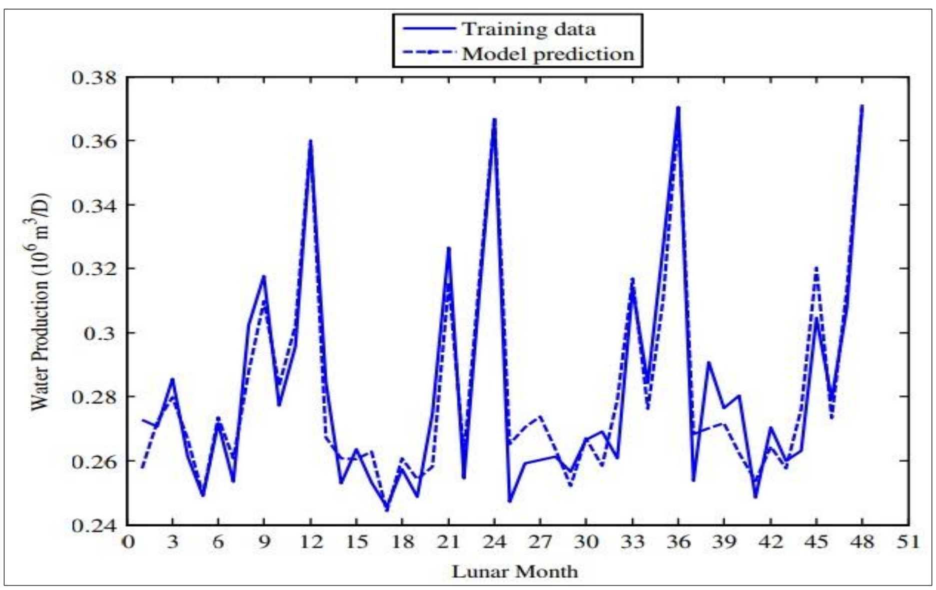

Figure (3-4) Neural model training result (Ajbar & Ali, 2015).

3.3. Long-term water demand forecasting

Rinaudo (2015) wrote a chapter that reviewed long term water demand forecasting methodologies. The literature review is enriched by economists, engineers and system modelers. The author illustrated how different tools can be used depending on the regulatory context, the water scarcity level, the geographic scale at which they are deployed and the technical background of water utilities and their consultants. His chapter represented three main challenges, the integration of land use planning with demand forecasting, accounting for climate change, and dealing with forecast uncertainty.

Some of the typology of water demand forecasting methodologies are: Temporal extrapolation models, Models based on 'unit water demand', Multivariate statistical models, Micro-component modeling, and Estimation based on projections for urbanization and land use.

The author concluded after studying the three examples that there is no single 'off the shelf' forecasting tool that can be applied universally. Different modeling tools can be used depending on the nature of the data, regulatory context, the water scarcity level, the geographic scale at which they are deployed, and the technical background of water utilities and their consultants.

Liu et. al (2015) also used data mining techniques to identifying determinants of urban water. Their approach is based on a genetic programming (GP) data mining technique, which automatically optimizes the structure of the function and parameters simultaneously. With historical urban water use as the target, the GP model identified the most relevant factors for 47 cities in northern China. They assume that housing area, water price and rainfall are affecting the forecasting model. They claim that the new model in their study could be helpful for urban water management, especially for cities that experience water scarcity.

Polebitski and Palmer (2009) developed regression-based(pooled, fixed, and random effects) water demand models capable of forecasting single-family residential water demands within individual census tracts at a bimonthly time-step. They used 12 years of deferent types of water consumption data associated with over 100 unique census tracts in Washington. They claim that the coefficient estimates developed in their research are appropriate for use in spatially disaggregate urban simulation models.

They noticed that water demand within a small geographic region was found to be highly variable. Using census tracts as the forecasting unit can provide utilities with greater insights on how water is consumed spatially and how specific demographics impact consumption across space and over time. They claim that the proposed approaches allow examination of spatially distributed demands within systems, identify the value of targeted conservation and infrastructure development, and improve understanding of the variables impacting demand in heterogeneous areas.

3.4. Researches in other domains

RISH (2015) forecasted the electricity consumption in the KhanYounis province from 2000 to 2010 using exponential smoothing and Box-Jenkins ARIMA methods. She compared between these two models to detect which model is more accurate to forecast electricity consumption of KhanYounis province. She concluded that exponential smoothing has MAPE 5.92% while Box-Jenkins has MAPE 6.24%.

Yu et. al (2017) used ARIMA for short term forecasting of water level in the Yangtze River to help prevent water floods. They trained the model over dataset of four years (2012-2015) and used data of 2016 to evaluate the performance of the model. They observed that the forecasting horizon influences the model performance and noted that the model accuracy decreases as the forecasting period is extended. The obtained MAPE of 3-day forecasting horizon is four times larger than 1-day forecasting horizon at Hankou station.

Kasturi Kanchymalay et al (2017) studied the relation between crude palm oil (CPO) price, selected some vegetable oil prices, crude oil and the monthly exchange rate. They noticed that the prediction results exhibit that the support vector regression had the best accuracy (MPE= 7.8%) compared to multi-layer perceptron (MPE= 19.9%) and Holt Winter exponential smoothing methods (MPE= 30.9%).

Wongsathan (2016) used a hybrid model from ARIMA and NN to forecast pollution estimation in Chiang Mai city. Their results showed clearly that hybrid model performance overcome single NNs and ARIMA model by average 65% and 50% respectively.

Khashei et al. (2009) report on the use of this combination to predict the future price movement in gold and US dollar/Iran rails financial markets. The workers reported favorable results in comparison to the techniques alone and suggest the method as having potential for accurate predictions of non-linear time series data.

In a similar study Zhang (2003) applied a combination of ARIMA and ANN to various data sets including the British pound/US dollar exchange rate. They observe that in the literature in general these two popular techniques are frequently compared in terms of predictive power with the reported results non-conclusive. Results from the three data sets modelled show that the combination of the two methods outperform the individual ones when the mean squared error (MSE) and mean absolute deviation (MAD) are used as the measure of forecasting accuracy.

Also, Wang and Meng (2012) has studied similar issue belongs to our research. They forecasted the energy consumption in Hebei province, China using a combination from ARIMA and NN. They observed that he hybrid model improved the accuracy of forecasting energy consumption with comparison to models used separately.

Fatima and Hussain (2008) studied the impact of hybrid approach in modelling short term predictions for the Karachi Stock Exchange index (KSE100). The authors reported comparison results for ANN, ARIMA and a hybrid of ARIMA/ANN. The hybrid model out-performed the separated ARIMA and ANN models. Usually, financial systems and data over time factor may have the nature of being non-linear, volatility patterns, political instabilities, general rumor and monetary policies, so one single algorithm may handle one aspect form the previous complex conditions affecting the financial forecasting task, but hybrid model of multi algorithm maybe more successful and accurate in forecasting the future patterns.

Other researches combined ARIMA with wavelet analysis in forecasting the price of electricity. Tan et al. (2010) reported results showing the enhanced predictive of the wavelet ARIMA hybrid model compared to GARCH and ARIMA models used individually. Also, Nury et al. (2015) studied the combination of wavelet/ARIMA and compared with wavelet/NN in temperature forecasting. Their data divided into training dataset (1957–2000) to build the models and a testing dataset (2001–2012) to evaluate their performance. Their results showed that wavelet/ARIMA is more effective than wavelet-NN.

Pai and Lin (2005) proposed a hybrid model from ARIMA and Support vector machine (SVM) to predict the prices of fifty stocks. The combination of ARIMA/SVM handles non-linear and linear time series data. Results from the work show that the hybrid method out-performs the ARIMA and SVM methods individually.

Rout et al. (2014) overcome the limitation of ARIMA by proposing a combination model from ARIMA and differential evolution (DE). They also compared accuracy of ARIMA/DE with other combinations like ARIMA/Particle Swarm Optimization (PSO), ARIMA/Cat Swarm Optimization (CSO), ARIMA/Bacterial Foraging Optimization (BFO) and ARIMA/Forward backward least mean square (FBLMS). They concluded that the performance of ARIMA/DE out-performed the other suggested combinations in predicting short and long terms.

Linear regression is used in many researches for forecasting time series data. Behrooz and Kandel (1985) studied the computer sales in the U.S. using fuzzy linear regression. The concluded to fuzzy linear regression can be applied to any model that uses abstract non-fuzzy linear regression by fuzzifying the linear regression model using a fuzzy linear function along with the fuzzy parameters of the type of triangular membership functions or any other type of membership functions.

Song et al. (2005) proposed a fuzzy linear regression method to forecast a short term 24 hourly loads of power in holidays for the years of 1996-1997, the suggested algorithm had average maximum percentage error equals 3.57%.

Also, Bianco et al. (2009) proposed a long-term electricity consumption forecasting model. They made a comparison with national forecasts, based on complex econometric models, such as Markal-Time, was performed, showing that the developed regressions are congruent with the official projections, with deviations of ±1% for the best case and ±11% for the worst. These deviations are to be considered acceptable in relation to the time span considered. They concluded to not consider the electricity price as explaining variable in forecasting models for Italian electricity consumption. They also expect that electricity consumption will increase in the following years in Italy by average rate 2% per year.

Mohamed and Chowdhury (2015) proposed a short-term multiple linear regression analysis model to generate forecasts of solar energy. They found that the forecasting hours with clear sky, the model's performance would be better for near forecasting horizon than farther horizon, but this is affected by cloudy hours, thereby the entire performance of the model. Also, they observed that more historical data will improve the model performance.

Iqelan (2017) used Singular Spectrum Analysis SSA in forecasting the monthly electricity consumption of the Middle Province in Gaza Strip in Palestine.

His data set is from November 2005 to December 2015 as a training sample and the remaining 12 observations from January 2016 to December 2016 are used as a testing data set to evaluate the electricity consumption forecasts. He compared SSA with ARIMA and exponential smoothing state space (ETS). He concluded to SSA (MAPE 9.38%) outperformed both ARIMA (MAPE 14.99%) and ETS (MAPE 15.63%).

Kumar and Jain (2010) used SSA to forecast gas consumption in India. The mean absolute percentage error (MAPE) for gas consumption forecasting equals 3.4%. While Hassani et al. (2015) studied the advantages of using Singular Spectrum Analysis (SSA) for forecasting tourist arrivals into the Unites States over the period 1996 to 2012. They compared the performance of SSA with ARIMA, exponential smoothing and ANN. They concluded that SSA outperformed the alternative methods. In terms of the MAPE, SSA reports the lowest average MAPE at 8% with comparison to the ARIMA 14%, ETS 17% and NN 22% MAPEs.

3.5. Conclusion

No one applied forecasting on water in any Palestinian area and most of previous related works studied and analyzed the water demand, but none have forecasted the water revenue and discussed the relation between water consumption and water revenue, is it positive relationship or not? We used some known forecasting algorithms for example ARIMA, Hybrid ARIMA, Auto-regression and SSA because they are the most popular and often-used technical indicators and is easy to calculate and, once plotted on a chart, it is a powerful visual trend-spotting tool which allows it to be employed quickly and easily.

We applied forecasting algorithms on KhanYounis city dataset and try to present results in relation views. Forecasting algorithms applied on water demand to predict the future of water needs by the growth of population with the time passage. Also, we forecasted the revenue and profits from water service by the growth of population with the time passage.

Chapter 4
Methodology and Model Development

Chapter 4
Methodology and Model Development

This chapter provides the methodology proposed for the forecasting municipal water consumption and revenues which involves four stages: (1) data collection, (2) data preprocessing, (3) building model, (4) testing and evaluating the model.

4.1. Research Approach

The research approach is divided into two categories: research methodology and work experimentation. The overall work completed for the applied model includes the experimentation methodology. The following sections give the overall view of the research model used for the research study and experimental methodology.

The proposed model is about forecasting water consumption and water revenue within customers and units' historical profiles registered in KhanYounis municipality KHM. The forecasting is going to be done for subareas and classes in KhanYounis city. The overall model methodology outlined in the thesis is shown in ***Figure (4-1***.

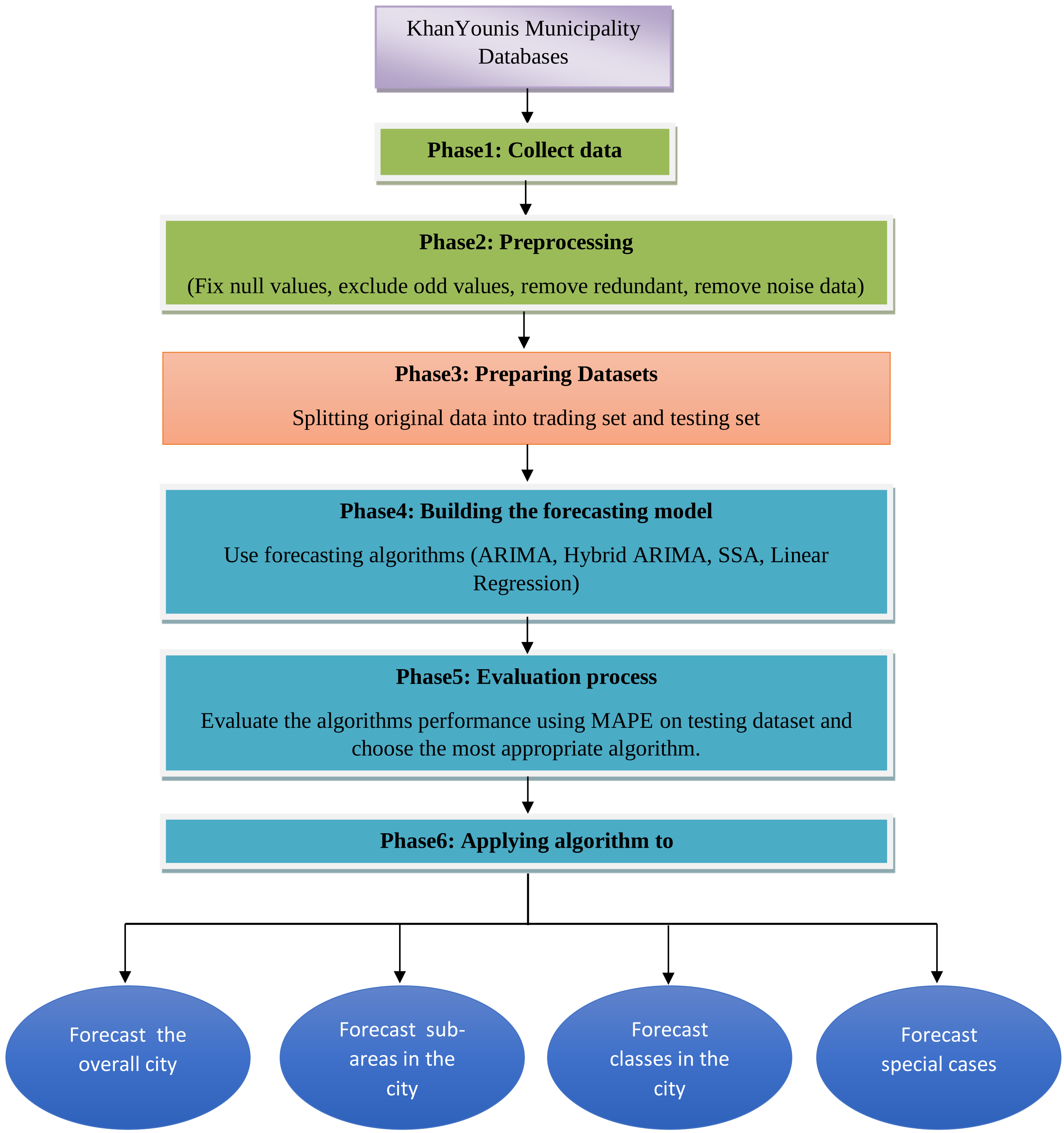

Figure (4-1) Proposed water consumption and revenue forecasting process

4.2. Data Collection

At this stage related data was collected from available databases in KhanYounis municipality. At KhanYounis municipality there are many internal systems, and more than such system deals and effects the water service data. So, related tables and column that belong to our case study were selected and combined into one dataset to prepare it for the next phase.

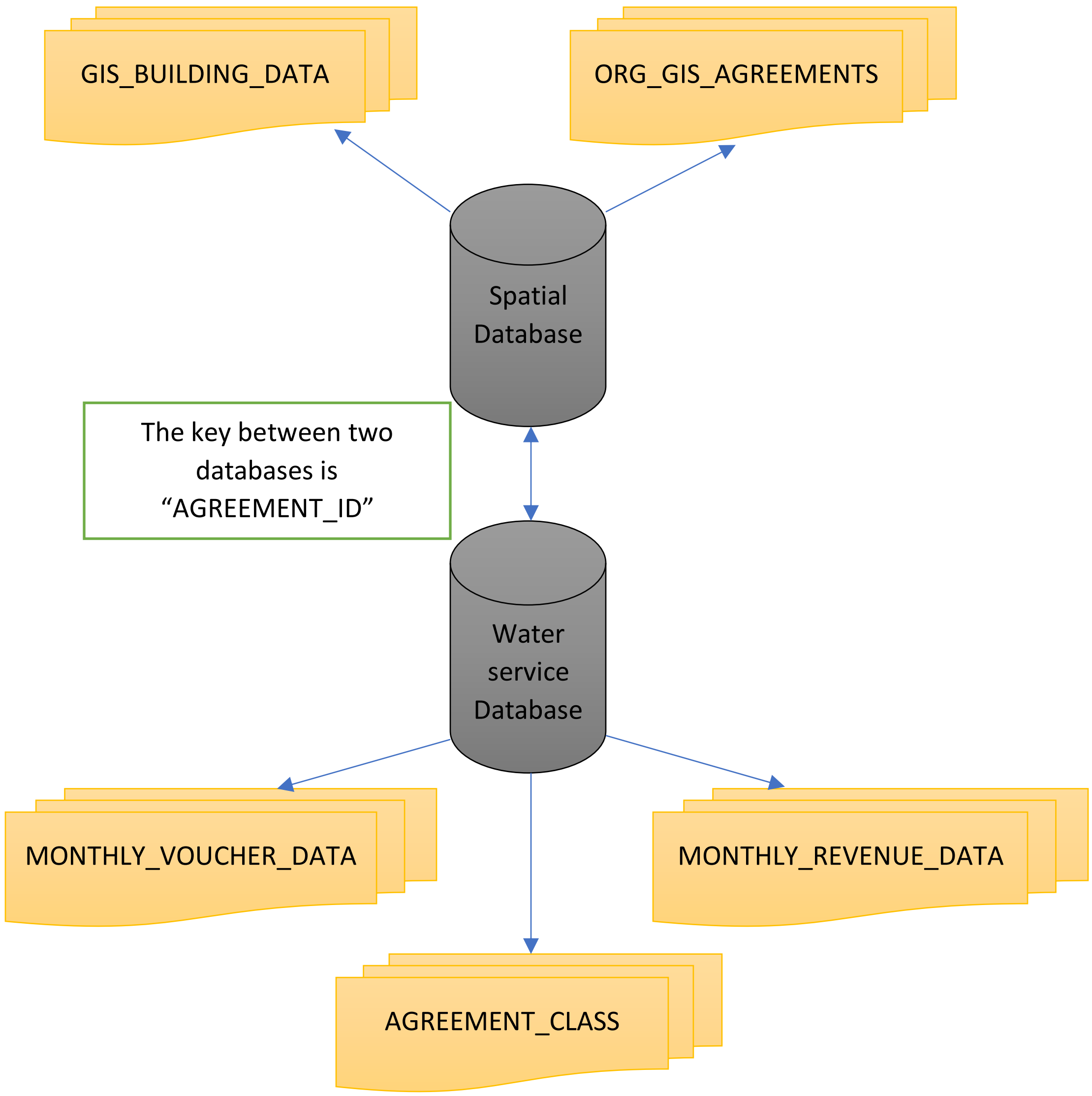

Figure (4-2) Overview of main databases used in KhanYounis Municipality belongs to water service system.

Figure (4-2 presents a main major database belongs to water service system and how data collection process is done from different datastores. First, 'Spatial Database' related to geolocation data for every unit benefited from water service this data includes x-coordinated, y-coordinated, counting zone, building number, neighborhood label and agreement-id which is the main key with the water service database. This database contains one table which registers one record for every existing building (regardless of number of floors) and another table registers all water agreements exists at every house unit at this building as shown in ***Table (4-1***. Second, 'Water Service Database' related to monthly transactions of water service operations like reading water meter devices, number of cubes consumed, date of reading, the status of meter device and the class nature of this water consumption unit. This database contains one table which registers one record for every existing water subscription (At KhanYounis city, there are 20949 water subscription) including some related managerial data like owner name, beneficiary name, registration date, class label etc. Many of these data were discarded because it is not important in our forecasting task and does not effects on the water consumption and water revenue numbers, another table which registers all water monthly transactions including the water consumption and water revenue values.

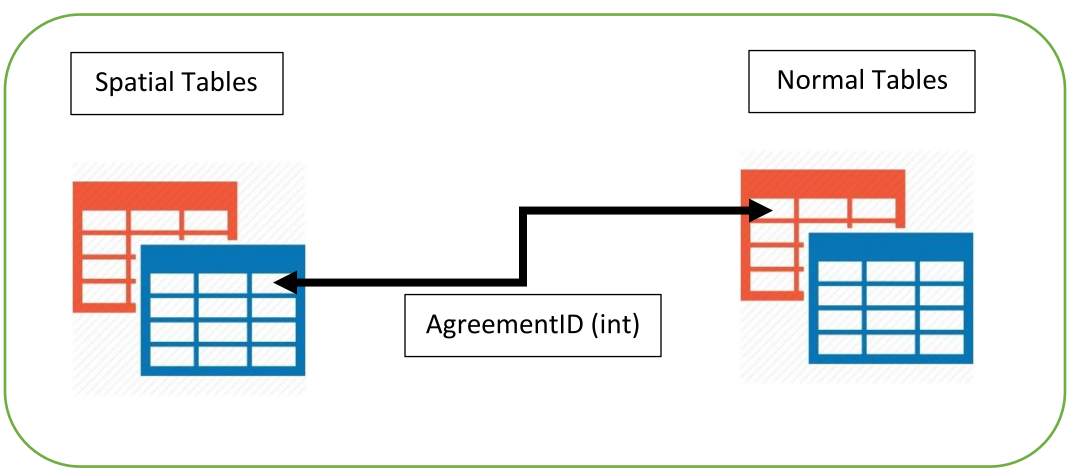

Figure (4-3) General view between spatial and normal tables with 'AgreementID' column.

