Ukrainian Catholic University

Bachelor Thesis

Optimization of schedule for distribution of dosimeter sets by the IAEA/WHO using machine learning

Author:
Anastasiia Vedernikova

Supervisors:
Yaroslav Pynda
Tomislav Bokulic

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2019

Declaration of Authorship

I, Anastasiia VEDERNIKOVA, declare that this thesis titled, “Optimization of schedule for distribution of dosimeter sets by the IAEA/WHO using machine learning” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Optimization of schedule for distribution of dosimeter sets by the IAEA/WHO using machine learning

by Anastasiia VEDERNIKOVA

Abstract

IAEA dosimetry laboratory (DOL) uses almost the same schedule for sending RPLD sets to hospitals around the world each year. Hospitals irradiate sets and send them back to the DOL for further analysis. The workload intensity of laboratory mostly depends on number of sets it receives each month. The goal of this project is to create more balanced schedule of irradiation windows, by minimizing the difference between received number of sets each month. The project consists of three main steps: forecasting waiting time, forecasting number of sets and scheduling. As a result, predictions for waiting time created by LSTM and ARIMA, together with predictions for number of sets created by Exponential Smoothing were used to generate more balanced schedule of irradiation windows using Linear Programming. New schedule satisfies all constraints and can be used next year for IAEA/WHO postal dose quality audits.

Acknowledgements

First of all, I would like to thank Pavel Kazantsev and Tomislav Bokulic, who have explained to me all processes connected with IAEA/WHO postal dose quality audits held by the laboratory in Seibersdorf, Austria. Fortunately, thanks to our discussions and analysis, we succeed to improve the data quality of 'Dosimetry Laboratory' database. Special thanks to Yaroslav Pynda, who was supporting and helping me with all stages of this project.
Also, I want to thank IAEA for being such a good place for an internship, that allowed me to contribute with my work to noble goals of dosimetry laboratory.
Finally, I am very grateful to Ukrainian Catholic University and Oles Dobsevych personally for the most challenging and exciting part of my life. For these four years of Bachelor studying, I have discovered a completely new world for me - a world of doing magic, a world of programming. I have got an enormous amount of knowledge and skills, made lifelong friends, made my dreams. Thank you, my second home, my native university.

Contents

List of Figures

List of Tables

Abbreviations and Definitions

IAEA	**I**nternational **A**tomic **E**nergy **A**gency
WHO	**W**orld **H**ealth **O**rganization
DOL	**D**osimetry **L**aboratory
SSDL	**S**econd **S**tandard **D**osimetry **L**aboratory
Radiophotoluminescence (RPL)	Property of certain substances to form intense luminescence (fluorescence) centres when irradiated with ionizing radiation; upon excitation with ultraviolet lightthe substance emits light in another light spectral region.
Radiophotoluminescentdosimeter (RPLD)	A dosimeter based on the radiophotoluminescence effect which is the property of certain substances (e.g. glass) to emit fluorescent light of larger wave length upon irradiation with ultraviolet light, when previously exposed to ionizing radiation. The intensity of the emitted fluorescent light is proportional to dose over a wide range of the irradiation dose. RPLD consists of the RPL glass rod and RPLD capsule.
RPLD capsule	A plastic capsule used for storage of a silver-activated metaphosphate glass rod; it consists of the capsule itself and a watertight plug. An RPLD capsule has an ID marking which matches the ID engraved on the glass inside the capsule.
RPLD batch	A group of dosimeters selected from the lot based on the dosimeter status and accumulated dose history. A batch dosimeter can be:Reference, LabBackground, UserBackground and UserCapsule.
Irradiation window	Period of 15 days when RPLD dosimeter should be irradiated.

Dedicated to my family who always supports me. If I said thank you to my parents from the bottom of my heart, I'd be lying; when it comes to expressing this kind of gratitude, my heart is bottomless.

Chapter 1

Introduction

1.1 Problem

Radiotherapy is one of the most effective forms of cancer treatment, and an integral component of a treatment strategy for one in two cancer patients. To ensure accurate radiation doses are administered to the patient, treatment machines like Co-60 and medical linear accelerators (linacs) need to be regularly calibrated.
The dose higher than the prescribed dose can cause damage to healthy tissue and may lead to serious health complications. On the other hand, tumour underdosage will limit the curative value of radiotherapy. Therefore, quality assurance programmes including regular participation in external dosimetry audits that independently check the beam calibration are imperative for hospitals to provide safe and effective cancer