Table (4-1) Sample of main GIS table for every existing building

AGREEMENT ID	FLATS COUNTS	ADDRESS	HOME NUMBER	ZONE NUMBER	X	Y
2636	2	حي الاسراء	31	217	34.29260479	31.3529276
8978	1	حاره الفرا	38	3	34.30477861	31.3419841
10478	1	الربوات الغربيه	150	129	34.30423448	31.35847237
10479	1	الربوات الغربيه بجوار الموظف مصطفى الاسطل	57	128	34.31200886	31.35631281
29125	1	السطر الغربي محيط مدرسة عيد الاغا	99	139	34.31930535	31.367178
10480	1	السطر الغربي محيط مدرسة عيد الاغا	99	139	34.31930535	31.367178

Table (4-2) Sample of main table for water subscriptions

AGREEMENT ID	ADDRESS	SUBSCRIPTION DATE
40928	تجاره عامه استيراد	28/01/2018
40927	قيزان ابو رشوان خلف مزرعة ياسر زعرب	25/01/2018
40926	مكتب هندسيالربوات الشرقيه	24/01/2018
40925	معمل معجنات	24/01/2018
40924	فوال المعسكر الجنوبي خلف مستشفي مبارك	22/01/2018
40923	تجاره عامه اجهزه كمبيوتر ش/ خالد بن الوليد	22/01/2018
40922	روضه جني المبدعه ش/ السوق شرق عياده الوكاله المعسكر الغربي	22/01/2018
40921	معن	21/01/2018
40920	شمال مبنى المحافظة	21/01/2018
40919	ورشه دهان سيارات ش/ المارس	21/01/2018
40918	فوال حي الامل العرايشيه	21/01/2018
40917	فوال حي الامل العرايشيه	21/01/2018
40916	المتارة شارع ابو زيد الفرا	21/01/2018
40915	ملابس مستعمله المحطه مقابل البلديه	21/01/2018
40914	فوال اسكان حي الامل العريشيه	20/01/2018
40913	البطن السمين - بجوار بير عيا	20/01/2018
40912	البطن السمين- بجوار بير عيا	20/01/2018
40911	معن-خلف شرطة المرور	20/01/2018
40910	بيع درجات هوائيه ش/5 التقاء مع ش/ جمال عبدالناصر	17/01/2018
40909	العاب كمبيوتر السطر الغربي	17/01/2018

Table (4-3) Sample of water subscriptions monthly transactions

AGREEMENT ID	VOUCHER DATE	VOUCHER AMOUNT (NIS)	CONSUMPTION QUANTITY (M^3)
1453	31/12/2017	45.75	14
1453	30/11/2017	45.75	13
1453	31/10/2017	45.75	7
1453	30/09/2017	45.75	18
1453	31/08/2017	45.75	20
1453	31/07/2017	54.38	25
1453	30/06/2017	120.83	37
1453	31/05/2017	45.75	11
1453	30/04/2017	187.65	24
1453	31/03/2017	75.08	37
1453	28/02/2017	25.89	35
1453	31/01/2017	45.75	20
1453	31/12/2016	105.55	51

Table (4-4) Sample of water subscription categorized over sub-regions

AGREEMENT ID	REGION NAME
1045	الأمل
64	الأمل
1077	الأمل
186	الأمل
13545	معن
14735	معن
11535	معن
13601	معن
13640	معن
14705	معن
13846	مركز المدينة
13851	مركز المدينة
16068	مركز المدينة
15175	مركز المدينة
14391	مركز المدينة
9391	الكتيبة
10490	الكتيبة
14087	الكتيبة
14195	الكتيبة

Table (4-5) Sample of water subscription categorized over classes

AGREEMENT ID	CLASS NAME
9996	روضة اطفال
1374	روضة اطفال
11287	روضة اطفال
10096	روضة اطفال
15629	روضة اطفال
1043	سكني
17312	سكني
1092	سكني
1151	سكني
1244	مساجد وزارة الاوقاف
17625	مساجد وزارة الاوقاف
122	مساجد وزارة الاوقاف
11724	مدارس ومرافق وزارة التربية والتعليم
16449	الوزارات والهيئات الأخرى المعتمدة
17726	مقرات الوكاله الغوث
20128	مقرات الوكاله الغوث
4065	مقرات الوكاله الغوث

As noticed from Table 4-1, Table 4-2, Table 4-3, Table 4-4 and Table 4-5 we used the primary key 'AGREEMENT ID' between tables and databases to collect and group our data into three categories, (1) monthly water data (Consumption and revenue) for all KhanYounis city as sample shown in Table ***(4-6***, (2) monthly water data for each sub-area (20 area) as sample shown in Table ***(4-7***, (3) monthly water data for each class (13 class) as sample shown in Table ***(4-8***.

Table (4-6) Sample of monthly water consumption and revenue for overall KhanYounis city.

VOUCHER DATE	WATER CONSUMPTION (M^3)	WATER REVENU (NIS)
01-2007	286273	450597.5
02-2007	293195	539147.2
03-2007	218840	429298.5
04-2007	312961	467045.8
05-2007	329750	510998.9
06-2007	357077	516393.4
07-2007	374079	538364.1
08-2007	389399	639859.2

09-2007	365029	624852.9
10-2007	397948	700426
11-2007	437264	740862.5
12-2007	211288	363127.5
01-2008	213842	398385.7
02-2008	182733	373401.5
03-2008	365713	527005.2

Table (4-7) Sample of monthly water consumption and revenue over regions

WATER CONSUMPTION (M^3)	WATER REVENUE (NIS)	VOUCHER DATE	REGION NAME
45022	65696.5	Jan-07	الأمل
28728	47303.5	Jan-07	البطن السمين
258	605.5	Jan-07	التحرير
16098	27186	Sep-09	الجلاء
11443	17282	Feb-09	السطر
90	450	Dec-08	السلام
13074	24566.5	Dec-07	الشيخ ناصر

Table (4-8) Sample of monthly water consumption and revenue over classes

WATER CONSUMPTION(M^3)	WATER REVENUE (NIS)	VOUCHER DATE	CLASS
245	636.0	01/2007	Ministries
107	136.5	01/2007	Banks
1050	2206.0	01/2007	Universities
10	160.0	01/2007	Family Office
41	100.0	01/2007	Kindergarten
266891	412086.5	01/2007	Houses
136	163.5	01/2007	Wedding halls
3878	7756.0	01/2007	MOE
4112	8348.0	01/2007	Municipal facilities
5688	11522.0	01/2007	MOH
1981	3850.5	01/2007	Societies and Institutions
1545	2484.5	01/2007	Mosques
574	1008.0	01/2007	UNRWA

Every dataset consists of four columns (water consumption *'double'*, water revenue *'double'*, voucher date *'date'*, class *'string'*). And to create this dataset for any level, we must read the water consumption and voucher date from the table which registers the monthly amount for every customer or unit record, and we read the paid water revenue from customers or units from table which registers the monthly payment transaction, and we read where this customer or unit in which sub-area and what is the class labeled for it from tables belongs to GIS system.

4.3. Data Preprocessing

At this stage preprocessing tasks were applied to prepare collected dataset to the forecasting algorithms, because data preparation is an important step for treating the data to be ready for the analysis and modeling steps. In some cases, some duplicated records for the same month with different values were found, this is due to the reading process from customers and units is done twice at the same month, and because our time series dataset assumes one single record for each month. So, we summed these duplicated records in one record only. **Table (*4-9*** shows an example of Al-Amal region which has two redundant records for the same month 2/2007, so we summed them into one single record which contains water consumption = 40597 and water revenue = 59442.5

Table (4-9) Sample of redundant records for the Al-Amal region at 2/2007

Water Consumption (M^3)	Water revenue (NIS)	Voucher date	Region name
31468	45532	2/2007	Al-Amal الامل
9129	13910.5	2/2007	Al-Amal الامل

Also, we found some outlier records in both attributes ‘Water Consumption’ and ‘Water Revenue’ in some sub-areas, and after asking the customer service department in KhanYounis municipality about this issue, they explained that these outliers record as result of wrong reading water meters or mis-typing while data entry process to water service system. They suggested to solve these outliers to take the average of same months for the last two years before this outlier month to be consistent in water use. **Figure (*4-4*** shows an example of Al-Salam region which has water consumption outlier record in 6/2009 with value 2043 and we processed it by taking the average of same months for the last two years (6/2008, 6/2007) which equals 45 as shown in **Figure (*4-5*.**

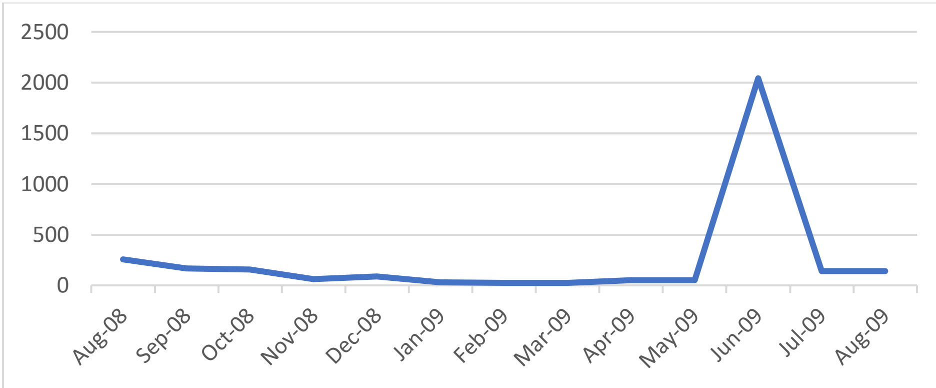

Figure (4-4) Water consumption outlier in Al-Salam region in 6/2009

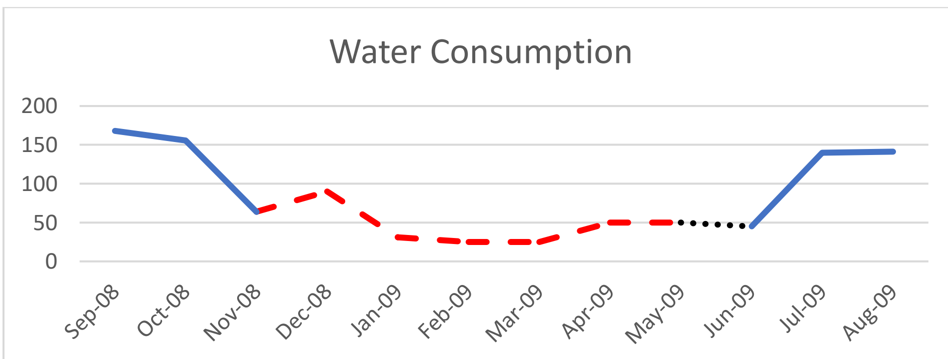

Figure (4-5) Processed water consumption value in Al-Salam region in 6/2009

In other cases, some sub-areas and classes were treated as special cases as explained in section 5.7, so the dataset for these special cases are presented yearly data not monthly data. **Figure (4-6** shows the monthly water consumption original data for Al-Jalaa region for 1/2007 to 10/2017, and **Figure (4-7** shows the yearly water consumption manipulated data for Al-Jalaa region for 2007 to 2017.

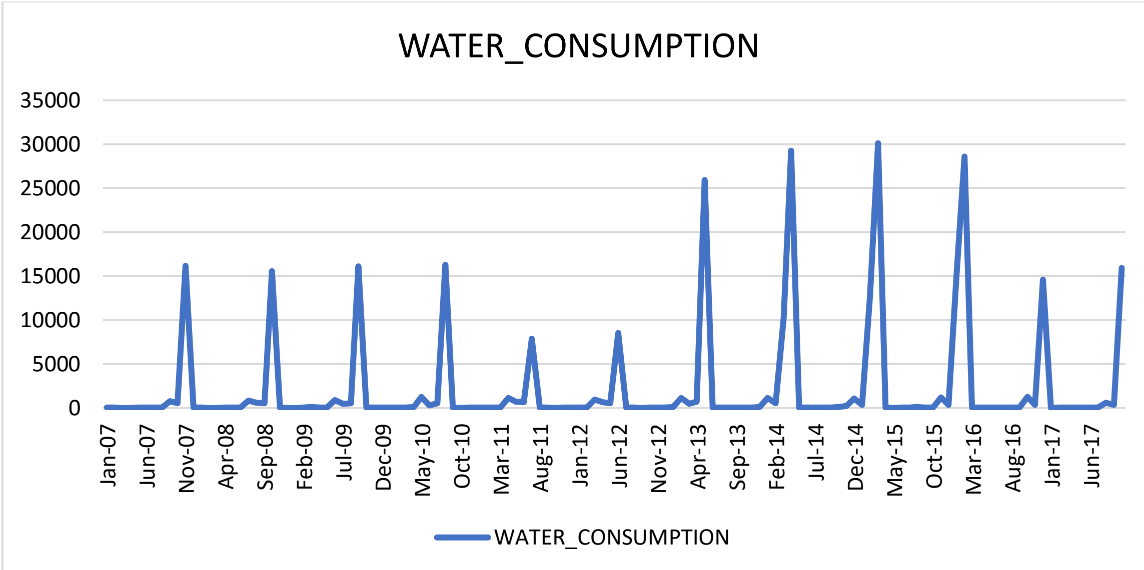

Figure (4-6) Al-Jalaa monthly water consumption data from 1/2007 to 10/2017

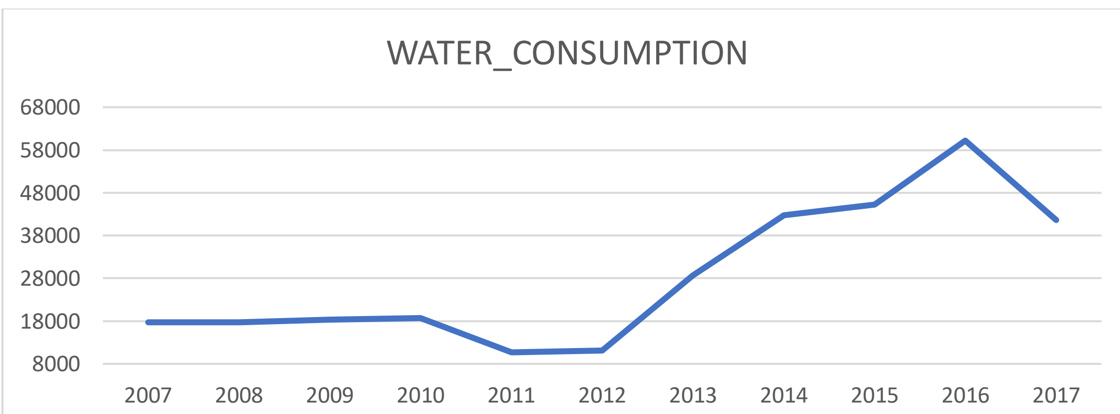

Figure (4-7) Al-Jalaa yearly water consumption data from 2007 to 2017

In class level we merged some sub-classes with main classes to have consistent categories of classes to ease to forecasting logic. For example, water consumption and revenue of private schools are summed with water consumption and revenue of governmental schools to have one consistent class called 'Schools and Educational facility' (مدرسة او مرفق تعليمي). Also, hospitals and primary care clinics centers are merged in one class called 'MOH' (مرافق وزارة الصحة).

4.4. Building Model

At this stage we built the model by applying four forecasting algorithms over the monthly water consumption and water revenue for the whole city to choose the appropriate algorithm depending on the lowest mean percentage error MPE. The four selected forecasting algorithms are Auto Regressive Integrated Moving Average (ARIMA), ARIMA combined with Neural Networks (Hybrid ARIMA), Singular Spectrum Analysis (SSA) and Linear Regression forecast. These algorithms are the most famous and recent algorithms in this field. We used four tools to build our model and test the selected algorithms over, these tools are RapidMiner Studio, R project, Anaconda and WEKA.

First step, we divided our original data into two parts, water consumption data and water revenue data as shown in ***Figure (4-8***, each part is separated into two datasets (training set and testing set). The training set for each starts from 01/2007 to 12/2015 which presets 83% from the original data and the testing set for each starts from 01/2016 to 10/2017 which presents 17% from the original data.

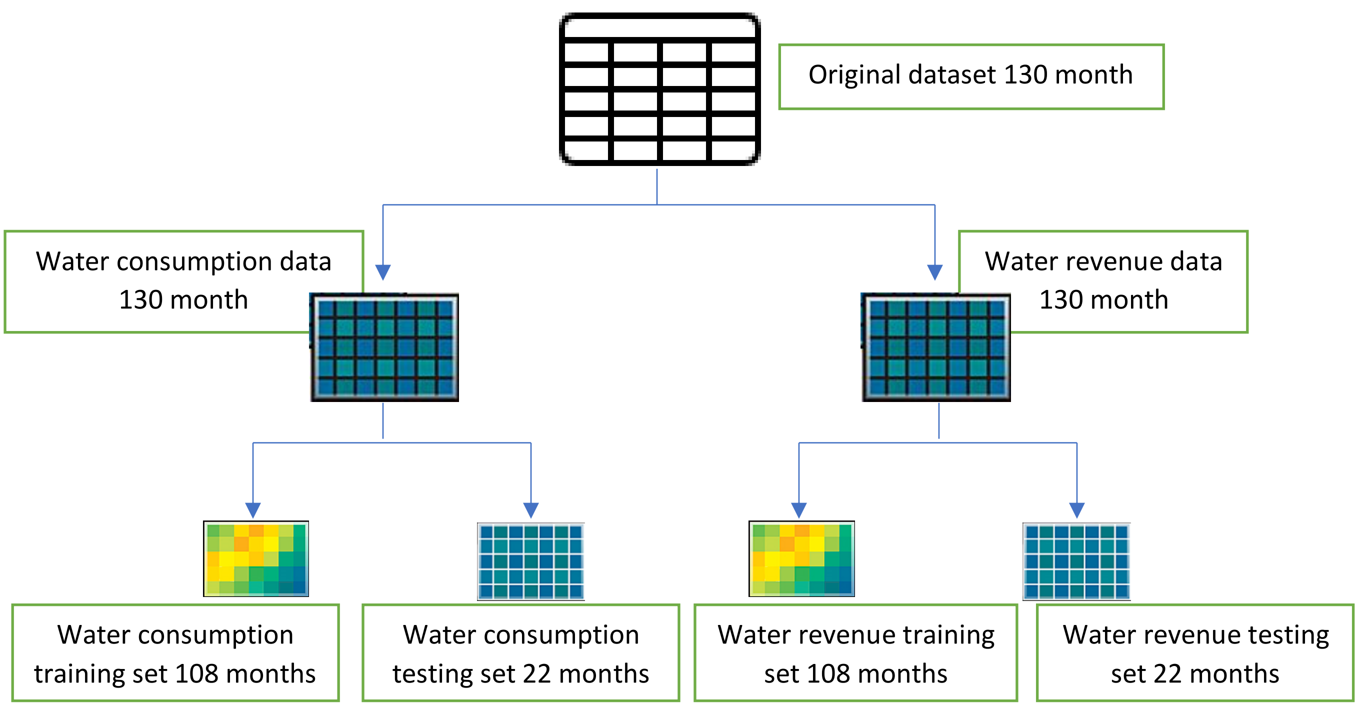

Figure (4-8) Original data splitting steps

Second step, we choose the most appropriate algorithms by running these four algorithms over the first part (water consumption data) on training set with horizon value 22 which means until 10/2017 to compare forecasting results with the testing set, and then apply the equation of MPE between the actual water consumption values in testing set and predicted values from forecasting results.

Third step, we run again the four algorithms over the second part (water revenue data) on training set with horizon value 22 which means until 10/2017 to compare forecasting results with the testing set, and then apply the equation of MPE between the actual water revenue values in testing set and predicted values from forecasting results.

Fourth step, we compared MPEs from different algorithms to decide which one is the most accurate algorithm to be used further in forecasting our data in three levels (overall city level, sub-regions level, class level).

4.5. Testing and Evaluating the Model

At this stage we evaluated the model accuracy using mean absolute percentage errors MAPE.

*(((Actual – Predicted) / Actual) *100)* ***Equation 4-1***

Also, the accuracy of forecasts is evaluated by comparing them with observed water demand and water revenue. This evaluation provides insights in recommending changes to existing models to reduce deviations in future forecasts.

After computing and comparing MPEs from different algorithms, we found that Hybrid ARIMA was the lowest MPE for both parts (water consumption part and water revenue part), which means the most appropriate algorithm.

4.6. Summary

Our model went into seven phases, from collecting data to applying forecasting algorithms for different levels. Data were collected from different databases, and some preprocessing tasks were held over collected data to prepare it for forecasting tasks. We build our model using several forecasting tools to run four forecasting algorithms over collected data. Finally, we tested our model using MAPE.

Chapter 5
Experimental Results and Discussion

Chapter 5
Experimental Results and Discussion

In this chapter we are going to describe experiments environment and the dataset, then we gave evaluation tasks of several time series forecasting algorithms to decide which one is the appropriate algorithm to be used forecasting five years in advance.

The chosen forecasting algorithms can be labeled as the most recent and famous algorithms used in time series analysis field. These algorithms are Auto Regressive Integrated Moving Average (ARIMA), ARIMA combined with Neural Networks (Hybrid ARIMA), Singular Spectrum Analysis (SSA) and Linear Regression forecast. We are going to run these four algorithms over our dataset to calculate the Mean Percentage Error (MPE) to detect which algorithms gives the lowest MPE, which mean the more accuracy algorithm to be used in further.

5.1. Experiment sets

Our experiments conducted using laptop HP Pavilion Core i7, Windows 10 Education version, 6 Giga RAM. We are going to use RapidMiner Studio 7.6.001, R studio, WEKA, Anaconda and Microsoft Excel in running all our experiments and plotting out results and charts.

5.2. Data set

Our data set is real data set collected from KhanYounis Municipality (KHM) – department of customer services, KhanYounis City, Gaza Strip, Palestine. The data set belongs to the monthly vouchers produced from KhanYounis municipality presenting the consumption amount of water in letters for each registered unit (double values), and the revenue fees of available water distributing service (double values) and voucher date in format month-year (mm-yyyy).

Our original data consists of four columns (Voucher Date, Area, Water Consumption, Water Revenue). The data sets containing almost eleven years of indices data (~11 years, 130 months) was divided into training set, containing the first nine years of data 01-2007 to 12-2015(9 years, 108 months), and testing set holding the remaining data which almost 2 years from 01-2016 to 10-2017 (~2 years, about 22 month). Models were trained on the training sets before being applied to the unseen data in the test sets. Table (5-1 gave a sample of training set for both attributes (consumption and revenue), and

Table (5-2 gave sample of testing set for both attributes (consumption and revenue).

Table (5-1) Sample of the training set from 1-2007 to 12-2015.

VOUCHER DATE	WATER CONSUMPTION (M^3)	WATER REVENUE (NIS)
01-2007	286273	450597.5
02-2007	293195	539147.2
03-2007	218840	429298.5
04-2007	312961	467045.8
05-2007	329750	510998.9
06-2007	357077	516393.4
07-2007	374079	538364.1

Table (5-2) Sample of the testing set from 01-2016 to 10-2017

VOUCHER DATE	WATER CONSUMPTION (M^3)	WATER REVENUE (NIS)
01-2016	459852	773740.5
02-2016	455574	777848.5
03-2016	264750	472531
04-2016	301128	488146.9
05-2016	388855	552890.7
06-2016	351510	510060.9

The testing set presents 17% of the overall dataset, and this is the period from 01/2016 to 10/2017 (22 months).

5.3. Evaluation forecasting algorithms

To choose the appropriate forecasting algorithms we first evaluated and compared the results from each algorithm with the original data we have. We run four forecasting algorithms on the training set and with horizon value 22 to reach 10-2017 to compare it with the testing set, at this stage we will only forecast the attribute ‘Water Consumption’ as sum of all areas over every month. These four forecasting algorithms are (ARIMA, Hybrid ARIMA, SSA and Linear Regression).

5.3.1. Evaluating algorithms over water consumption

First step we evaluated the selected four algorithms over the attribute ‘Water Consumption’ and compute the mean percentage error MPE for each algorithm.

5.3.1.1. ARIMA Evaluation

ARIMA refers to Auto Regressive Integrated Moving Average. An ARIMA model is defined by its three factors, p, d, q. p specifies the number of Auto Regressive terms in the model. d specifies the number of differentiations applied on the Time Series values. q specifies the number of Moving Average terms in the model. We used RapidMiner “Arima Trainer” Operator with values (p=1, d=0, q=1).

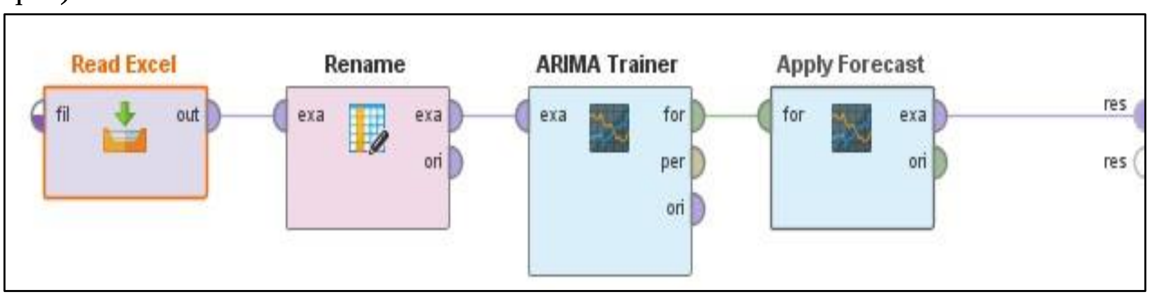

Figure (5-1) ARIMA RapidMiner Process

Figure (5-1 Presents RapidMiner process consists from four components. The first component is ‘Read Excel’ to read the data set. The second component is ‘Rename’ to change the input attributes to one single label to be unique for the rest of all components. The third component is ‘ARIMA Trainer’ is used to train the

ARIMA model over the dataset. Finally, the 'Apply Forecast' component is used to forecast the selected attribute for the detected horizon value.

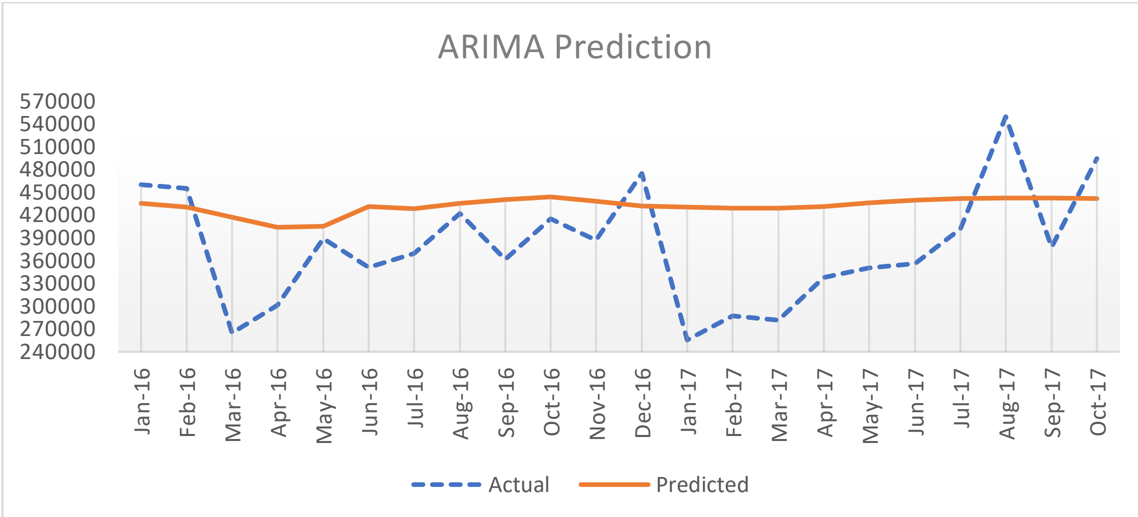

Figure (5-2) ARIMA Evaluation for water consumption (Actual vs Predicted).

Figure (5-2 shows that there are no compatible lines between the actual and predicted values and ARIMA gives predicted values almost horizontal line with few curves. After computing MAPE we found it 22.85%.

5.3.1.2. Hybrid ARIMA Evaluation

In this evaluation task we combined ARIMA with Neural Network (NN) to have Hybrid ARIMA with equal weight for each. While ARIMA may not be adequate with complex nonlinear problems and on the other hand NN yielded mixed results with linear problems. So, the combination handles effectively vary types of dataset nature.

```
library(forecastHybrid)
f=read.csv("Consumption 1-2007 to 12-2015.csv",sep=";")
ts1=ts(f,frequency=12,start=c(2007,1))
print(ts1)
plot.ts(ts1)
mod1 <- hybridModel(ts1,models = "an")
fc1 <- forecast(mod1)
print(fc1)
plot(fc1, ylab = "Water Consumption")
```

Figure (5-3) Hybrid ARIMA R code.

From ***Figure (5-3*** we can see R code that presents Hybrid ARIMA by using the attribute (models='an') in line 6, which 'a' refers to 'auto.arima' model and 'n' refers to Neural Network.

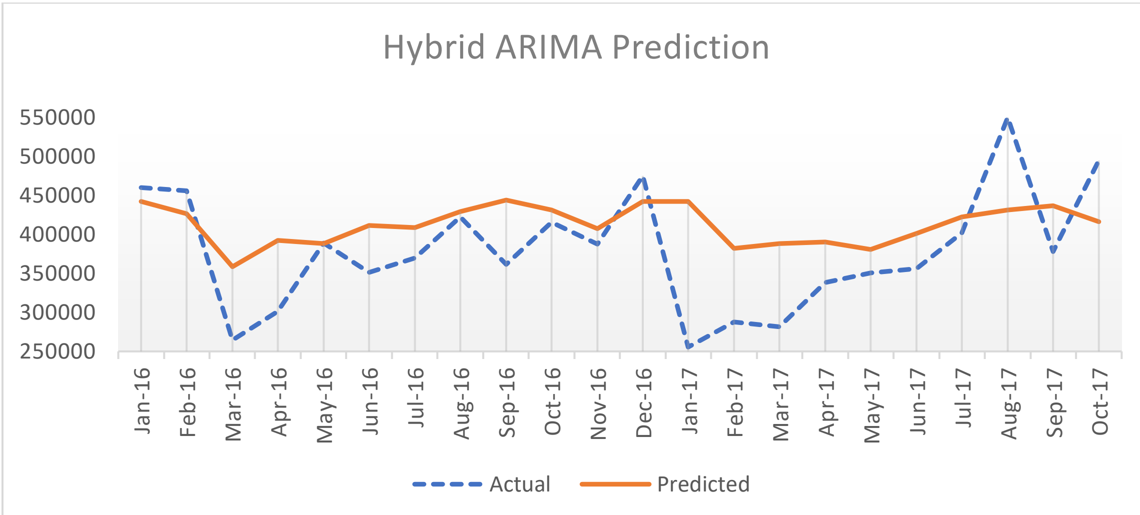

Figure (5-4) Hybrid ARIMA Evaluation for water consumption (Actual vs Predicted).

From **Figure *(5-4*** we can see that Hybrid ARIMA gives predicted values which are almost close with the actual value. Even though, there are some intervals which are differentiated, for example the predicted values are below the actual values from Jan-2016 to Feb-2016, also from Jul-2017 to Sep-2017 and in Dec-2016. After computing MAPE we found it 17.38%.

5.3.1.3. Singular Spectrum Analysis Evaluation

Singular Spectrum Analysis (SSA) is one of the modern non-parametric method for the analysis of time series. It provides a set of fast and reliable implementations of various routines to perform forecasting, decomposition and reconstruction. We used Python code in Anaconda software explained at **Appendix R**.

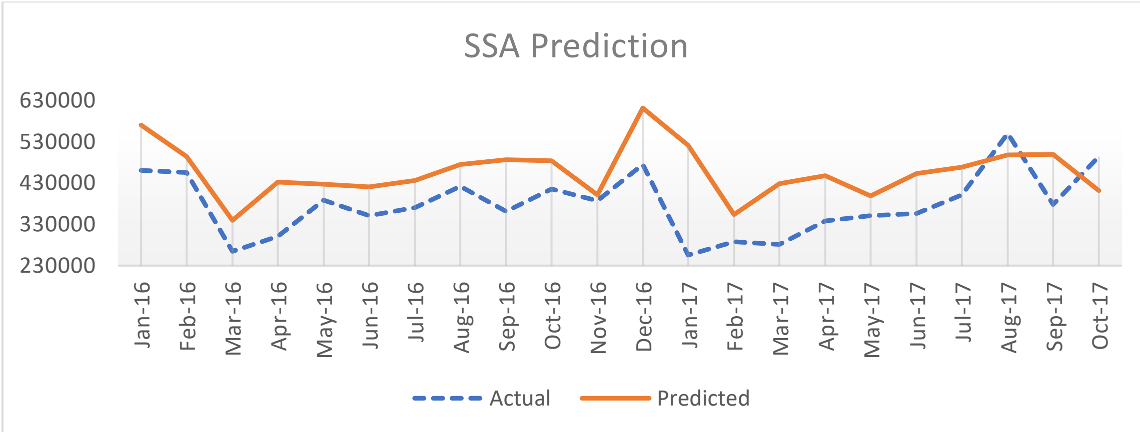

Figure (5-5) SSA Evaluation for water consumption (Actual vs Predicted).

From ***Figure (5-5*** we can see that there are miss-matching lines between the actual and predicted values and SSA gives predicted values which are differentiated with actual values specially between Feb-2017 and Mar-2017, also between Apr-2017 and May-2017, also between Sep-2017 and Oct-2017. After computing MAPE we found it 26.07%.

5.3.1.4. Linear Regression Evaluation

Linear Regression is a technique used for numerical prediction. Linear Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable (i.e. the label attribute) and a series of other changing variables known as independent variables (regular attributes). We used WEKA software to run built-in Linear Regression algorithm.

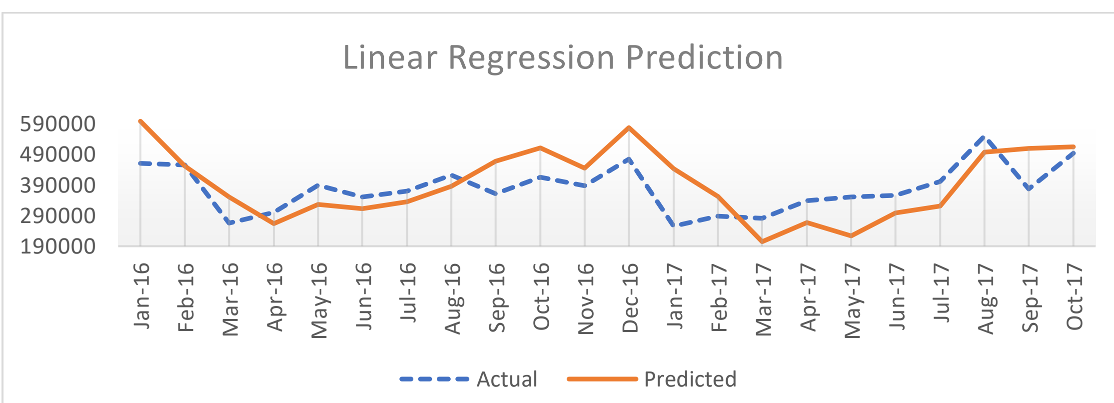

Figure (5-6) Linear Regression Evaluation for water consumption (Actual vs Predicted).

From ***Figure (5-6*** we can see that Linear Regression gives predicted values which are sometimes below the actual values specially on the interval between Apr-2016 and Aug-2016, and on the other hand there are predicted values above the actual values on the interval between Aug-2016 to Feb-2017. After computing MAPE we found it 18.8%.

5.3.1.5. Comparing Methods accuracy over 'Water Consumption'

Table (5-3) Comparing Methods MPE over 'Water Consumption'.

Algorithm	MPE%
ARIMA	22.85
Hybrid ARIMA	17.38
Linear Regression	18.80
SSA	26.07

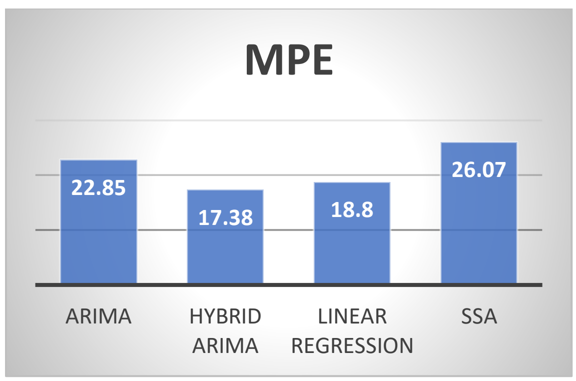

Figure (5-7) MPA percentages for four forecasting algorithms over water consumption attribute.

From ***Table (5-3*** and ***Figure (5-7*** which summarize the MPE results of the four algorithms, we can notice that Hybrid ARIMA is the most appropriate algorithms with least MPE 17.38% which means the forecasting values of predicting set are close to data of the testing set, while SSA is out of consideration with very high MPE 26.07% which means the forecasting values of predicting set are far away from data of the testing set. Linear regression has 18.8% and this is close to Hybrid ARIMA, but in general our datasets are multiple trends and non-linearity, that's why simple linear regression in our case study is often inappropriate for time-series work, and the proper use of linear regression depends on the data and the goals of the forecaster which doesn't fit in this research. So, at this stage Hybrid ARIMA is the appropriate algorithms which can be used further, but we have repeated and conducted all the above work over the attribute 'Water Revenue' in order to evalute and assess more the four algorithms on our data.

5.3.2. Evaluating algorithms over water revenue

After running the four algorithms on the attribute 'Water Consumption' and we noticed that hybrid ARIMA was the most accurate algorithm to be used in real forecasting, now repeated the past work on the attribute 'Water Revenue'.

5.3.2.1. ARIMA Evaluation

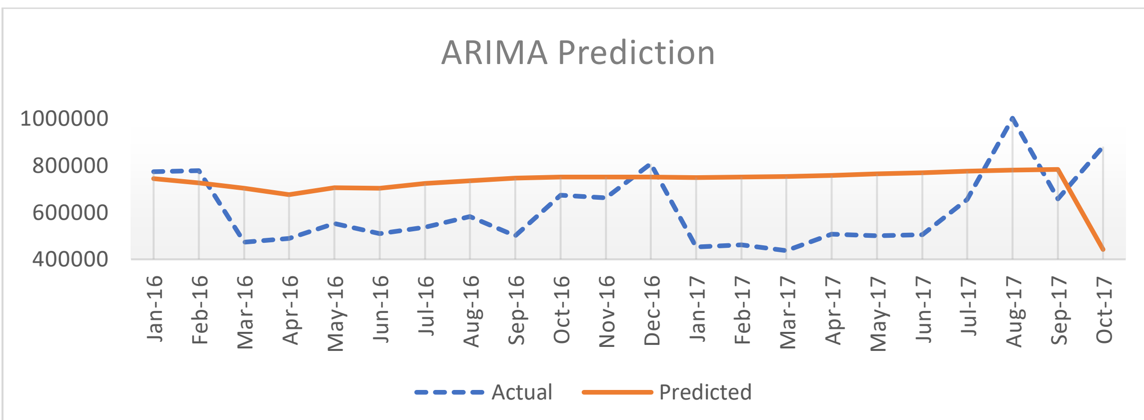

Figure (5-8) ARIMA Evaluation for water revenue (Actual vs Predicted).

After computing MAPE we found it 35.03%.

5.3.2.2. Hybrid ARIMA Evaluation

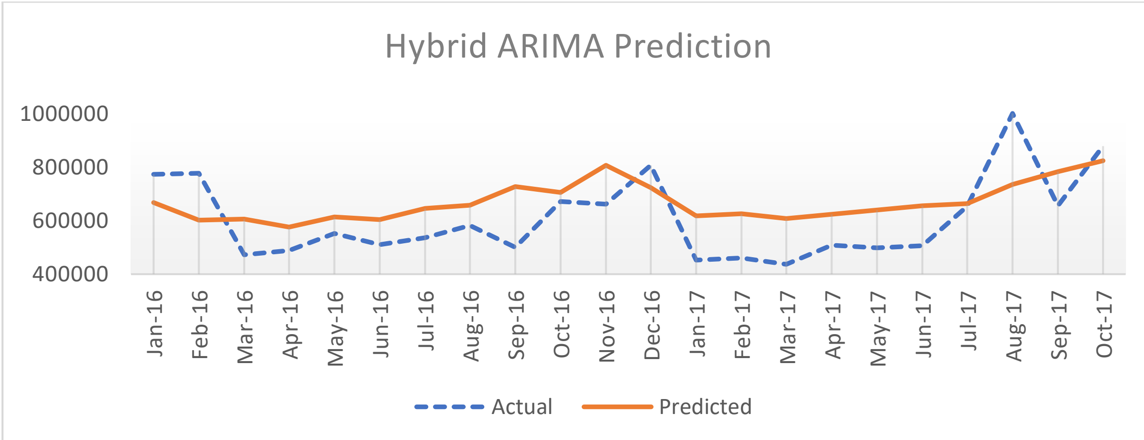

Figure (5-9) Hybrid ARIMA Evaluation for water revenue (Actual vs Predicted).

After computing MAPE we found it 21.57%.

5.3.2.3. Singular Spectrum Analysis Evaluation

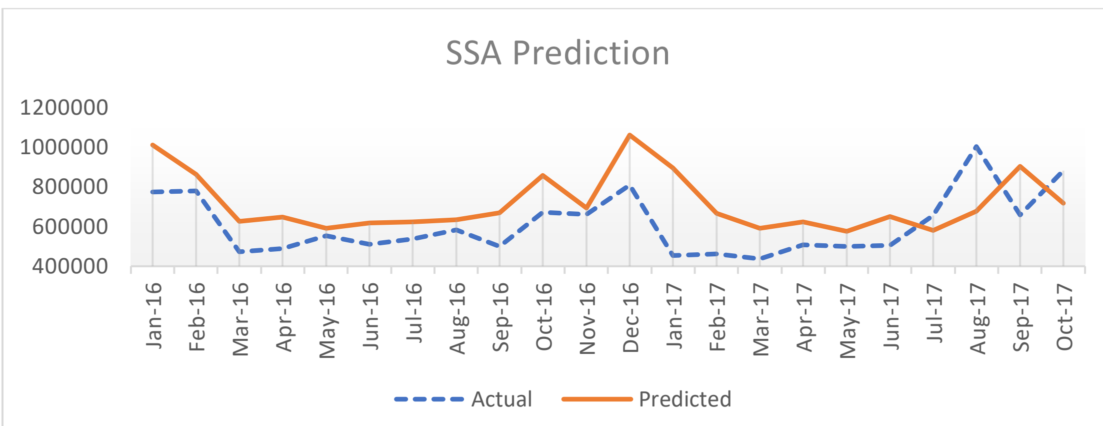

Figure (5-10) SSA Evaluation for water revenue (Actual vs Predicted).

After computing MAPE we found it 27.27%.

5.3.2.4. Linear Regression Evaluation

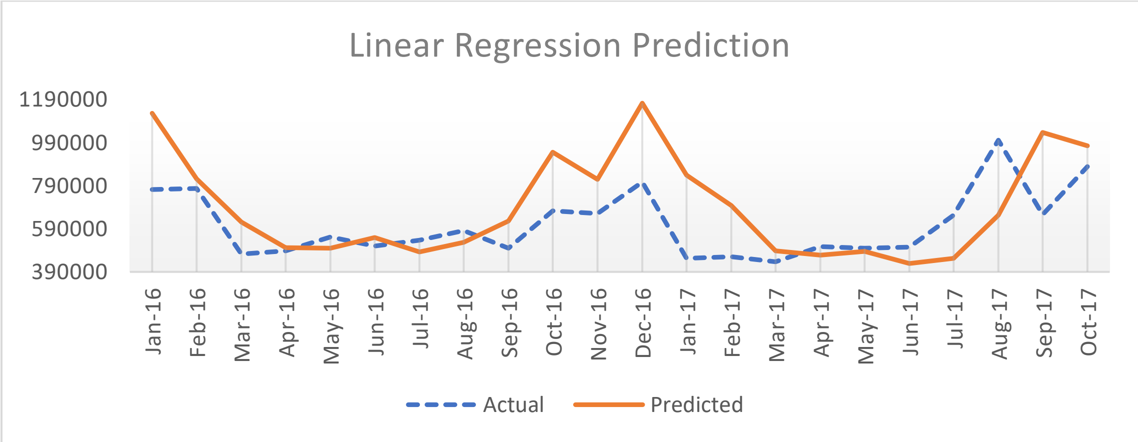

Figure (5-11) Linear Regression Evaluation for water revenue (Actual vs Predicted).

After computing MAPE we found it 25.77%.

5.3.2.5. Comparing Methods accuracy over 'Water Revenue'

Table (5-4) Comparing algorithms accuracy over 'Water Revenue'.

Algorithm	MPE%
ARIMA	35.03
Hybrid ARIMA	21.57
Linear Regression	25.77
SSA	27.27

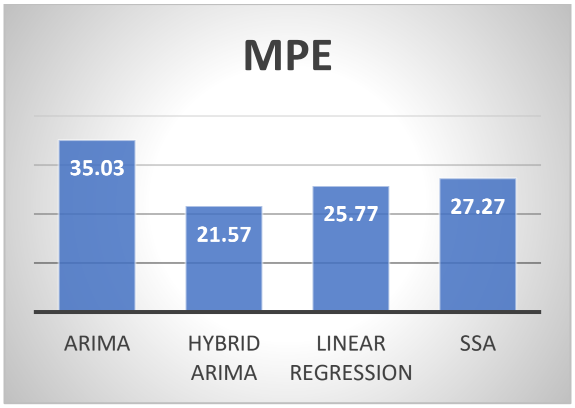

Figure (5-12) MPA percentages for four forecasting algorithms over water revenue attribute.

From ***Table (5-4*** and ***Figure (5-12*** which summarize the MPE results of the four methods, we again noticed that Hybrid ARIMA is the most appropriate algorithms with least MPE 21.57% which means the forecasting values of training set are close to data of the testing set, while SSA is out of consideration with very high MPE 27.27% which means the forecasting values of training set are far away from data of the testing set.

Finally, clearly illustrated that Hybrid ARIMA is the appropriate algorithms which will be used further in forecasting future 5 years over the attributes ‘Water Consumption and ‘Water Revenue’. We propose to take a combining approach to time series forecasting. The linear ARIMA model and the nonlinear ANN model are used jointly, aiming to capture different forms of relationship in the time series data. The hybrid model takes advantage of the unique strength of ARIMA and ANN in linear and nonlinear modeling. For complex problems that have both linear and nonlinear correlation structures, the combination method can be an effective way to improve forecasting performance.

After choosing Hybrid ARIMA to continue our forecasting tasks, we forecasted three levels, 1^{st} level is the whole city, 2^{nd} level is the sub-areas and the 3^{rd} level is the classes inside KhanYounis city.

5.4. Forecasting the overall city

Now we forecast the whole city in advance 5 years regardless of sub-areas or classes within KhanYounis city. Then we computed the deviation percentage for every forecasted year and compare it with year 2017.

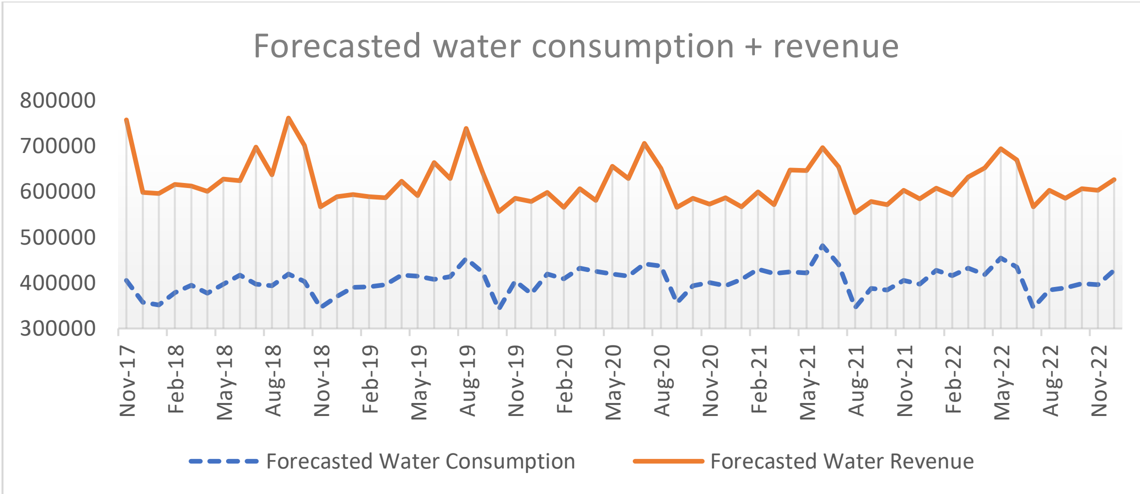

Figure (5-13) Five years forecasting water consumption and revenue for KhanYounis city.

From ***Figure (5-13*** we can notice that the two curves almost having the same trend, which means when forecasted water consumption increases the forecasted revenue also increases and that is logical and vice versa. But there are some exceptions such as the interval from 01-2019 to 03-2019, the curve of water consumption is decreasing but the curve of revenue in increasing.

5.4.1. Consumption and revenue future deviations

We compared the last year of the original dataset with the forecasted data for both 'Water Consumption' and 'Water Revenue', to help decision makers in planning water distributing and production systems and give them future vision of how much KhanYounis city will be using water.

Table (5-5) Consumption deviation in comparison to 2017.

Year	Water Consumption (M^3)	Water Consumption Deviation %
2017	4555059	0
2018	4693144	3.03
2019	4764013	4.59
2020	4928643	8.20
2021	4938253	8.41
2022	4901387	7.60

Table (5-6) Revenue deviation in comparison to 2017

Year	Water Revenue (NIS)	Water Revenue Deviation %
2017	7527434	0
2018	7828069	3.99
2019	7367335	-2.13
2020	7306228	-2.94
2021	7242724	-3.78
2022	7396991	-1.73

From Table (5-5 we can notice that water consumption amounts will be increasing in all next 5 years in comparison to 2017, but on the other hand, from ***Figure (5-6*** water revenues are increasing only in 2018 but decreasing from 2019 to 2022. Generally, we can say that after 5 years the maximum water consumption will increase with 8.4% (1.7% per year) in comparison with 2017 but the minimum water revenue will be decreased with 3.8% (0.76% per year).

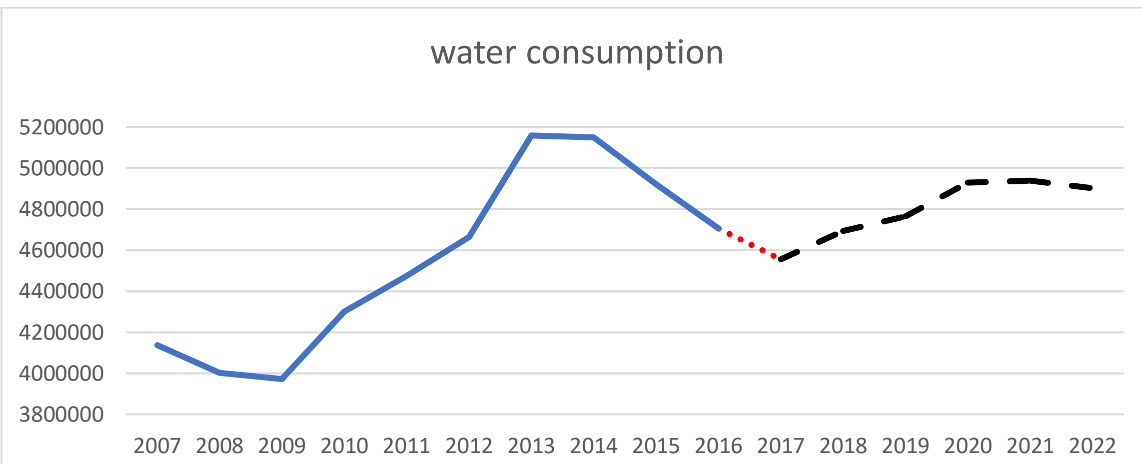

Figure (5-14) Water consumption data (Actual + forecasted) from 2007 to 2022

Figure (*5-14* shows the whole yearly consumption data (actual + forecasted), the dotted line shows 2017 consumption data, while the dashed line shows the 5 years forecasted data from 2018 to 2022. By visual it's clear that the dashed line trend is going up with comparison to dotted line, which means the average of water consumptions for the next five years will increase.

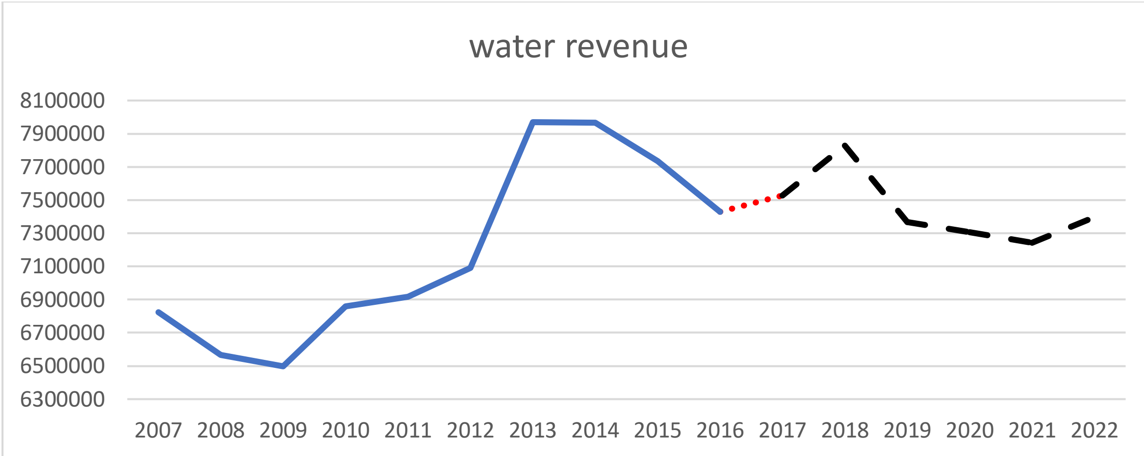

Figure (5-15) Water revenue data (Actual + forecasted) from 2007 to 2022

Figure (5-15 shows the whole yearly consumption data (actual + forecasted), the dotted line shows 2017 consumption data, while the dashed line shows the 5 years forecasted data from 2018 to 2022. By visual, the dashed line trend is going down with comparison to dotted line, which means the average of water revenues for the next five years will decrease.

5.5. Forecasting sub-regions

After choosing Hybrid ARIMA to continue our forecasting tasks, we forecasted 'Water Consumption' and 'Water Revenue' for all sub-areas in KhanYounis city (20 sub-area) with horizon value 62 (~5 years, about 62 month) until the end of 12-2022, these sub-areas are different in its nature, some of them have high population density, other has low population, while others are agricultural nature, and some of them are artificial crafts and business traffic nature. For example, the City Center (مركز المدينة) and Al Moaskar (المعسكر) regions are famous of their business traffic, while Al Amal(الامل), Al Moaskar (المعسكر) and Al Sheikh Naser (الشيخ ناصر) regions are famous of their high residential nature. South Mawase (المواصي الجنوبي), North Mawase (المواصي الشمالي) and Al Tahrir(التحرير) regions are famous of their agricultural and very low population density.

Every dataset for each region consists of 130 records (~10 years, about 130 months) starting from 01-2007 to 10-2017 with the attributes 'Water Consumption', ‘Water Revenue’, ‘Region Name’ and ‘Voucher Date’, and the target end of forecasting is to 12-2022 for every region.

Before starting real forecast for the 20 sub-areas, we presented the water consumption and revenue percentages of every area from the overall consumption rate and revenue rate for the year 2017 where water consumption equals 3693231 and water revenue equals 6057109 to show which areas weights high and which weights low in both attributes ‘Water Consumption’ and ‘Water Revenue’.

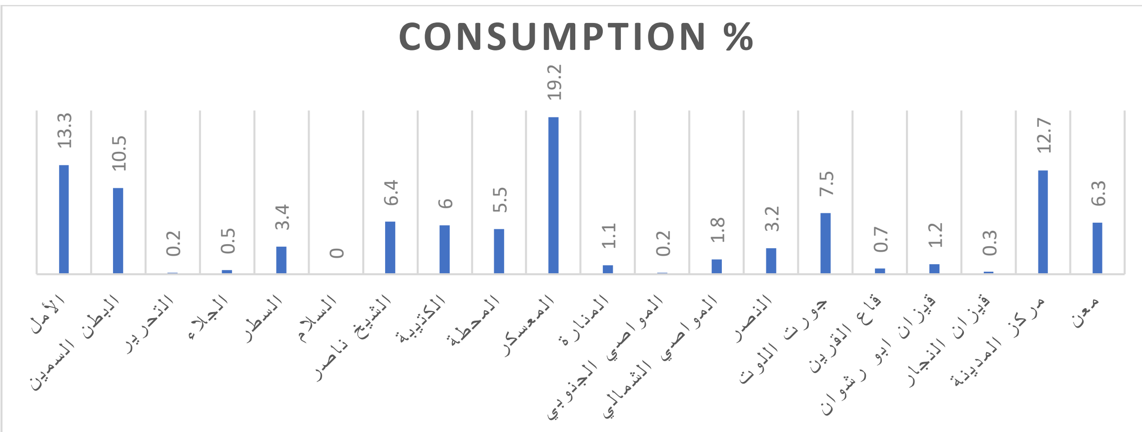

Figure (5-16) Water consumption rates for regions in 2017

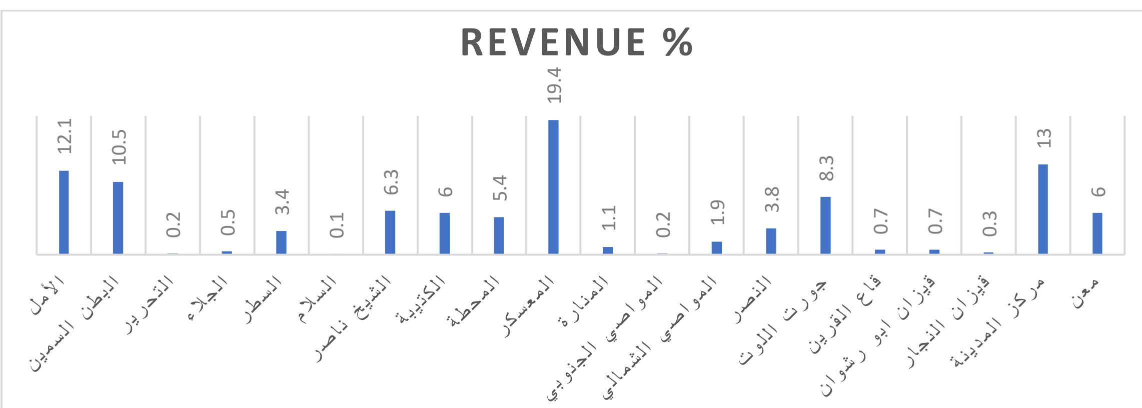

Figure (5-17) Water revenue rates for regions in 2017

Figure (*5-16* and **Figure (*5-17*** shows that Al-Moaskar area weights the highest area in both attributes ‘Water Consumption’ and ‘Water Revenue’, while Al-Tahrir, Al-Salam and South-Mawase are weighting the lowest areas in both attributes ‘Water Consumption’ and ‘Water Revenue’.

In ***Table (5-7*** we display the results of last forecasted month for each sub-area for the attributes 'Water Consumption' and 'Water Revenue' for every region, and all forecasted data will be available on Appendix A.

Table (5-7) Last month of the forecasted water consumption and revenue.

Region Name	Voucher Date	Forecasted Water Consumption (M^3)	Forecasted Water Revenue (NIS)
الأمل	12-2022	50918	89408
البطن السمين	12-2022	42152	64740
التحرير	12-2022	644	1332
الجلاء	12-2022	21925	-29765
السطر	12-2022	13512	22137
السلام	12-2022	129	313
الشيخ ناصر	12-2022	27604	38903
الكتيبة	12-2022	24582	37145
المحطة	12-2022	21110	34861
المعسكر	12-2022	84749	117238
المنارة	12-2022	5568	9226
المواصي الجنوبي	12-2022	1123	1736
المواصي الشمالي	12-2022	7460	11774
النصر	12-2022	9419	15356
جورت اللوت	12-2022	25789	38764
قاع القرين	12-2022	5158	13293
قيزان ابو رشوان	12-2022	2764	4017
قيزان النجار	12-2022	1168	2054
مركز المدينة	12-2022	52949	84600
معن	12-2022	26240	36972

Below we display charts of some regions showing the forecasting results of the attribute 'Water Consumption'.

1. Al Amal(الامل)

Figure (*5-18* shows the forecast results for Al-Amal region between 2018 and 2022, and we found that the water consumption deviation percentage is 5.40%, while the water revenue deviation percentage is 2.36% as shown in **Table 5-8**.

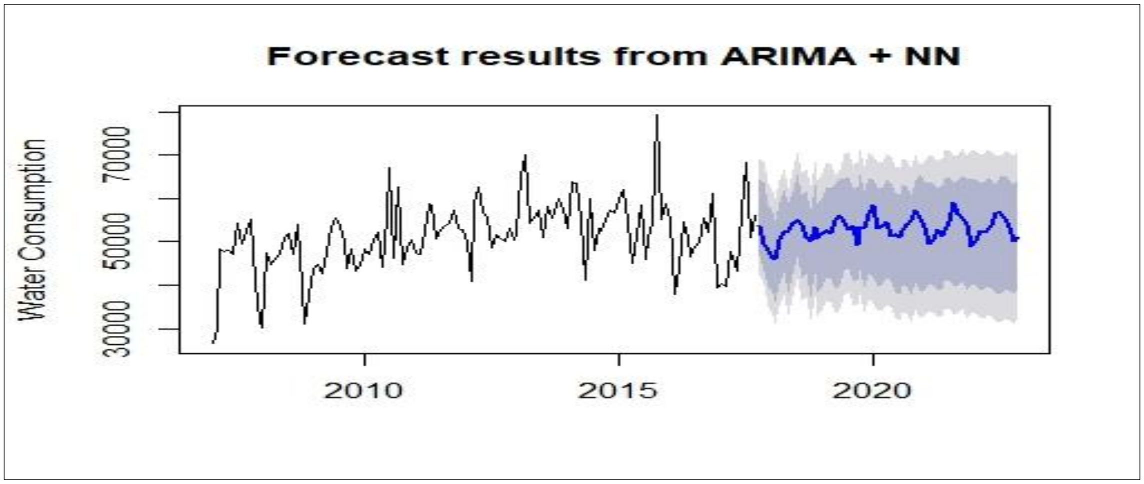

Figure (5-18) Al Amal forecasting water consumption results from 2018 to 2022.

2. Al Tahrir(التحرير)

Figure (*5-19* shows the forecast results for Al-Tahrir region between 2018 and 2022, and we found that the water consumption deviation percentage is 2.95%, while the water revenue deviation percentage is -14.47% as shown in **Table 5-8**.

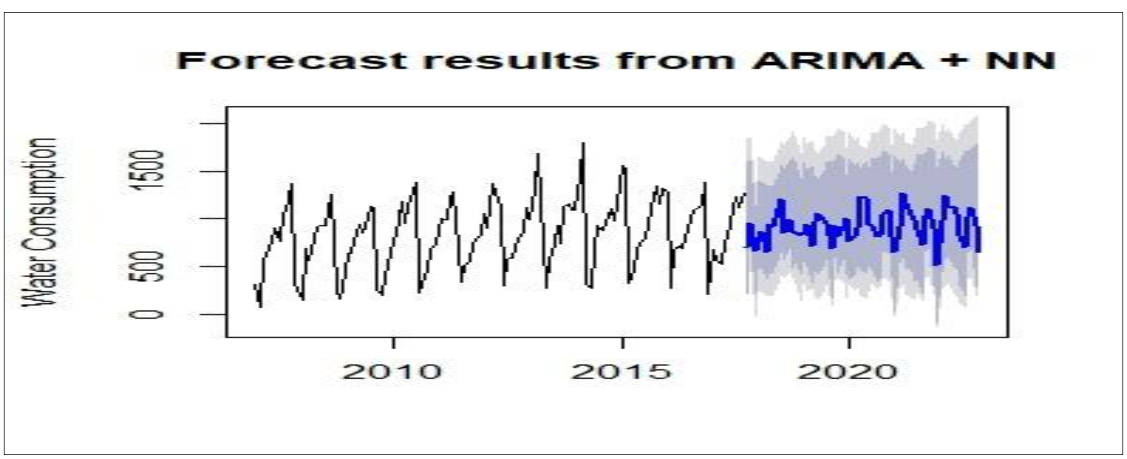

Figure (5-19) Al Tahrir forecasting water consumption results from 2018 to 2022.

3. Al Salam(السلام)

Figure (*5-20* shows the forecast results for Al-Salam region between 2018 and 2022, and we found that the water consumption deviation percentage is -16.48%, while the water revenue deviation percentage is -1.96% as shown in **Table 5-8**.

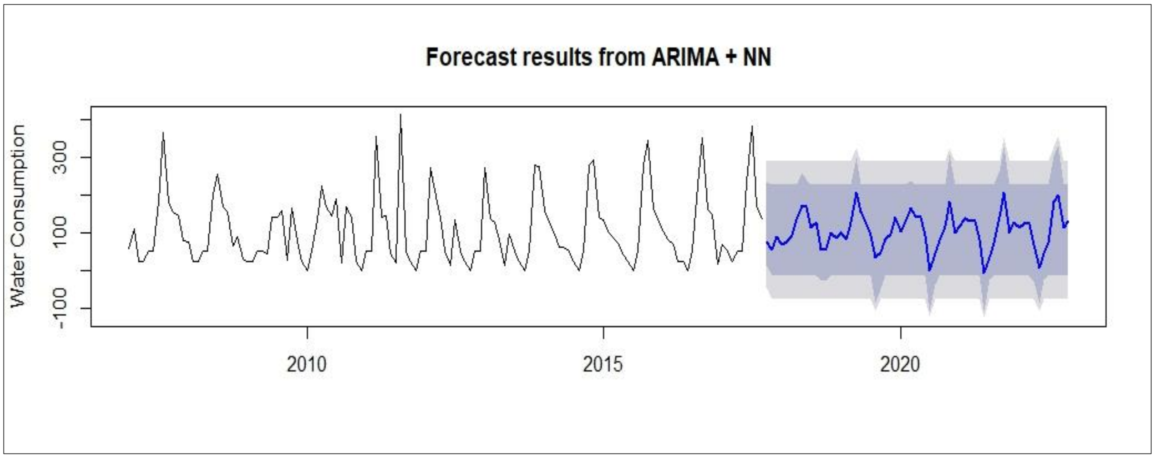

Figure (5-20) Al Salam forecasting water consumption results from 2018 to 2022.

4. Al Moaskar(المعسكر)

Figure (*5-21* shows the forecast results for Al- Moaskar region between 2018 and 2022, and we found that the water consumption deviation percentage is 5.92%, while the water revenue deviation percentage is 0.93% as shown in **Table 5-8**.

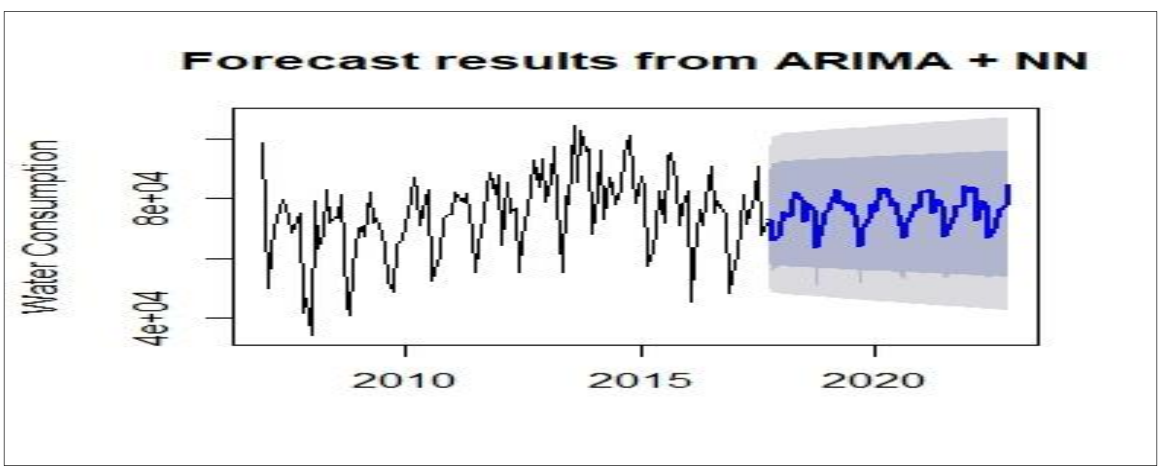

Figure (5-21) Al Moaskar forecasting water consumption results from 2018 to 2022.

5. Al-Batn (البطن السمين)

Figure (*5-22* shows the forecast results for Al- Batn region between 2018 and 2022, and we found that the water consumption deviation percentage is -2.80%, while the water revenue deviation percentage is -4.37% as shown in **Table 5-8**.

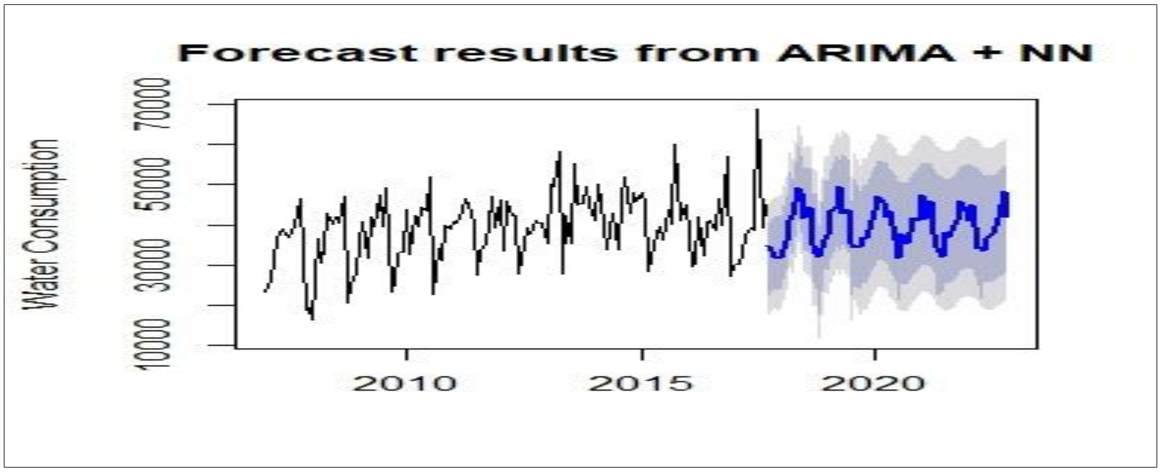

Figure (5-22) Al Batn forecasting water consumption results from 2018 to 2022.

6. Al-Jalaa(الجلاء)

Figure (*5-23* shows the forecast results for Al- Jalaa region between 2018 and 2022, and we found that the water consumption deviation percentage is -39.33%, while the water revenue deviation percentage is 27.32% as shown in **Table 5-8**.

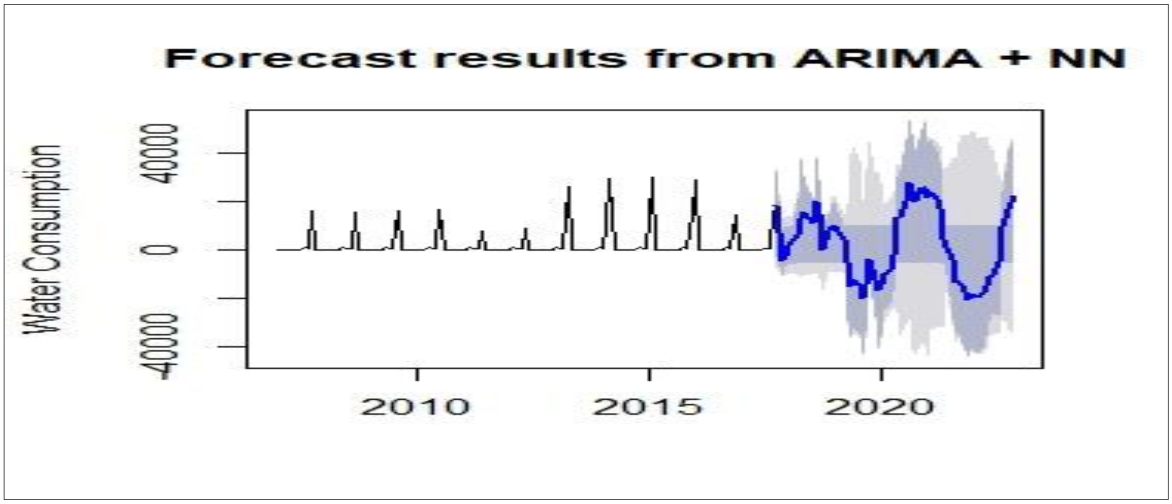

Figure (5-23) Al Jalaa forecasting water consumption results from 2018 to 2022.

7. Al-Satar(السطر)

Figure (*5-24* shows the forecast results for Al- Satar region between 2018 and 2022, and we found that the water consumption deviation percentage is 8.34%, while the water revenue deviation percentage is 1.69% as shown in **Table 5-8**.

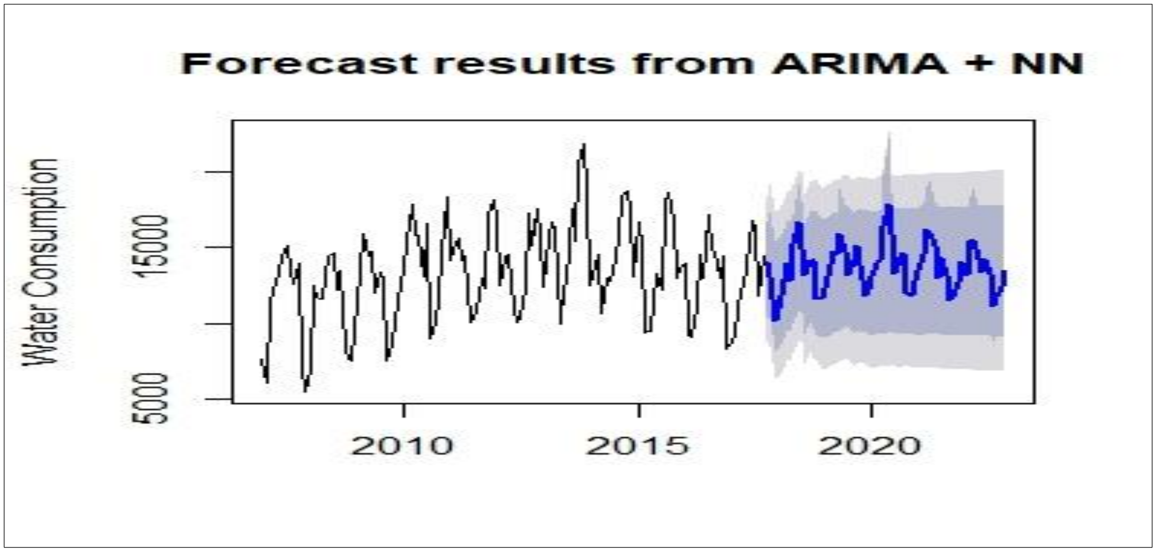

Figure (5-24) Al Satar forecasting water consumption results from 2018 to 2022.

8. Al-Sheikh-Naser(الشيخ ناصر)

Figure (*5-25* shows the forecast results for Al- Sheikh-Naser region between 2018 and 2022, and we found that the water consumption deviation percentage is 0.90%, while the water revenue deviation percentage is 1.25% as shown in **Table 5-8**.

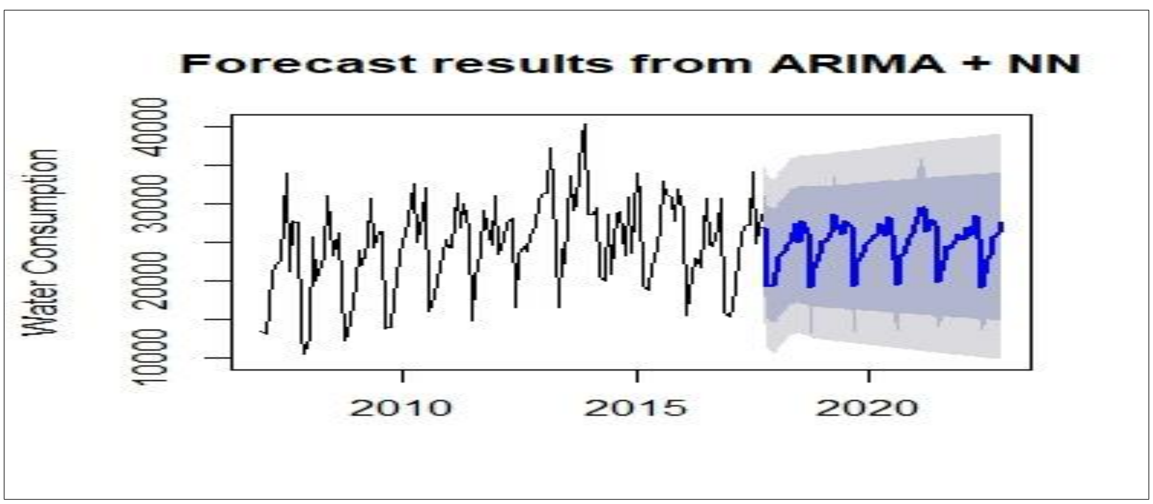

Figure (5-25) Al Sheikh-Naser forecasting water consumption results from 2018 to 2022.

9. Al-Kateeba(الكتيبة)

Figure (*5-26* shows the forecast results for Al-Kateeba region between 2018 and 2022, and we found that the water consumption deviation percentage is 6.94%, while the water revenue deviation percentage is 1.30% as shown in **Table 5-8**.

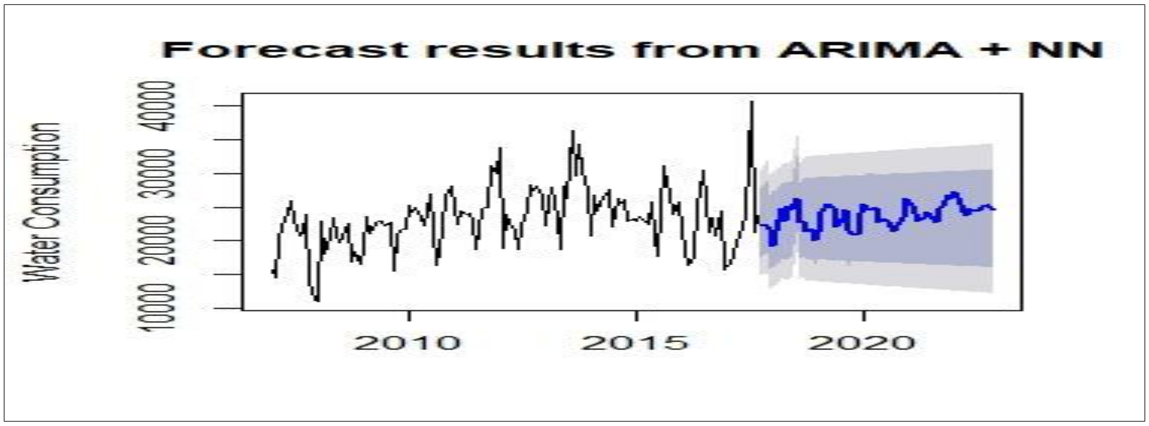

Figure (5-26) Al Kateeba forecasting water consumption results from 2018 to 2022.

10. Al-Mahata(المحطة)

Figure (*5-27* shows the forecast results for Al- Mahata region between 2018 and 2022, and we found that the water consumption deviation percentage is 0.14%, while the water revenue deviation percentage is 0.11% as shown in **Table 5-8**.

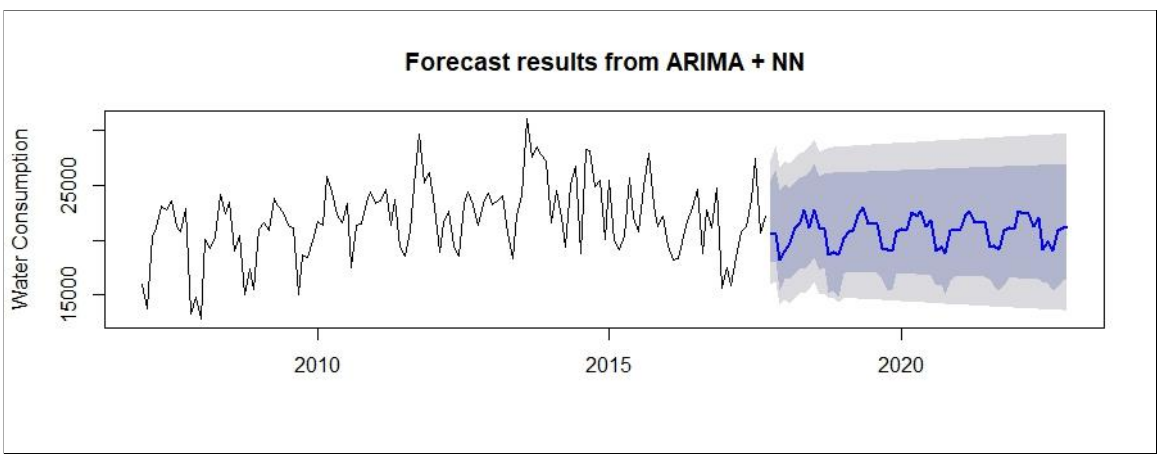

Figure (5-27) Al Mahata forecasting water consumption results from 2018 to 2022.

11. Al-Manara (المنارة)

Figure (*5-28* shows the forecast results for Al-Manara region between 2018 and 2022, and we found that the water consumption deviation percentage is -1.46%, while the water revenue deviation percentage is -4.46% as shown in **Table 5-8**.

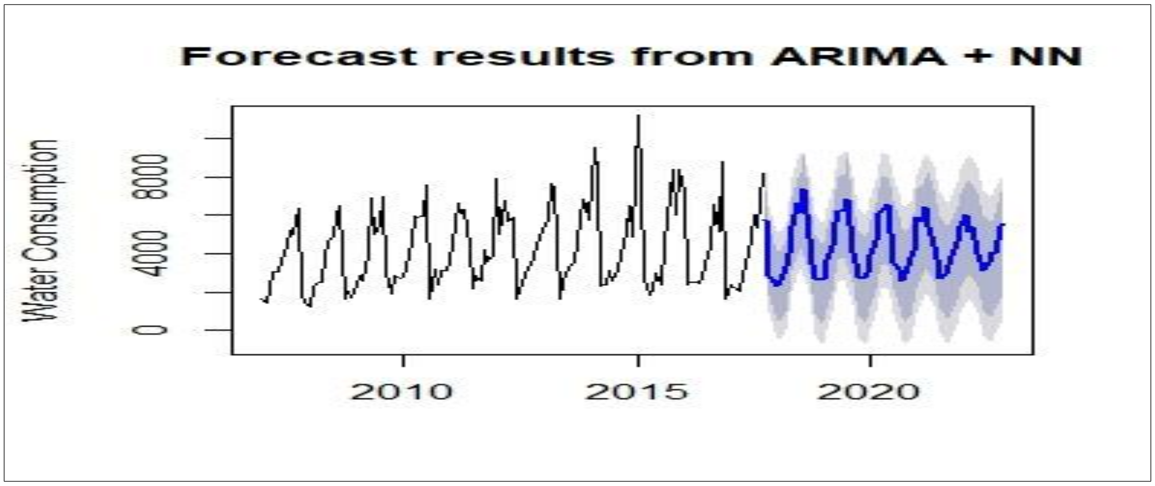

Figure (5-28) Al-Manara forecasting water consumption results from 2018 to 2022.

12. South-Mawase (المواصي الجنوبي)

Figure (*5-29* shows the forecast results for South-Mawase region between 2018 and 2022, and we found that the water consumption deviation percentage is -5.54%, while the water revenue deviation percentage is -13.25% as shown in **Table 5-8**.

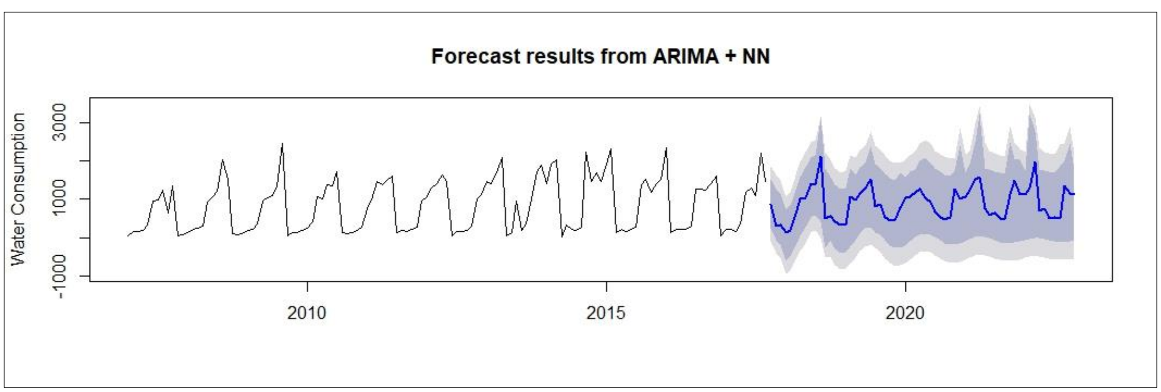

Figure (5-29) South-Mawase forecasting water consumption results from 2018 to 2022.

13. North-Mawase (المواصي الشمالي)

Figure (*5-30* shows the forecast results for North-Mawase region between 2018 and 2022, and we found that the water consumption deviation percentage is 3.93%, while the water revenue deviation percentage is -3% as shown in **Table 5-8**.

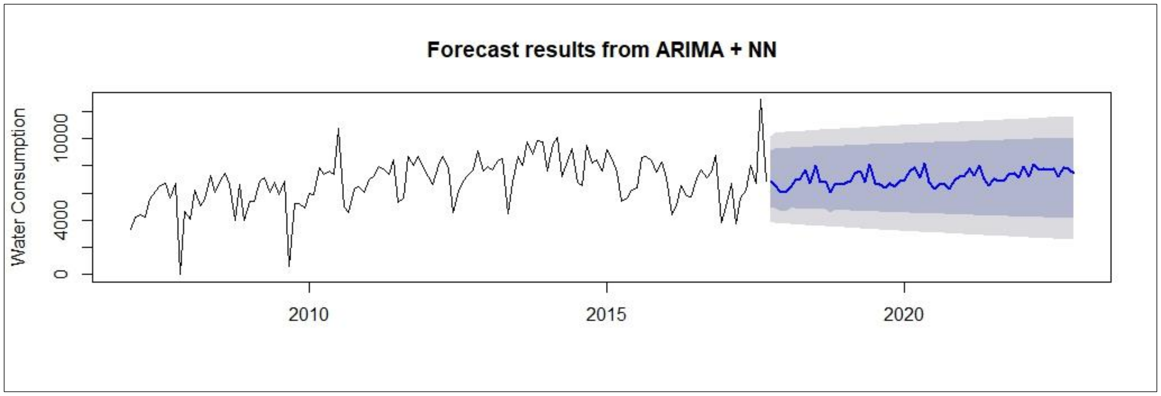

Figure (5-30) North -Mawase forecasting water consumption results from 2018 to 2022.

14. Al-Nasr (النصر)

Figure (*5-31* shows the forecast results for Al-Nasr region between 2018 and 2022, and we found that the water consumption deviation percentage is -30.19%, while the water revenue deviation percentage is 14.35% as shown in **Table 5-8**.

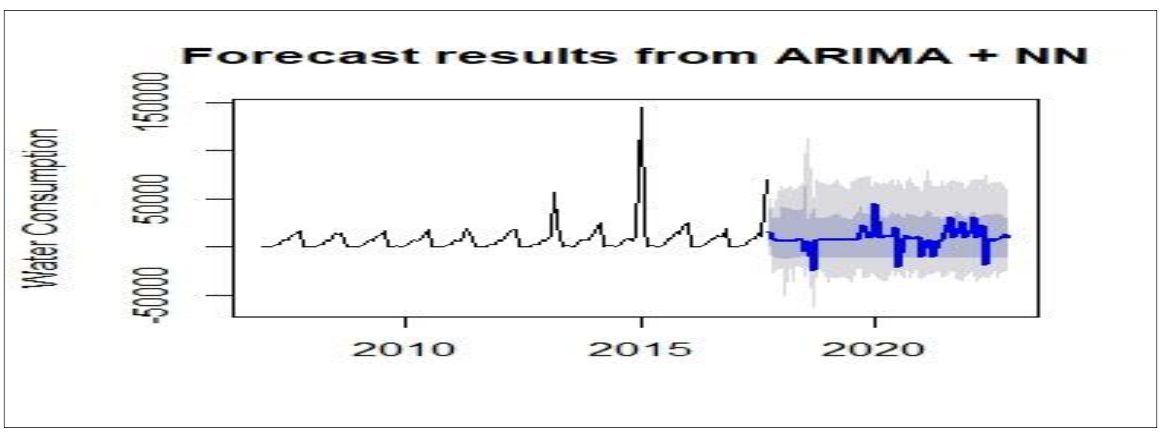

Figure (5-31) Al-Nasr forecasting water consumption results from 2018 to 2022.

15. Al-Jora (جورة اللوت)

Figure (*5-32* shows the forecast results for Al- Jora region between 2018 and 2022, and we found that the water consumption deviation percentage is -12.47%, while the water revenue deviation percentage is -16.90% as shown in **Table 5-8**.

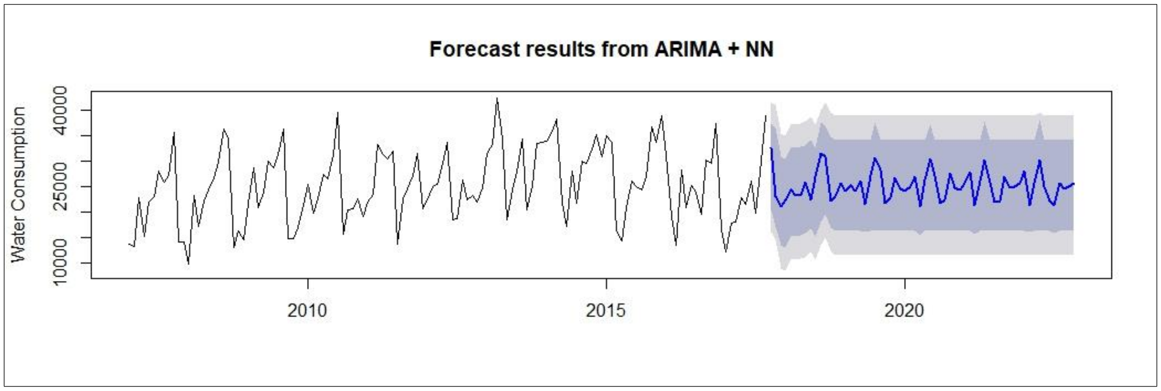

Figure (5-32) Al- Jora forecasting water consumption results from 2018 to 2022.

16. Al-Qreen (قاع القرين)

Figure (*5-33* shows the forecast results for Al- Qreen region between 2018 and 2022, and we found that the water consumption deviation percentage is 12.72%, while the water revenue deviation percentage is 49.23% as shown in **Table 5-8**.

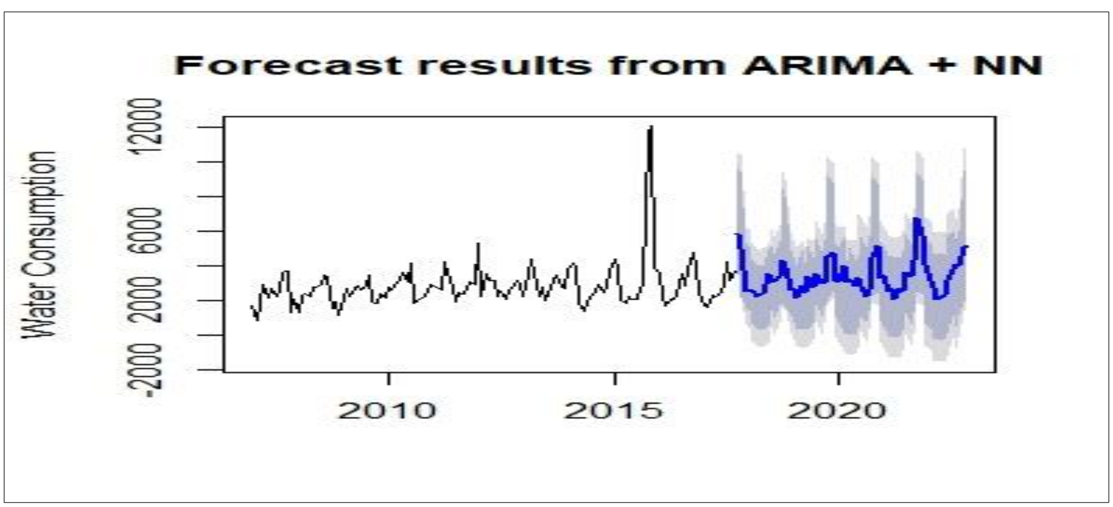

Figure (5-33) Al- Qreen forecasting water consumption results from 2018 to 2022.

17. Al-Rashwan (قيزان ابو رشوان)

Figure (*5-34* shows the forecast results for Al- Rashwan region between 2018 and 2022, and we found that the water consumption deviation percentage is -13.72%, while the water revenue deviation percentage is -8.45% as shown in **Table 5-8**.

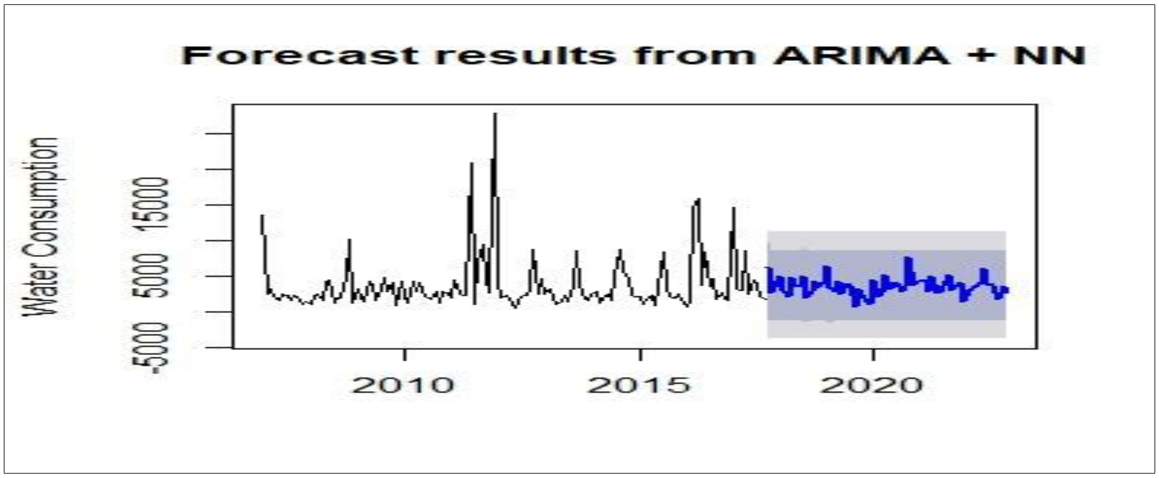

Figure (5-34) Al- Rashwan forecasting water consumption results from 2018 to 2022

18. Al-Najar (قيزان النجار)

Figure (*5-35* shows the forecast results for Al- Najar region between 2018 and 2022, and we found that the water consumption deviation percentage is 7.38%, while the water revenue deviation percentage is -5.97% as shown in **Table 5-8**.

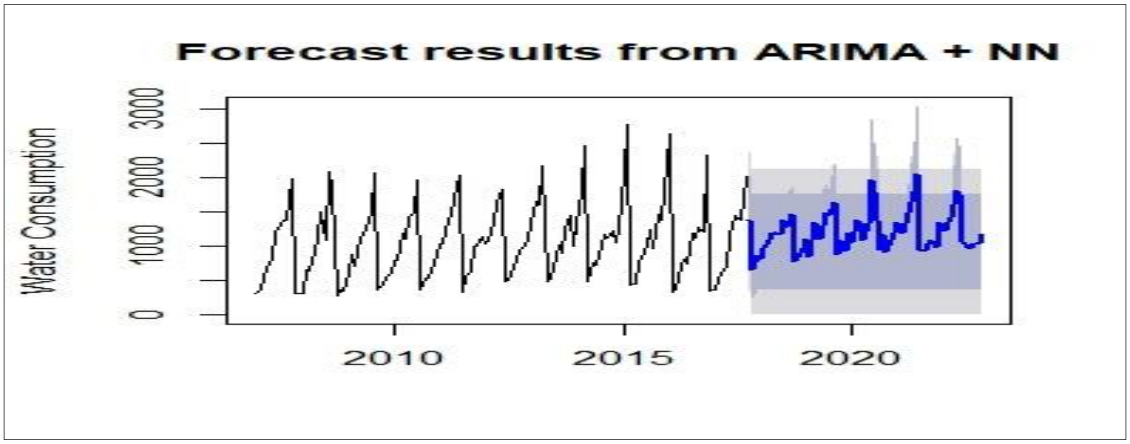

Figure (5-35) Al- Najar forecasting water consumption results from 2018 to 2022

19. City-Center (مركز المدينة)

Figure (*5-36* shows the forecast results for City-Center region between 2018 and 2022, and we found that the water consumption deviation percentage is 10.95%, while the water revenue deviation percentage is 0.76% as shown in **Table 5-8**.

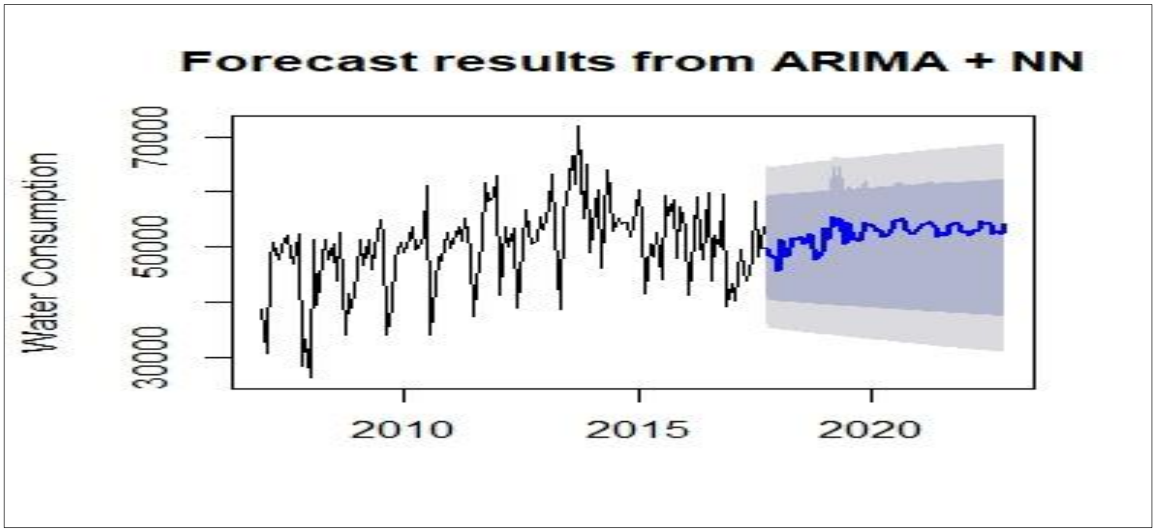

Figure (5-36) City-Center forecasting water consumption results from 2018 to 2022

20. Maan (معن)

Figure (*5-37* shows the forecast results for Maan region between 2018 and 2022, and we found that the water consumption deviation percentage is 5.48%, while the water revenue deviation percentage is -3.57% as shown in **Table 5-8**.

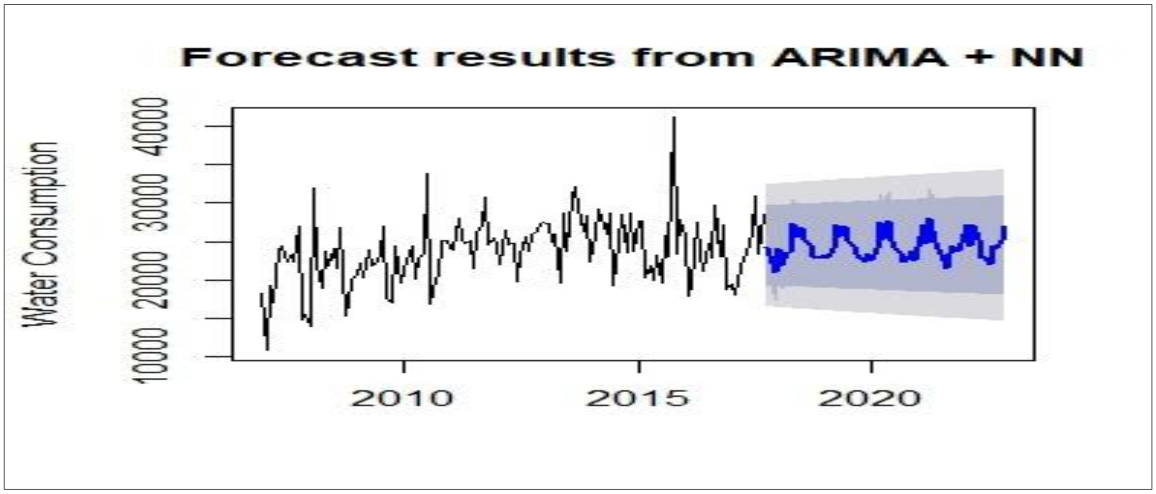

Figure (5-37) Maan forecasting water consumption results from 2018 to 2022

In **Table 5-8**, we present the increase/decrease percentage for the attributes 'Water Consumption' and 'Water Revenue' for every region between 2017 and the maximum forecasted value to 2022.

Table 5-8) Computing the deviation percentage of water consumption and revenue over regions between 2017 and 2022.

Region Name	Water Con. 2017	Water Con. Avg. 2018 - 2022	Deviation %	Water Rev. 2017	Water Rev. Avg. 2018 - 2022	Deviation %
الأمل	603082	635652	5.40	905790	927172	2.36
البطن السمين	486137	472532	-2.80	807531	73193	-4.37
التحرير	10816	11135	2.95	17322	14816	-14.47
الجلاء	32101	19476	-39.33	57489	73193	27.32
السطر	151919	164585	8.34	252040	256310	1.69
السلام	1495	1249	-16.48	4376	4290	-1.96
الشيخ ناصر	295687	298343	0.90	478963	484961	1.25
الكتيبة	266112	284583	6.94	438286	443980	1.30
المحطة	249165	249517	0.14	410338	410770	0.11
المعسكر	858168	908930	5.92	143515 7	1448464	0.93
المنارة	52816	52044	-1.46	89781	85774	-4.46
المواصي الجنوبي	11283	10658	-5.54	19565	16972	-13.25
المواصي الشمالي	81822	85040	3.93	142370	138096	-3.00
النصر	146943	102580	-30.19	285518	326481	14.35
جورت اللوت	345508	302418	-12.47	617950	513496	-16.90
قاع القرين	35552	40075	12.72	60687	90561	49.23

قيزان ابو رشوان	50322	43419	-13.72	50200	45957	-8.45
قيزان النجار	13534	14532	7.38	24307	22857	-5.97
مركز المدينة	579012	642422	10.95	986368	993875	0.76
معن	283621	299166	5.48	443403	427572	-3.57

From **Table 5-8**, we can classify the deviations over regions between year 2017 and the average of forecasted data from 2018 to 2022 into four categories as shown in **Figure (*5-38***. First category, increasing in both water consumption and revenue ($\wedge consumption + \wedge revenue$) included in 8 regions of 20 region which presents 66.7% of water consumption in 2017. Second category, decreasing in both water consumption and revenue ($\vee consumption + \vee revenue$) included in 6 regions of 20 region which presents 20.8% of water consumption in 2017. Third category, increasing in water consumption and decreasing in water revenue ($\wedge consumption + \vee revenue$) included in 4 regions of 20 region which presents 8.6% of water consumption in 2017. Fourth category, decreasing in water consumption and increasing in water revenue ($\vee consumption + \wedge revenue$) included in 2 regions of 20 region which presents 3.9% of water consumption in 2017.

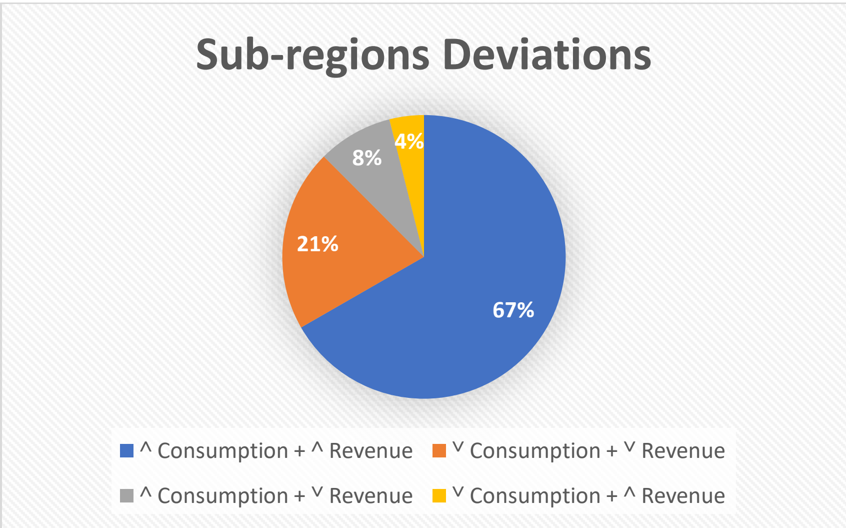

Figure (5-38) Sub-regions forecasted average deviations with comparison with 2017

Also, ***Table 5-8*** shows that the region Al Jalaa'الجلاء' and Al-Nasr have irregular deviation in water consumption average from year 2018 to 2022 in comparison with 2017 with deviation percentage -39.33%, -30.19% respectively and this deviation is high, not accepted and not logical.

5.6. Forecasting classes

As we did in section 5.5 about sub-areas and computing the consumption and revenue percentages over the overall city, we computed the consumption and revenue percentages for classes within KhanYounis city as show in ***Figure (5-39*** and ***Figure (5-40***.

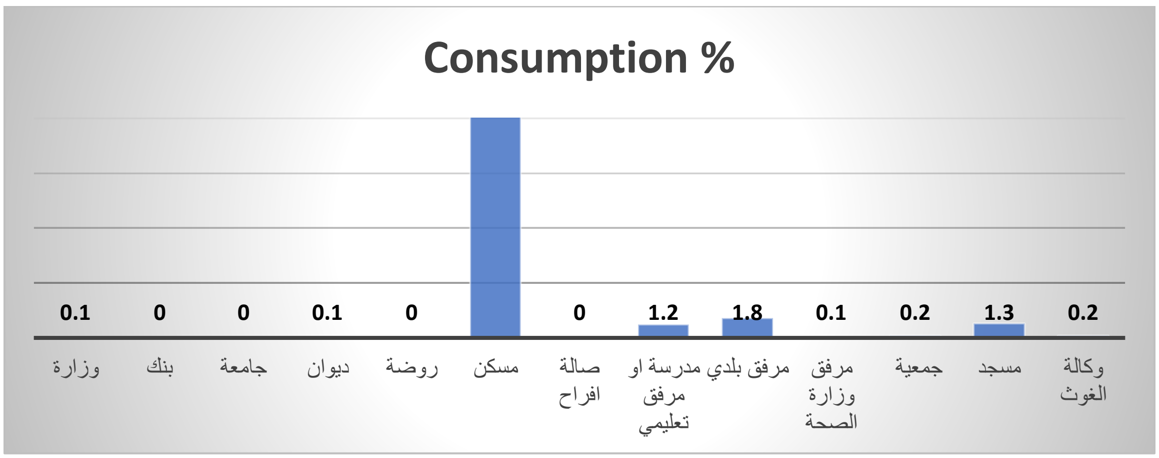

Figure (5-39) Water consumption rates for 13 class in year 2017

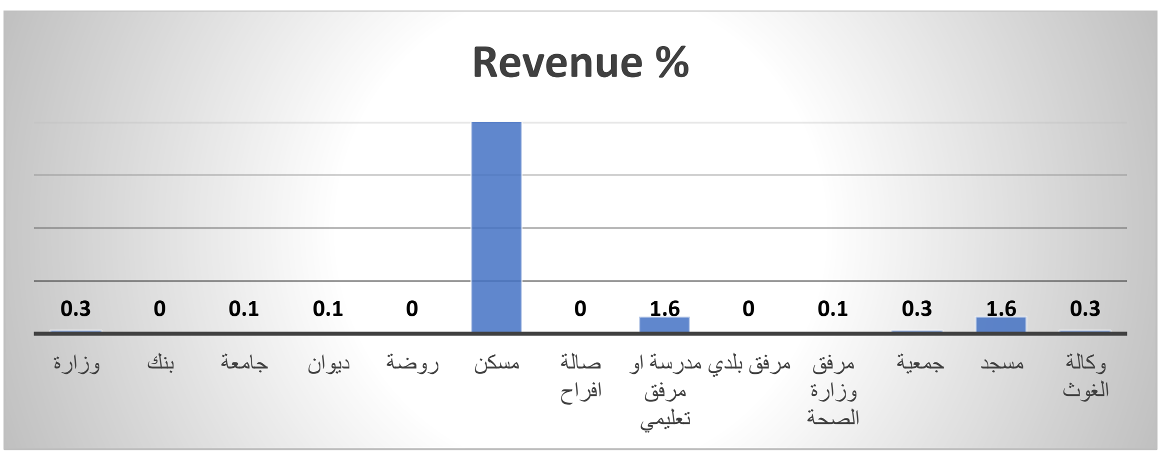

Figure (5-40) Water revenue rates for 13 class in year 2017

Figure (*5-39* and **Figure (*5-40*** shows that houses class weights the highest class in both attributes 'Water Consumption' and 'Water Revenue', while other 9 classes of 13 are weighting almost zero in both attributes 'Water Consumption' and 'Water Revenue'.

After conducting forecasting algorithms over regions, we forecasted classes (for example public building, private buildings, schools, banks, …. etc.) over months, and we will display results of last forecasted months for each class for the attributes 'Water Consumption' and 'Water Revenue' for every class.

Table (5-9) Results of last forecasted months over classes on KhanYounis city.

Region Name	Voucher Date	Forecasted Water Consumption (M^3)	Forecasted Water Revenue (NIS)
وزارة	12-2022	566	1957
بنك	12-2022	118	237
جامعة	12-2022	414	82
ديوان	12-2022	260	227
روضة	12-2022	128	228
مسكن	12-2022	458679	944665
صالة افراح	12-2022	156	258
مدرسة او مرفق تعليمي	12-2022	4165	8725
مرفق بلدي	12-2022	-1640	-25428
مرفق وزارة الصحة	12-2022	-1147	-1998
جمعية	12-2022	1008	2002
مسجد	12-2022	6654	9944
وكالة الغوث	12-2022	6985	20533

From **Table (*5-9*** we can notice that the municipality and Ministry of Health classes are presenting minus forecasted water consumption and revenue and this is not logical, so these two classes will be treated as special cases in section 5.7.

At **Table (*5-10***, we are going to present the deviation percentage for the attributes 'Water Consumption' and 'Water Revenue' for every class between 2017 and 2022.

Table (5-10) Computing the deviation percentage of water consumption and water revenue over classes between 2017 and 2022.

Class Name	Water Cons. 2017	Water Cons. Avg. 2018 - 2022	Deviation %	Water Rev. 2017	Water Rev. Avg. 2018 - 2022	Deviation %
وزارة	6062	7667	26.49	20144	23405	16.19
بنك	1295	1210	-6.55	3092	3105	0.43
جامعة	1293	4585	254.61	3633	2868	-21.04
ديوان	2527	3095	22.48	5201	4210	-19.05
روضة	1414	1419	0.39	2558	4144	62.01
مسكن	5272772	5522580	4.74	9187973	10613666	15.52
صالة افراح	1237	1978	59.91	1936	3083	59.26
مدرسة او مرفق تعليمي	56563	57387	1.46	122319	119566	-2.25
مرفق بلدي	69089	68016	-1.55	0	-95733	-95733
مرفق وزارة الصحة	2809	-6161	-319.32	8805	-9193	-204.41
جمعية	11193	11578	3.44	22948	24302	5.90
مسجد	60287	63979	6.12	127671	122438	-4.10
وكالة الغوث	6609	58600	786.67	18073	166490	821.21

From **Table (*5-10***, we can classify the deviations over classes between year 2017 and the last forecasted year 2022 into four categories as shown in ***Figure (5-41***. First category, increasing in both water consumption and revenue (∧*consumption* + ∧*revenue*) included in 6 regions of 13 region which presents 96.47% of water consumption in 2017. Second category, decreasing in both water consumption and revenue ($\vee consumption + \vee revenue$) included in 2 regions of 13 region which presents 1.31% of water consumption in 2017. Third category, increasing in water consumption and decreasing in water revenue ($\wedge consumption + \vee revenue$) included in 4 regions of 13 region which presents 2.2% of water consumption in 2017. Fourth category, decreasing in water consumption and increasing in water revenue ($\vee consumption + \wedge revenue$) included in one region of 13 region which presents 0.02% of water consumption in 2017.

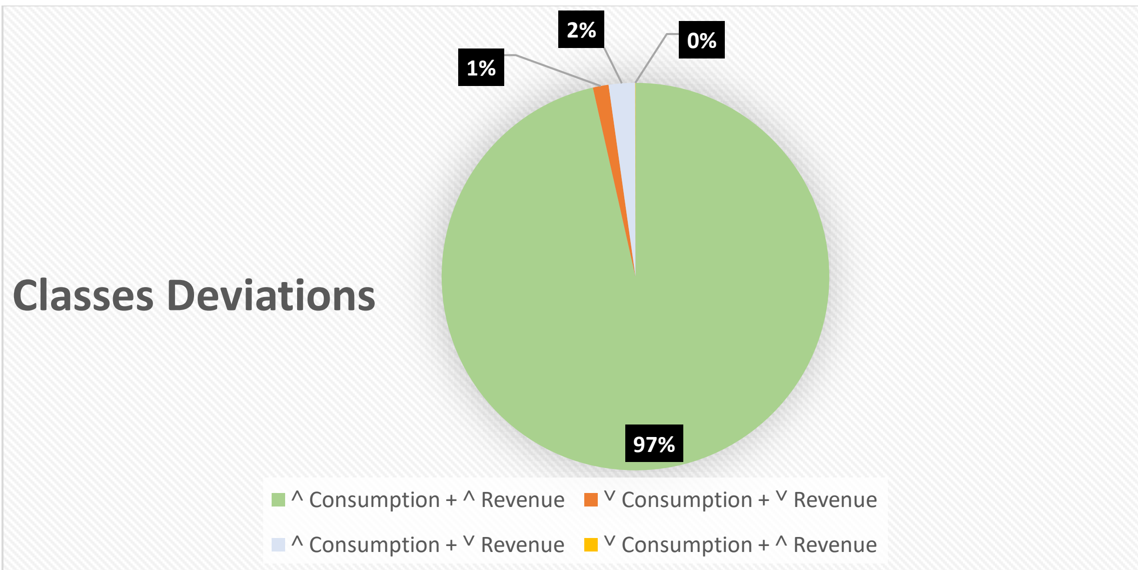

Figure (5-41) Classes forecasted average deviations with comparison with 2017.

5.7. Forecasting special cases

From previous work we found some special cases in both forecasting levels regions and classes, and we treated every special case separately in the following sections.

5.7.1. Forecasting special cases in regions level

The normal process of reading water meters is periodically once per month for all units, but in some cases and some sub areas, this process is done in such month once per year, which lets the water consumption for 11 months of the year have large different water consumption readings compared with this single month, this issue is said also about water revenue, which lets our data sets inconsistent like the rest of other sub areas. Due to these issues in collecting and reading the monthly consumption water values to compute the water revenues in two sub areas in KhanYounis city (Al-Nasr, Al-Jalaa) which presents only 3.7% of the overall city water consumption as shown in ***Figure (5-16***, we had to forecast these areas yearly not monthly. So, our original data for these two sub-area consists only of eleven records, one record for every year from 2007 to 2017, and the forecasting horizon equals 5 which means until 2022.

Our original data as we described before starts from 1-2007 to 10-2017, and to have complete 11 year of data set, we must complete two missing months in year 2017, so to handle these two missing months November and December in year 2017, we calculated them for both water consumption and water revenue by computing the average Novembers and Decembers of the last three years (2016,2015,2014).

5.7.1.1. Al-Jalaa

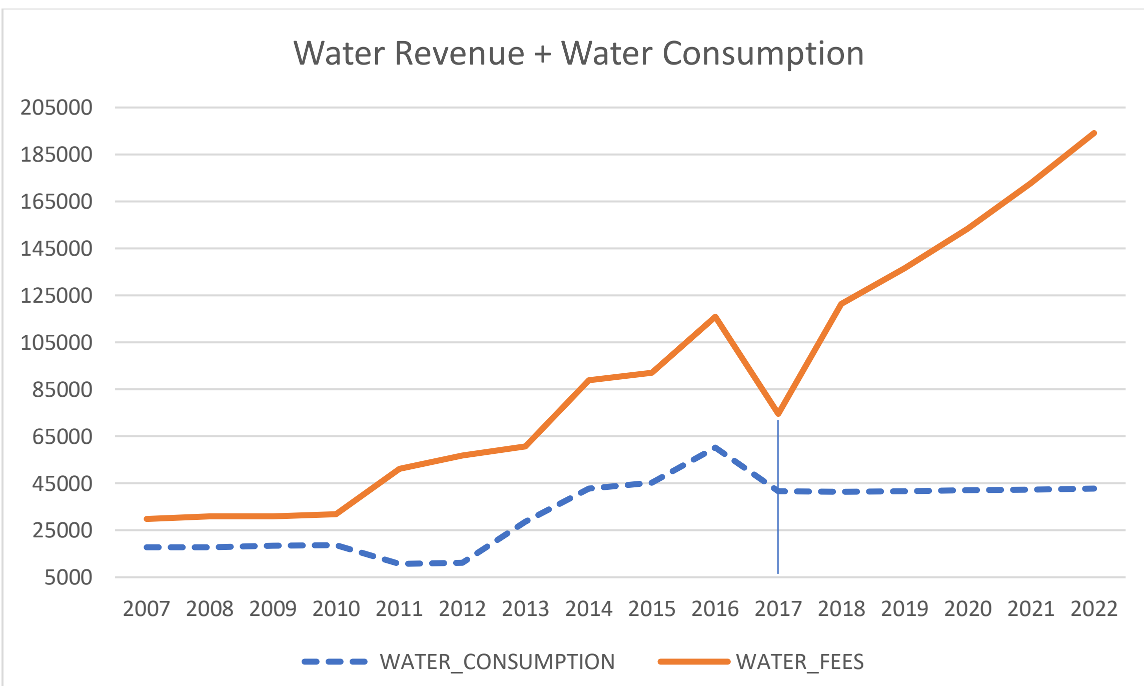

Figure (5-42) Al-Jalaa forecasting water consumption and revenue results from 2018 to 2022.

5.7.1.2. Al-Nasr

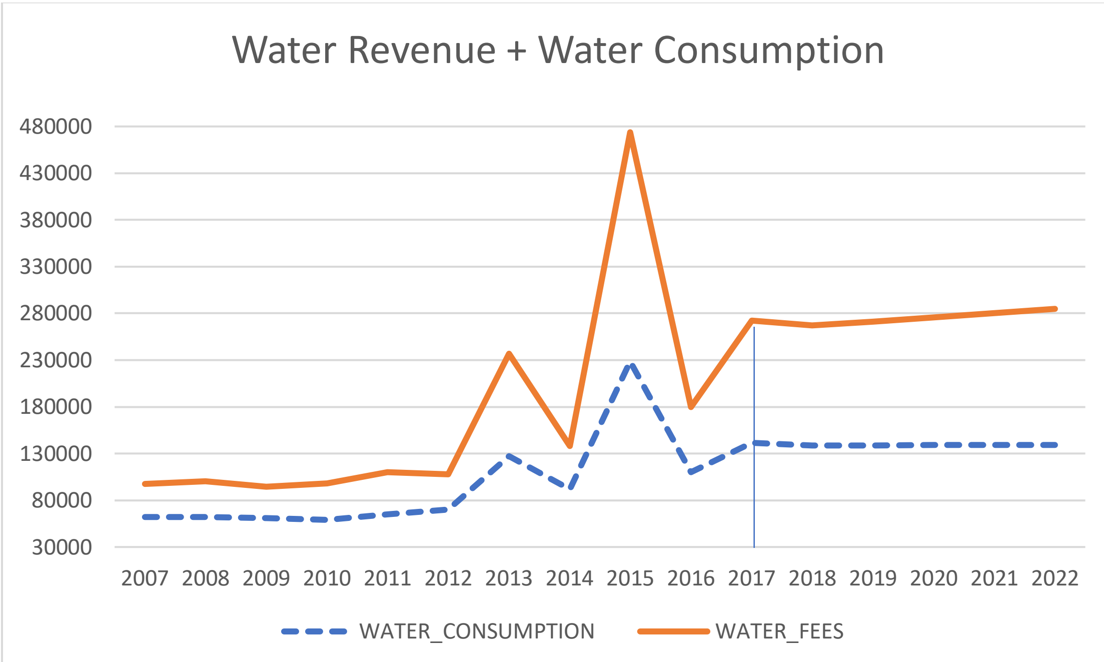

Figure (5-43) Al-Nasr forecasting water consumption and revenue results from 2018 to 2022.

5.7.2. Forecasting special cases in classes level

In this part we have two special cases in classes level, these two classes are Municipality buildings and Ministry of Health building which present only 1.9% of the overall city water consumption in 2017 as shown in ***Figure (5-39***.

5.7.2.1. Municipality buildings

Because KhanYounis municipality is the only institute responsible for providing and serving water services in KhanYounis city, it stopped account itself since 3-2009 for what water consumption it uses in all its distributed building. So, this point let this class contains 80% of last part of its dataset for the attribute 'Water Revenue' is zeros, and automatically the forecasting will results in also zeros for the next five year. That's why we will not include this class in our forecasting tasks for the next five years at this level.

5.7.2.2. Ministry of Health building(MOH)

In this special case to forecast the next five years, the original dataset presented as yearly values not monthly values as we did before in special cases in regions. As we did before in filling the two months missed in the original data to have complete 11 years we calculated November and December in 2017 for both water consumption and water revenue by computing the average Novembers and Decembers of the last three years (2016,2015,2014) for both attributes.

This dramaturgically decrease in water consumption and water revenue from 2007 to 2017 due to most of MOH buildings are depending heavily on their own water resources like ground wells and water desalination plants.

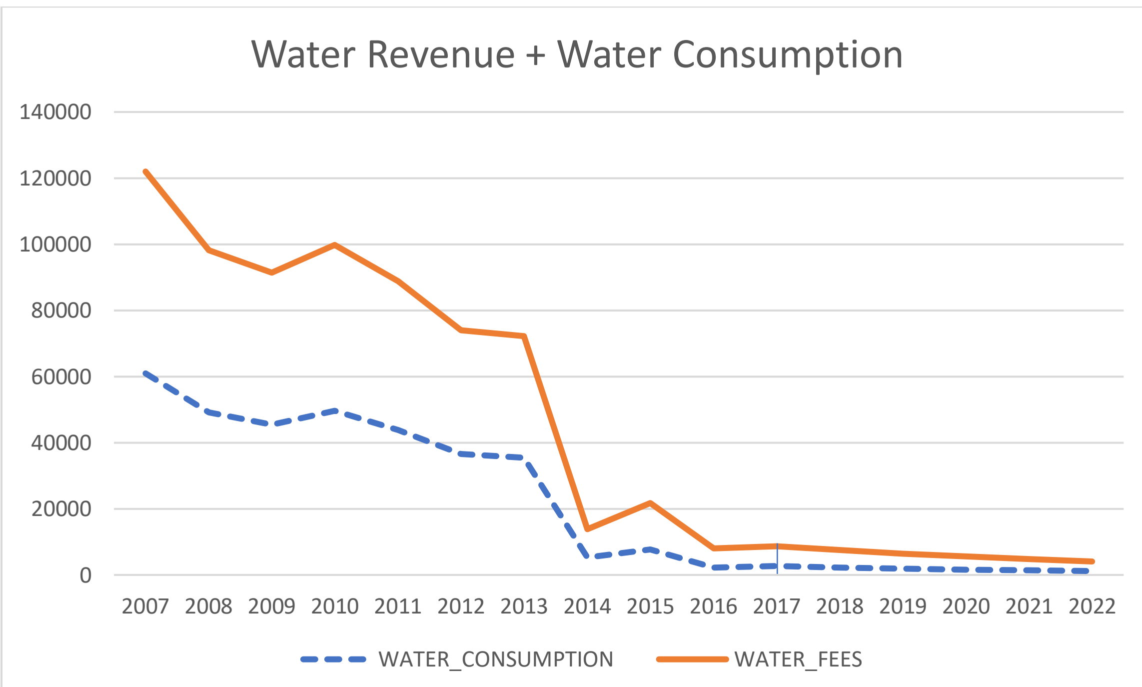

Figure (5-44) MOH forecasting water consumption and revenue results from 2018 to 2022.

From **Figure *(5-44*** we can conclude that water consumption and revenue for class MOH in the next five years is going to almost zero.

5.8. Discussion

From **Figure (*5-14*** and **Figure (*5-15***, we observed that water consumptions for the next five years will increase, and this is logical result, because the population is increasing by the time which means more people will demand water for basics needs and day life use. But on the other hand, water revenues will decrease for the next five years, and this contrary with the previous observation about water consumption, while the nature relationship between water consumption and water revenue is positive relationship, which means when water demand increases, water revenues also should increase and vice versa. But water revenues are decreasing maybe due to several reasons, (1) The suffocating economic crisis, which affecting directly on people financial status, which lets them not pay for the water voucher produced by KhanYounis municipality, (2) blockage from sea, land and air which effect on the society projects and income, (3) Political differences and their impact on employees' salaries, so much so that a class of employees received 40% of the salary for a long time, (4) KhanYounis municipality do not count the quantities of water it's buildings consume and which definitely affect their revenues in return. All these reasons and other effects on the municipality revenues from water service.

5.9. Summary

In this chapter three levels of forecasting are done, 1st overall city forecasting, 2nd sub-areas forecasting, 3rd classes forecasting. We noticed that the overall city forecasting results, in general, for the next 5 years water consumption will increase with 8.4% (1.7% per year) but the water revenue will be decreased with 3.8% (0.76% per year).

Forecasting sub-areas and classes are done monthly from 1-2007 to 12-2022, but there were some special cases in both categories were treated yearly from 2007 to 2022, except the class 'Municipality class', because it's the organization which provides the water service, so it stopped counting the water amounts it consumes.

Chapter 6
Conclusion and Future Works

Chapter 6

Conclusion and Future Works

This chapter summarizes the thesis results and presents the major important points discovered while conducting our experiments.

6.1. Summary

Four of the most famous forecasting algorithms (ARIMA, Hybrid ARIMA, Linear regression and SSA), are used widely in time series forecasting tasks and many researches and papers enriched their performance and business usage in real life applications. It was very clearly observed that no one applied any forecasting task on water consumption or water revenue in any Palestinian area and none have discussed the relation between water consumption and water revenue, is it positive relationship or not?

We applied forecasting algorithms on KhanYounis city dataset and try to present results in relation views as shown in ***Figure (6-1***. Forecasting algorithms applied on water demand to predict the future of water needs by the growth of population with the time passage. Also, we forecasted the revenue and profits from water service by the growth of population with the time passage.

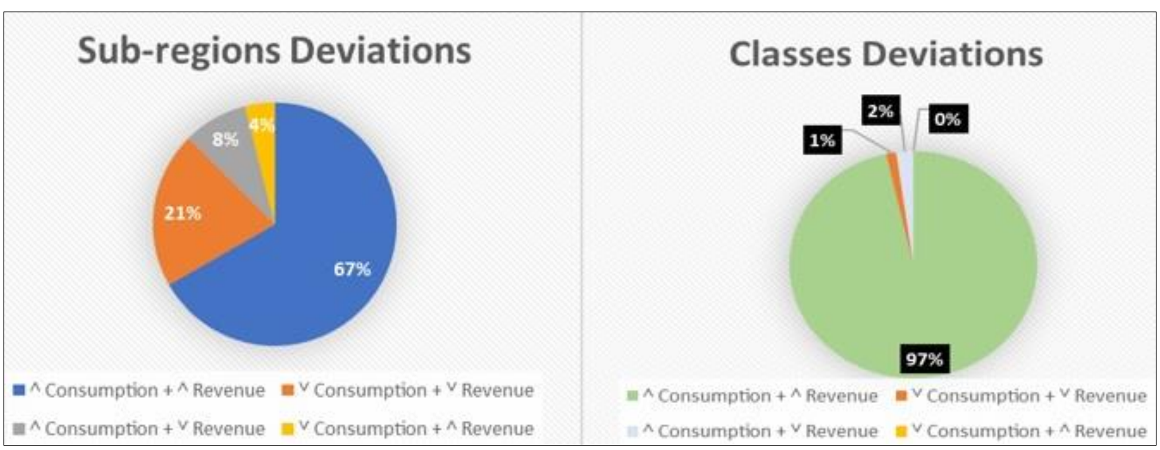

Figure (6-1) Relational views for forecasted results for sub-regions and classes deviation

Three levels of forecasting have been conducted, 1st overall city forecasting, 2nd sub-areas forecasting, 3rd classes forecasting. For the next 5 years water consumption for the overall city will increase with 8.4% (1.7% per year) but on the other hand, the water revenue will be decreased with 3.8% (0.76% per year).

Forecasting sub-areas and classes are done monthly from 1-2007 to 12-2022, but there were some special cases in both categories were treated yearly from 2007 to 2022.

6.2. Conclusion

The main concern of this research study was to evaluate the forecasting algorithms over our datasets, then choose the most appropriate one to forecast 5 years in advance.

From all discussions of this study the following conclusions can be drawn:

1. Data of water consumption and revenue is non-stationary time series.
2. The model cover ~11 years (130 months) of customer's historical water consumptions and revenue data.
3. Our datasets used in the research are (1) the whole city dataset, (2) sub-areas dataset, (3) classes dataset, and every dataset starts from 1-2007 to 10-2017.
4. The best model for water consumption after treatment non-stationary is Hybrid ARIMA, this was supported by the MPE 17.38%.
5. The best algorithm for forecasting water revenue after treatment non-stationary is also Hybrid ARIMA, this was supported by the MPE 21.57%.
6. The linear ARIMA model and the nonlinear ANN model are used jointly, aiming to capture different forms of relationship in the time series data.
7. Generally, for the whole city we can say that after five years water consumption will approximately increase with 8.4% but the water revenue will approximately decrease with 3.8%.

8. Forecast is done in three level, 1st level is the whole city, 2nd level is the sub-areas and the 3rd level is the classes inside KhanYounis city.
9. Some special cases in two levels were treated as yearly not monthly forecast due to the original data is yearly because of preprocessing tasks.

6.3. Future works

As future work, we can use long term forecasting approaches to predict the water consumption and water revenue for more than 10 years. Also, we can forecast data using more variables and attributes such as population.

More deep learning models can be used specially to solve our problem or achieve our goal. Also, other resources can be forecasted in KhanYounis municipality KHM.

This model could be applied for other municipalities datasets to forecast water demand and water revenue or any other resources available. Finally, we can use forecasting methods to predict individual water bills.

Bibliography

Abuella, M., & Chowdhury, B. (2015). *Solar power probabilistic forecasting by using multiple linear regression analysis.* Paper presented at the SoutheastCon 2015.

Aggarwal, C. C. (2015). *Data mining: the textbook*: Springer.

Ajbar, A., & Ali, E. M. (2015). Prediction of municipal water production in touristic Mecca City in Saudi Arabia using neural networks. *Journal of King Saud University-Engineering Sciences, 27*(1), 83-91.

Alvarado-Valencia, J., Barrero, L. H., Önkal, D., & Dennerlein, J. T. (2017). Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. *International Journal of Forecasting, 33*(1), 298-313.

Azam, F. (2000). *Biologically inspired modular neural networks.*

Bails, D., & Peppers, L. C. (1993). *Business fluctuations: forecasting techniques and applications*: Prentice Hall.

Bakker, M., Van Duist, H., Van Schagen, K., Vreeburg, J., & Rietveld, L. (2014). Improving the performance of water demand forecasting models by using weather input. *Procedia Engineering, 70*, 93-102.

Barakat, R. (2013). *Water in Palestine*.

Bianco, V., Manca, O., & Nardini, S. (2009). Electricity consumption forecasting in Italy using linear regression models. *Energy, 34*(9), 1413-1421.

Billings, R. B., & Jones, C. V. (2011). *Forecasting urban water demand*: American Water Works Association.

Bitzer, S., & Kiebel, S. J. (2012). Recognizing recurrent neural networks (rRNN): Bayesian inference for recurrent neural networks. *Biological cybernetics*, 1-17.

Bougadis, J., Adamowski, K., & Diduch, R. (2005). Short-term municipal water demand forecasting. *Hydrological Processes, 19*(1), 137-148.

Broomhead, D. S., & King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena, 20*(2-3), 217-236.

Ceballos, A., & Walke, A. G. (2016). Short-Term Forecasting Analysis for Municipal Water Demand. *Journal: American Water Works Association, 108*(1).

Chen, J., & Boccelli, D. (2014). Demand forecasting for water distribution systems. *Procedia Engineering, 70*, 339-342.

Delen, D., & Olson, D. (2008). Advanced data mining techniques. *Springer. doi, 10*, 9780470172339.

Donkor, E. A., Mazzuchi, T. A., Soyer, R., & Alan Roberson, J. (2012). Urban water demand forecasting: review of methods and models. *Journal of Water Resources Planning and Management, 140*(2), 146-159.

Elsner, J. B., & Tsonis, A. A. (2013). *Singular spectrum analysis: a new tool in time series analysis*: Springer Science & Business Media.

EWASH. (2016). WATER AND SANITATION IN PALESTINE. *EWASH*, 2.

Fatima, S., & Hussain, G. (2008). Statistical models of KSE100 index using hybrid financial systems. *Neurocomputing, 71*(13), 2742-2746.

Golyandina, N., Nekrutkin, V., & Solntsev, V. (2001). „Caterpillar‟-SSA Technique for Analysis of Time Series in Economics. *New Models of Business: Managerial Aspects and Enabling Technology*, 198.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Hassani, H. (2007). Singular spectrum analysis: methodology and comparison.

Hassani, H., Webster, A., Silva, E. S., & Heravi, S. (2015). Forecasting US tourist arrivals using optimal singular spectrum analysis. *Tourism Management, 46*, 322-335.

Heshmaty, B., & Kandel, A. (1985). Fuzzy linear regression and its applications to forecasting in uncertain environment. *Fuzzy sets and systems, 15*(2), 159-191.

Humaid, E. H. S. (2012). *A Data Mining Based Fraud Detection Model for Water Consumption Billing System in MOG.* Islamic University of Gaza.

Iqelan, B. M. (2017). A Singular Spectrum Analysis Technique to Electricity Consumption Forecasting. *Int. Journal of Engineering Research and Application*, 9.

Kanchymalay, K., Salim, N., Sukprasert, A., Krishnan, R., & Hashim, U. R. a. (2017). *Multivariate Time Series Forecasting of Crude Palm Oil Price Using Machine Learning Techniques.* Paper presented at the IOP Conference Series: Materials Science and Engineering.

Khashei, M., & Bijari, M. (2011). A new hybrid methodology for nonlinear time series forecasting. *Modelling and Simulation in Engineering, 2011*, 15.

Khashei, M., Bijari, M., & Ardali, G. A. R. (2009). Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (ANNs). *Neurocomputing, 72*(4), 956-967.

Kumar, U., & Jain, V. (2010). Time series models (Grey-Markov, Grey Model with rolling mechanism and singular spectrum analysis) to forecast energy consumption in India. *Energy, 35*(4), 1709-1716.

Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism management, 59*, 57-66.

Liu, Y., Zhao, J., & Wang, Z. (2015). Identifying determinants of urban water use using data mining approach. *Urban Water Journal, 12*(8), 618-630.

Mills, T. C. (2011). *The foundations of modern time series analysis*: Springer.

Nosedal, A. (2011). Time Series Analysis and Forecasting.

Nury, A. H., Hasan, K., & Alam, M. J. B. (2015). Comparative study of wavelet-ARIMA and wavelet-ANN models for temperature time series data in northeastern Bangladesh. *Journal of King Saud University-Science*.

Oracle. (2016). Data Mining Concepts. 9.

Ozyasar, H. (2011). Application of Regression Analysis in Business.

Pacchin, E., Alvisi, S., & Franchini, M. (2017). A Short-Term Water Demand Forecasting Model Using a Moving Window on Previously Observed Data. *Water, 9*(3), 172.

Pai, P.-F., & Lin, C.-S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega, 33*(6), 497-505.

Pandey, B., Jain, T., Kothari, V., & Grover, T. (2012). Evolutionary Modular Neural Network Approach for Breast Cancer Diagnosis. *International Journal of Computer Science Issues, 9*(1), 219-225.

PCBS. (2016). Population increase rate in KhanYounis city

Polebitski, A. S., & Palmer, R. N. (2009). Seasonal residential water demand forecasting for census tracts. *Journal of Water Resources Planning and Management, 136*(1), 27-36.

Rinaudo, J.-D. (2015). Long-term water demand forecasting *Understanding and Managing Urban Water in Transition* (pp. 239-268): Springer

Ripley, R. (2015). Why Resource Forecasting is Important: Preparing for Projects. *workamajig*.

RISH, R. M. A. A. (2015). Electricity Consumption Forecasting in the Khan Younis Province

Using Exponential Smoothing and Box - Jenkins Methods

: A Modeling Viewpoint. 95.

Rout, M., Majhi, B., Majhi, R., & Panda, G. (2014). Forecasting of currency exchange rates using an adaptive ARMA model with differential evolution based training. *Journal of King Saud University-Computer and Information Sciences, 26*(1), 7-18.

Shabani, S., Yousefi, P., & Naser, G. (2017). Support Vector Machines in Urban Water Demand Forecasting Using Phase Space Reconstruction. *Procedia Engineering, 186*, 537-543.

Shen, A., Tong, R., & Deng, Y. (2007). *Application of classification models on credit card fraud detection.* Paper presented at the Service Systems and Service Management, 2007 International Conference on.

Sibanda, W., & Pretorius, P. (2012). Artificial neural networks-a review of applications of neural networks in the modeling of hiv epidemic. *International Journal of Computer Applications, 44*(16), 1-9.

So, M. K., & Chung, R. S. (2014). Dynamic seasonality in time series. *Computational Statistics & Data Analysis, 70*, 212-226.

Solutions, S. (2013). What is Linear Regression.

Song, K.-B., Baek, Y.-S., Hong, D. H., & Jang, G. (2005). Short-term load forecasting for the holidays using fuzzy linear regression method. *IEEE transactions on power systems, 20*(1), 96-101.

Steel, A. (2014). *Predictions in Financial Time Series Data.* Institute of Technology Blanchardstown.

Su, H., Li, W., Wang, K., & Ding, X. (2011). Stability analysis for stochastic neural network with infinite delay. *Neurocomputing, 74*(10), 1535-1540.

Tan, Z., Zhang, J., Wang, J., & Xu, J. (2010). Day-ahead electricity price forecasting using wavelet transform combined with ARIMA and GARCH models. *Applied Energy, 87*(11), 3606-3610.

Tanaka, H., & Asai, K. (1984). Fuzzy linear programming problems with fuzzy numbers. *Fuzzy sets and systems, 13*(1), 1-10.

Wang, X., & Meng, M. (2012). A Hybrid Neural Network and ARIMA Model for Energy Consumption Forcasting. *JCP, 7*(5), 1184-1190.

Week, B. (1982). Artificial Intelligence: The second computer age begins. *March, 8*, 66-75.

Wen, Y. Y., Huang, W. M., Wu, J., Chen, Y., & Song, J. Q. (2013). *Water consumption analysis system based on data mining.* Paper presented at the Applied Mechanics and Materials.

Wongsathan, R., & Seedadan, I. (2016). A Hybrid ARIMA and Neural Networks Model for PM-10 Pollution Estimation: The Case of Chiang Mai City Moat Area. *Procedia Computer Science, 86*, 273-276.

Yu, Z., Lei, G., Jiang, Z., & Liu, F. (2017). *ARIMA modelling and forecasting of water level in the middle reach of the Yangtze River.* Paper presented at the Transportation Information and Safety (ICTIS), 2017 4th International Conference on.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159-175.

Appendix

Summary of Results

Appendix A

Computing MPE for ARIMA algorithm over water consumption (M^3).

VOUCHER DATE	WATER CONSUMPTION (Y)	ARIMA Forecasted WATER CONSUMPTION (Y')	ERROR ABS (Y - Y'))/ Y
01-2016	459852	435254.6	0.05
02-2016	455574	430740.9	0.05
03-2016	264750	417487.3	0.58
04-2016	301128	403983.0	0.34
05-2016	388855	405325.4	0.04
06-2016	351510	431101.2	0.23
07-2016	369680	428520.7	0.16
08-2016	422076	435797.5	0.03
09-2016	361463	440333.7	0.22
10-2016	415381	444049.7	0.07
11-2016	387266	438466.4	0.13
12-2016	474598	431798.0	0.09
01-2017	255514	430714.8	0.69
02-2017	287409	429039.1	0.49
03-2017	281615	429350.2	0.52
04-2017	338134	431387.2	0.28
05-2017	350639	436130.1	0.24
06-2017	356330	439843.8	0.23
07-2017	401674	441812.1	0.10
08-2017	549701	442874.4	0.19
09-2017	377772	442823.0	0.17
10-2017	494443	441899.5	0.11
		MPE	**22.85**

Appendix B

Computing MPE for Hybrid ARIMA algorithm over water consumption (M^3)..

VOUCHER DATE	WATER CONSUMPTION (Y)	Hybrid ARIMA Forecasted WATER CONSUMPTION (Y')	ERROR ABS (Y - Y'))/ Y
01-2016	459852	441960.6	0.04
02-2016	455574	426307.4	0.06
03-2016	264750	358440.7	0.35
04-2016	301128	392045.4	0.30
05-2016	388855	387966.9	0.00
06-2016	351510	411199.6	0.17
07-2016	369680	408528.8	0.11
08-2016	422076	429077.9	0.02
09-2016	361463	443861.9	0.23
10-2016	415381	431147.9	0.04
11-2016	387266	407437.1	0.05
12-2016	474598	442028.5	0.07
01-2017	255514	441903	0.73
02-2017	287409	381865.3	0.33
03-2017	281615	388087.4	0.38
04-2017	338134	390173.8	0.15
05-2017	350639	380717.7	0.09
06-2017	356330	401033.5	0.13
07-2017	401674	422395.4	0.05
08-2017	549701	430837.4	0.22
09-2017	377772	436651.7	0.16
10-2017	494443	416034.3	0.16
		MPE	**17.38**

Appendix C

Computing MPE for SSA algorithm over water consumption (M^3)..

VOUCHER DATE	WATER CONSUMPTION (Y)	SSA Forecasted WATER CONSUMPTION (Y')	ERROR ABS (Y - Y'))/ Y
01-2016	459852	569425.7	0.24
02-2016	455574	493849.6	0.08
03-2016	264750	339470.9	0.28
04-2016	301128	432129.2	0.44
05-2016	388855	426663.1	0.10
06-2016	351510	420844.4	0.20
07-2016	369680	436311.8	0.18
08-2016	422076	474900.5	0.13
09-2016	361463	485748.9	0.34
10-2016	415381	484235.3	0.17
11-2016	387266	401520.4	0.04
12-2016	474598	611258.5	0.29
01-2017	255514	520468.6	1.04
02-2017	287409	353257.1	0.23
03-2017	281615	427691.4	0.52
04-2017	338134	446934.8	0.32
05-2017	350639	398030.7	0.14
06-2017	356330	453102.0	0.27
07-2017	401674	467762.4	0.16
08-2017	549701	498155.5	0.09
09-2017	377772	499123.1	0.32
10-2017	494443	410855.1	0.17
		MPE	**26.07**

Appendix D

Computing MPE for Liner Regression algorithm over water consumption (M^3)..

VOUCHER DATE	WATER CONSUMPTION (Y)	Linear Regression Forecasted WATER CONSUMPTION (Y')	ERROR ABS (Y - Y'))/ Y
01-2016	459852	598391.4	0.30
02-2016	455574	452009.2	0.01
03-2016	264750	350461.8	0.32
04-2016	301128	263132.9	0.13
05-2016	388855	326810.8	0.16
06-2016	351510	312964.0	0.11
07-2016	369680	335091.6	0.09
08-2016	422076	384898.3	0.09
09-2016	361463	466815.0	0.29
10-2016	415381	510465.4	0.23
11-2016	387266	444923.5	0.15
12-2016	474598	576553.5	0.21
01-2017	255514	443132.2	0.73
02-2017	287409	353356.1	0.23
03-2017	281615	204802.4	0.27
04-2017	338134	267330.7	0.21
05-2017	350639	224157.3	0.36
06-2017	356330	299381.9	0.16
07-2017	401674	321403.5	0.20
08-2017	549701	496863.3	0.10
09-2017	377772	508797.6	0.35
10-2017	494443	513858.6	0.04
		MPE	**18.8**

Appendix E

Computing MPE for ARIMA algorithm over water revenue (NIS).

VOUCHER DATE	WATER REVENUE (Y)	ARIMA Forecasted WATER REVENUE (Y')	ERROR ABS (Y - Y'))/ Y
01-2016	773740.5	743017.7	0.04
02-2016	777848.5	724952.5	0.07
03-2016	472531	702612.3	0.49
04-2016	488146.9	676432.6	0.39
05-2016	552890.7	706113.3	0.28
06-2016	510060.9	702522.6	0.38
07-2016	536841.6	723984.1	0.35
08-2016	581807.7	734581.0	0.26
09-2016	499719.4	747196.9	0.50
10-2016	672519.5	751840.1	0.12
11-2016	661813.5	749936.2	0.13
12-2016	808512	750804.5	0.07
01-2017	452750	748404.2	0.65
02-2017	461185.9	750039.9	0.63
03-2017	436950	752430.7	0.72
04-2017	508122.4	757850.5	0.49
05-2017	499459.7	763993.8	0.53
06-2017	505578.3	769914.1	0.52
07-2017	654230	775509.0	0.19
08-2017	1001983.4	779786.0	0.22
09-2017	656727	783444.6	0.19
10-2017	880122	441899.5	0.50
		MPE	**35.03**

Appendix F

Computing MPE for Hybrid ARIMA algorithm over water revenue (NIS).

VOUCHER DATE	WATER REVENUE (Y)	Hybrid ARIMA Forecasted WATER REVENUE (Y')	ERROR ABS (Y - Y'))/ Y
01-2016	773740.5	668558.2	0.14
02-2016	777848.5	602046	0.23
03-2016	472531	605297	0.28
04-2016	488146.9	576305.5	0.18
05-2016	552890.7	613804.1	0.11
06-2016	510060.9	603700.1	0.18
07-2016	536841.6	645224.5	0.20
08-2016	581807.7	659114.9	0.13
09-2016	499719.4	727304.1	0.46
10-2016	672519.5	706948.4	0.05
11-2016	661813.5	808012.1	0.22
12-2016	808512	723624.4	0.10
01-2017	452750	617408.7	0.36
02-2017	461185.9	627035.9	0.36
03-2017	436950	607504.6	0.39
04-2017	508122.4	624280.7	0.23
05-2017	499459.7	640574.8	0.28
06-2017	505578.3	657063.3	0.30
07-2017	654230	664214.7	0.02
08-2017	1001983.4	736634.9	0.26
09-2017	656727	783255.8	0.19
10-2017	880122	824803.4	0.06
		MPE	**21.57**

Appendix G

Computing MPE for SSA algorithm over water revenue (NIS).

VOUCHER DATE	WATER REVENUE (Y)	SSA Forecasted WATER REVENUE (Y')	ERROR ABS (Y - Y'))/ Y
01-2016	773740.5	1009871.00	0.31
02-2016	777848.5	862609.70	0.11
03-2016	472531	626941.00	0.33
04-2016	488146.9	646824.00	0.33
05-2016	552890.7	590844.80	0.07
06-2016	510060.9	618845.90	0.21
07-2016	536841.6	622761.30	0.16
08-2016	581807.7	633077.80	0.09
09-2016	499719.4	668093.80	0.34
10-2016	672519.5	856688.80	0.27
11-2016	661813.5	692501.80	0.05
12-2016	808512	1059995.00	0.31
01-2017	452750	893950.50	0.97
02-2017	461185.9	666129.90	0.44
03-2017	436950	589633.40	0.35
04-2017	508122.4	623804.30	0.23
05-2017	499459.7	575569.60	0.15
06-2017	505578.3	651175.30	0.29
07-2017	654230	580066.10	0.11
08-2017	1001983.4	676566.80	0.32
09-2017	656727	902254.10	0.37
10-2017	880122	715729.20	0.19
		MPE	**27.27**

Appendix H

Computing MPE for Liner Regression algorithm over water revenue (NIS).

VOUCHER DATE	WATER REVENUE (Y)	Linear Regression Forecasted WATER REVENUE (Y')	ERROR ABS (Y - Y'))/ Y
01-2016	773740.5	1127998.7	0.46
02-2016	777848.5	821095.2	0.06
03-2016	472531	621350.1	0.31
04-2016	488146.9	502419.7	0.03
05-2016	552890.7	500477.7	0.09
06-2016	510060.9	549162.4	0.08
07-2016	536841.6	482106.6	0.10
08-2016	581807.7	528315.7	0.09
09-2016	499719.4	626271.1	0.25
10-2016	672519.5	947226.9	0.41
11-2016	661813.5	819252.5	0.24
12-2016	808512	1173982.8	0.45
01-2017	452750	840635.1	0.86
02-2017	461185.9	697627.6	0.51
03-2017	436950	488733.0	0.12
04-2017	508122.4	467037.1	0.08
05-2017	499459.7	485610.1	0.03
06-2017	505578.3	428945.8	0.15
07-2017	654230	452845.5	0.31
08-2017	1001983.4	652507.3	0.35
09-2017	656727	1037094.5	0.58
10-2017	880122	976970.0	0.11
		MPE	**25.77**

Appendix I

Results of forecasting ‘water consumption’ and ‘water revenue’ for KhanYounis city using Hybrid ARIMA.

Voucher Date	Forecasted Water Consumption (M^3)	Forecasted Water Revenue (NIS)
11-2017	405736.6	757069
12-2017	357225.5	598151.6
01-2018	351279	595841.3
02-2018	378950.6	615910.3
03-2018	395127.8	612504.9
04-2018	377133.3	600204.1
05-2018	397295.2	627884.3
06-2018	416395	623408.3
07-2018	397062.6	697137.9
08-2018	393847.8	637310.3
09-2018	419874.2	761340.2
10-2018	403216.4	701306.3
11-2018	345551.3	566954
12-2018	369660	588976.9
01-2019	390635.6	594117.5
02-2019	391809.4	588963.9
03-2019	396134.7	586746.4
04-2019	416500.7	622313.7
05-2019	414578	591040.4
06-2019	407552.4	663859.9
07-2019	413369	628801.2
08-2019	452861.4	738267
09-2019	423318.4	641408
10-2019	342041.7	555886
11-2019	404363.4	585272.9
12-2019	376502.9	577765.3
01-2020	418973.6	598520
02-2020	408250.7	565936.4
03-2020	432691.1	606861.7
04-2020	425372.4	580448.3
05-2020	419795.4	655054.6
06-2020	414829.3	628577.2
07-2020	441410.2	705712.3
08-2020	436727.6	651429.9
09-2020	356538.1	565829

10-2020	393188.2	584820.3
11-2020	400891.1	572438.2
12-2020	394037.5	586723.6
01-2021	407638.6	567085.1
02-2021	429607.4	599602.3
03-2021	420056.5	571137.1
04-2021	424315.2	647009.6
05-2021	421811.5	646644
06-2021	481446.2	696034.5
07-2021	440836.3	653781.2
08-2021	345651.4	553630
09-2021	388102.2	577839.4
10-2021	383859.1	570798.9
11-2021	404733.6	602388.3
12-2021	397222.5	584560.7
01-2022	427431	607610.8
02-2022	415305.5	592952.2
03-2022	432436.9	631869.8
04-2022	417747.7	651798.8
05-2022	454244.2	693911.2
06-2022	434768.2	670040.7
07-2022	345945.4	566906.5
08-2022	383999.3	603408.5
09-2022	388941.5	585482.6
10-2022	398611	606060.6
11-2022	396194.9	603389.7
12-2022	427224	625970.5

Appendix J

Water consumption and revenue rates for 20 sub-area in year 2017

Region Name	Consumption in 2017	Revenue in 2017	Consumption %	Revenue %
الأمل	489775	729968	13.3	12.1
البطن السمين	389311	636530	10.5	10.5
التحرير	8291	12913	0.2	0.2
الجلاء	17168	29458	0.5	0.5
السطر	124460	203640	3.4	3.4

السلام	1183	3742	0.0	0.1
الشيخ ناصر	238016	383121	6.4	6.3
الكتيبة	220886	361518	6.0	6.0
المحطة	203299	329422	5.5	5.4
المعسكر	707384	1177428	19.2	19.4
المنارة	38881	67159	1.1	1.1
المواصي الجنوبي	8249	14380	0.2	0.2
المواصي الشمالي	65520	117459	1.8	1.9
النصر	116945	232232	3.2	3.8
جورت اللوت	278448	501601	7.5	8.3
قاع القرين	26781	44330	0.7	0.7
قيزان ابو رشوان	45803	42643	1.2	0.7
قيزان النجار	9989	17574	0.3	0.3
مركز المدينة	469408	789654	12.7	13.0
معن	233434	362343	6.3	6.0

Appendix K

Last three months of the forecasted water consumption and revenue.

Region Name	Voucher Date	Forecasted Water Consumption (M^3)	Forecasted Water Revenue (NIS)
الأمل	10-2022	53523	82186
الأمل	11-2022	50124	84816
الأمل	12-2022	50918	89408
البطن السمين	10-2022	41597	76716
البطن السمين	11-2022	48322	62470
البطن السمين	12-2022	42152	64740
التحرير	10-2022	1112	1112
التحرير	11-2022	928	1014
التحرير	12-2022	644	1332
الجلاء	10-2022	12889	25161
الجلاء	11-2022	16451	-1236
الجلاء	12-2022	21925	-29765
السطر	10-2022	11682	20525

السطر	11-2022	12393	21448
السطر	12-2022	13512	22137
السلام	10-2022	117	340
السلام	11-2022	117	308
السلام	12-2022	117	313
الشيخ ناصر	10-2022	25676	38013
الشيخ ناصر	11-2022	26215	39811
الشيخ ناصر	12-2022	27604	38903
الكتيبة	10-2022	25364	34859
الكتيبة	11-2022	25240	36819
الكتيبة	12-2022	24582	37145
المحطة	10-2022	22761	36280
المحطة	11-2022	22761	36280
المحطة	12-2022	22761	36280
المعسكر	10-2022	76449	120561
المعسكر	11-2022	77835	119332
المعسكر	12-2022	84749	117238
المنارة	10-2022	4032	6164
المنارة	11-2022	5406	7121
المنارة	12-2022	5567	9226
المواصي الجنوبي	10-2022	557	882
المواصي الجنوبي	11-2022	831	869
المواصي الجنوبي	12-2022	1147	1717
المواصي الشمالي	10-2022	7400	11387
المواصي الشمالي	11-2022	7400	11361
المواصي الشمالي	12-2022	7400	11391
النصر	10-2022	8544	15357
النصر	11-2022	13500	43369
النصر	12-2022	9419	15356
جورت اللوت	10-2022	25272	47215
جورت اللوت	11-2022	25798	46733
جورت اللوت	12-2022	31259	49703
قاع القرين	10-2022	4053	10272
قاع القرين	11-2022	5057	14245
قاع القرين	12-2022	5158	13293
قيزان ابو رشوان	10-2022	1808	5263
قيزان ابو رشوان	11-2022	3510	7414
قيزان ابو رشوان	12-2022	2764	4017
قيزان النجار	10-2022	1028	1945
قيزان النجار	11-2022	1032	2082
قيزان النجار	12-2022	1168	2054

مركز المدينة	10-2022	52648	78705
مركز المدينة	11-2022	51393	79770
مركز المدينة	12-2022	52949	84600
معن	10-2022	25212	33371
معن	11-2022	27822	32720
معن	12-2022	26240	36972

Appendix L

Water consumption and revenue rates for 13class in year 2017.

Class Name	Consumption in 2017	Revenue in 2017	Consumption %	Revenue %
وزارة	4715	15987	0.1	0.3
بنك	1145	2642	0.0	0.0
جامعة	1086	3099	0.0	0.1
ديوان	2250	4711	0.1	0.1
روضة	1156	2106	0.0	0.0
مسكن	3508569	5790596	95.0	95.6
صالة افراح	966	1537	0.0	0.0
مدرسة او مرفق تعليمي	44604	97351	1.2	1.6
مرفق بلدي	65455	0	1.8	0.0
مرفق وزارة الصحة	2383	7525	0.1	0.1
جمعية	9083	19027	0.2	0.3
مسجد	46726	98058	1.3	1.6
وكالة الغوث	5675	15410	0.2	0.3

Appendix M

Results of last three forecasted months over classes on KhanYounis city.

Region Name	Voucher Date	Forecasted Water Consumption (M^3)	Forecasted Water Revenue (NIS)
وزارة	10-2022	584	2861
وزارة	11-2022	600	2110
وزارة	12-2022	566	1957
بنك	10-2022	102	191
بنك	11-2022	100	261

بنك	12-2022	118	237
جامعة	10-2022	414	82
جامعة	11-2022	414	82
جامعة	12-2022	414	82
ديوان	10-2022	304	227
ديوان	11-2022	281	227
ديوان	12-2022	260	227
روضة	10-2022	124	314
روضة	11-2022	127	232
روضة	12-2022	128	228
مسكن	10-2022	459910	934104
مسكن	11-2022	478992	938303
مسكن	12-2022	458679	944665
صالة افراح	10-2022	157	259
صالة افراح	11-2022	156	258
صالة افراح	12-2022	156	258
مدرسة او مرفق تعليمي	10-2022	4704	8053
مدرسة او مرفق تعليمي	11-2022	4018	7800
مدرسة او مرفق تعليمي	12-2022	4165	8725
مرفق بلدي	10-2022	24866	-28102
مرفق بلدي	11-2022	11388	-28951
مرفق بلدي	12-2022	-1640	-25428
مرفق وزارة الصحة	10-2022	-1107	-1920
مرفق وزارة الصحة	11-2022	-1127	-1959
مرفق وزارة الصحة	12-2022	-1147	-1998
جمعية	10-2022	1270	2408
جمعية	11-2022	1072	2383
جمعية	12-2022	1008	2002
مسجد	10-2022	6166	10587
مسجد	11-2022	4822	9970
مسجد	12-2022	6654	9944
وكالة الغوث	10-2022	7805	20632
وكالة الغوث	11-2022	9200	29445
وكالة الغوث	12-2022	6985	20533

Appendix N

Special cases dataset from 2007 to 2017 for water consumption and revenue.

Area	Type	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Al-Nasr	con.	62142	62064	61135	59048	64770	70121	127167	91849	228529	109798	141380
	rev.	97494	100158	94518	97917	110024	107725	236632	137768	473791	179747	272109
Al-Jalaa	con.	17773	17778	18368	18740	10694	11085	28666	42751	45221	60200	41648
	rev.	29825	31032	31045	31797	51048	56836	60709	88899	91977	115935	74544

Appendix O

Results of forecasting ‘water consumption’ and ‘water revenue’ for Al-Jalaa sub-area using ARIMA.

WATER_CONSUMPTION	WATER_REVENUE	VOUCHER_DATE	REGION_NAME
17773	29825	2007	الجلاء
17778	31032.5	2008	الجلاء
18368	31045.5	2009	الجلاء
18740	31797	2010	الجلاء
10694	51048	2011	الجلاء
11085	56836	2012	الجلاء
28666	60709.5	2013	الجلاء
42751	88899	2014	الجلاء
45221	91977.5	2015	الجلاء
60200	115935.5	2016	الجلاء
41648	74544	2017	الجلاء
41401	121378	2018	الجلاء
41727	136496	2019	الجلاء
42055	153497	2020	الجلاء
42386	172615	2021	الجلاء
42720	194114	2022	الجلاء

Appendix P

Results of forecasting ‘water consumption’ and ‘water revenue’ for Al-Nasr sub-area using ARIMA.

WATER_CONSUMPTION	WATER_ REVENUE	VOUCHER_DATE	REGION_NAME
62142	97494	2007	النصر
62064	100159	2008	النصر

61135	94519	2009	النصر
59048	97917	2010	النصر
64770	110025	2011	النصر
70121	107726	2012	النصر
127167	236632	2013	النصر
91849	137769	2014	النصر
228529	473791	2015	النصر
109798	179747	2016	النصر
141380	272109	2017	النصر
138839	267030	2018	النصر
138877	271386	2019	النصر
138915	275813	2020	النصر
138953	280313	2021	النصر
138990	284886	2022	النصر

Appendix Q

Results of forecasting 'water consumption' and 'water revenue' for MOH class using ARIMA.

WATER_CONSUMPTION	WATER_ REVENUE	VOUCHER_DATE	REGION_NAME
60957	122000	2007	وزارة الصحة
49156	98227	2008	وزارة الصحة
45510	91358	2009	وزارة الصحة
49683	99874	2010	وزارة الصحة
43852	88902	2011	وزارة الصحة
36517	74044	2012	وزارة الصحة
35384	72247	2013	وزارة الصحة
5254	13841	2014	وزارة الصحة
7643	21724	2015	وزارة الصحة
2211	8105	2016	وزارة الصحة
2690	8707	2017	وزارة الصحة
2268	7592	2018	وزارة الصحة
1925	6498	2019	وزارة الصحة
1634	5562	2020	وزارة الصحة
1387	4760	2021	وزارة الصحة
1178	4074	2022	وزارة الصحة

Appendix R

Python code for SSA

===

```
import numpy as np
import pandas as pd
from numpy import matrix as m
from pandas import DataFrame as df
from scipy import linalg
try:
    import seaborn
except:
    pass
from matplotlib.pylab import rcParams
rcParams['figure.figsize'] = 11, 4

class mySSA(object):
    '''Singular Spectrum Analysis object'''
    def __init__(self, time_series):

        self.ts = pd.DataFrame(time_series)
        self.ts_name = self.ts.columns.tolist()[0]
        if self.ts_name==0:
            self.ts_name = 'ts'
        self.ts_v = self.ts.values
        self.ts_N = self.ts.shape[0]
        self.freq = self.ts.index.inferred_freq

    @staticmethod
    def _printer(name, *args):
        '''Helper function to print messages neatly'''
```

```
    print('-'*40)
    print(name+':')
    for msg in args:
        print(msg)

@staticmethod
def _dot(x,y):
    '''Alternative formulation of dot product to allow missing values in arrays/matrices'''
    pass

@staticmethod
def get_contributions(X=None, s=None, plot=True):
    '''Calculate the relative contribution of each of the singular values'''
    lambdas = np.power(s,2)
    frob_norm = np.linalg.norm(X)
    ret = df(lambdas/(frob_norm**2), columns=['Contribution'])
    ret['Contribution'] = ret.Contribution.round(4)
    if plot:
        ax = ret[ret.Contribution!=0].plot.bar(legend=False)
        ax.set_xlabel("Lambda_i")
        ax.set_title('Non-zero contributions of Lambda_i')
        vals = ax.get_yticks()
        ax.set_yticklabels(['{:3.2f}%'.format(x*100) for x in vals])
        return ax
    return ret[ret.Contribution>0]

@staticmethod
def diagonal_averaging(hankel_matrix):
    '''Performs anti-diagonal averaging from given hankel matrix
    Returns: Pandas DataFrame object containing the reconstructed series'''
```

```
        mat = m(hankel_matrix)
        L, K = mat.shape
        L_star, K_star = min(L,K), max(L,K)
        new = np.zeros((L,K))
        if L > K:
            mat = mat.T
        ret = []

        #Diagonal Averaging
        for k in range(1-K_star, L_star):
            mask = np.eye(K_star, k=k, dtype='bool')[::-1][:L_star,:]
            mask_n = sum(sum(mask))
            ma = np.ma.masked_array(mat.A, mask=1-mask)
            ret+=[ma.sum()/mask_n]

        return df(ret).rename(columns={0:'Reconstruction'})

    def view_time_series(self):
        '''Plot the time series'''
        self.ts.plot(title='Original Time Series')

    def embed(self, embedding_dimension=None, suspected_frequency=None, verbose=False,
return_df=False):
        '''Embed the time series with embedding_dimension window size.
        Optional: suspected_frequency changes embedding_dimension such that it is divisible by
suspected frequency'''
        if not embedding_dimension:
            self.embedding_dimension = self.ts_N//2
        else:
            self.embedding_dimension = embedding_dimension
        if suspected_frequency:
```

```
            self.suspected_frequency = suspected_frequency
            self.embedding_dimension =
(self.embedding_dimension//self.suspected_frequency)*self.suspected_frequency


        self.K = self.ts_N-self.embedding_dimension+1
        self.X = m(linalg.hankel(self.ts, np.zeros(self.embedding_dimension))).T[:,:self.K]
        self.X_df = df(self.X)
        self.X_complete = self.X_df.dropna(axis=1)
        self.X_com = m(self.X_complete.values)
        self.X_missing = self.X_df.drop(self.X_complete.columns, axis=1)
        self.X_miss = m(self.X_missing.values)
        self.trajectory_dimentions = self.X_df.shape
        self.complete_dimensions = self.X_complete.shape
        self.missing_dimensions = self.X_missing.shape
        self.no_missing = self.missing_dimensions[1]==0


        if verbose:
            msg1 = 'Embedding dimension\t:  {}\nTrajectory dimensions\t: {}'
            msg2 = 'Complete dimension\t: {}\nMissing dimension     \t: {}'
            msg1 = msg1.format(self.embedding_dimension, self.trajectory_dimentions)
            msg2 = msg2.format(self.complete_dimensions, self.missing_dimensions)
            self._printer('EMBEDDING SUMMARY', msg1, msg2)


        if return_df:
            return self.X_df


    def decompose(self, verbose=False):
        '''Perform the Singular Value Decomposition and identify the rank of the embedding
subspace
        Characteristic of projection: the proportion of variance captured in the subspace'''
        X = self.X_com
```

```
    self.S = X*X.T
    self.U, self.s, self.V = linalg.svd(self.S)
    self.U, self.s, self.V = m(self.U), np.sqrt(self.s), m(self.V)
    self.d = np.linalg.matrix_rank(X)
    Vs, Xs, Ys, Zs = {}, {}, {}, {}
    for i in range(self.d):
        Zs[i] = self.s[i]*self.V[:,i]
        Vs[i] = X.T*(self.U[:,i]/self.s[i])
        Ys[i] = self.s[i]*self.U[:,i]
        Xs[i] = Ys[i]*(m(Vs[i]).T)
    self.Vs, self.Xs = Vs, Xs
    self.s_contributions = self.get_contributions(X, self.s, False)
    self.r = len(self.s_contributions[self.s_contributions>0])
    self.r_characteristic = round((self.s[:self.r]**2).sum()/(self.s**2).sum(),4)
    self.orthonormal_base = {i:self.U[:,i] for i in range(self.r)}

    if verbose:
        msg1 = 'Rank of trajectory\t\t: {}\nDimension of projection space\t: {}'
        msg1 = msg1.format(self.d, self.r)
        msg2 = 'Characteristic of projection\t: {}'.format(self.r_characteristic)
        self._printer('DECOMPOSITION SUMMARY', msg1, msg2)

def view_s_contributions(self, adjust_scale=False, cumulative=False, return_df=False):
    '''View the contribution to variance of each singular value and its corresponding signal'''
    contribs = self.s_contributions.copy()
    contribs = contribs[contribs.Contribution!=0]
    if cumulative:
        contribs['Contribution'] = contribs.Contribution.cumsum()
    if adjust_scale:
        contribs = (1/contribs).max()*1.1-(1/contribs)
```

```
        ax = contribs.plot.bar(legend=False)
        ax.set_xlabel("Singular_i")
        ax.set_title('Non-zero{} contribution of Singular_i {}'.\
                format(' cumulative' if cumulative else '', '(scaled)' if adjust_scale else ''))
        if adjust_scale:
            ax.axes.get_yaxis().set_visible(False)
        vals = ax.get_yticks()
        ax.set_yticklabels(['{:3.0f}%'.format(x*100) for x in vals])
        if return_df:
            return contribs


    @classmethod
    def view_reconstruction(cls, *hankel, names=None, return_df=False, plot=True,
symmetric_plots=False):
        '''Visualise the reconstruction of the hankel matrix/matrices passed to *hankel'''
        hankel_mat = None
        for han in hankel:
            if isinstance(hankel_mat,m):
                hankel_mat = hankel_mat + han
            else:
                hankel_mat = han.copy()
        hankel_full = cls.diagonal_averaging(hankel_mat)
        title = 'Reconstruction of signal'
        if names or names==0:
            title += ' associated with singular value{}: {}'
            title = title.format('' if len(str(names))==1 else 's', names)
        if plot:
            ax = hankel_full.plot(legend=False, title=title)
            if symmetric_plots:
                velocity = hankel_full.abs().max()[0]
```

```
            ax.set_ylim(bottom=-velocity, top=velocity)
    if return_df:
        return hankel_full

def _forecast_prep(self, singular_values=None):
    self.X_com_hat = np.zeros(self.complete_dimensions)
    self.verticality_coefficient = 0
    self.forecast_orthonormal_base = {}
    if singular_values:
        try:
            for i in singular_values:
                self.forecast_orthonormal_base[i] = self.orthonormal_base[i]
        except:
            if singular_values==0:
                self.forecast_orthonormal_base[0] = self.orthonormal_base[0]
            else:
                raise('Please pass in a list/array of singular value indices to use for forecast')
    else:
        self.forecast_orthonormal_base = self.orthonormal_base
    self.R = np.zeros(self.forecast_orthonormal_base[0].shape)[:-1]
    for Pi in self.forecast_orthonormal_base.values():
        self.X_com_hat += Pi*Pi.T*self.X_com
        pi = np.ravel(Pi)[-1]
        self.verticality_coefficient += pi**2
        self.R += pi*Pi[:-1]
    self.R = m(self.R/(1-self.verticality_coefficient))
    self.X_com_tilde = self.diagonal_averaging(self.X_com_hat)

def forecast_recurrent(self, steps_ahead=12, singular_values=None, plot=False,
return_df=False, **plotargs):
```

```
        '''Forecast from last point of original time series up to steps_ahead using recurrent
methodology

        This method also fills any missing data from the original time series.'''

        try:
            self.X_com_hat
        except(AttributeError):
            self._forecast_prep(singular_values)
        self.ts_forecast = np.array(self.ts_v[0])
        for i in range(1, self.ts_N+steps_ahead):
            try:
                if np.isnan(self.ts_v[i]):
                    x = self.R.T*m(self.ts_forecast[max(0,i-self.R.shape[0]): i]).T
                    self.ts_forecast = np.append(self.ts_forecast,x[0])
                else:
                    self.ts_forecast = np.append(self.ts_forecast,self.ts_v[i])
            except(IndexError):
                x = self.R.T*m(self.ts_forecast[i-self.R.shape[0]: i]).T
                self.ts_forecast = np.append(self.ts_forecast, x[0])
        self.forecast_N = i+1
        new_index = pd.date_range(start=self.ts.index.min(),periods=self.forecast_N, freq=self.freq)
        forecast_df = df(self.ts_forecast, columns=['Forecast'], index=new_index)
        forecast_df['Original'] = np.append(self.ts_v, [np.nan]*steps_ahead)
        if plot:
            forecast_df.plot(title='Forecasted vs. original time series', **plotargs)
        if return_df:
            return forecast_df


if __name__=='__main__':
    from mySSA import mySSA
    from pandas import DataFrame as df
```

```
import pandas as pd
import numpy as np
from matplotlib.pylab import rcParams

# Construct the data with gaps
ts = pd.read_csv('AirPassengers.csv', parse_dates=True, index_col='Month')
ts_ = ts.copy()
ts_.ix[67:79] = np.nan
ts_ = ts_.set_value('1961-12-01','#Passengers', np.nan).asfreq('MS')
ssa = mySSA(ts_)

# Plot original series for reference
ssa.view_time_series()

ssa.embed(embedding_dimension=36, suspected_frequency=12, verbose=True)
ssa.decompose(True)
ssa.view_s_contributions(adjust_scale=True)

# Component Signals
components = [i for i in range(13)]
rcParams['figure.figsize'] = 11, 2
for i in range(5):
    ssa.view_reconstruction(ssa.Xs[i], names=i, symmetric_plots=i!=0)
rcParams['figure.figsize'] = 11, 4

# RECONSTRUCTION
ssa.view_reconstruction(*[ssa.Xs[i] for i in components], names=components)
# FORECASTING
ssa.forecast_recurrent(steps_ahead=48, plot=True)
```
==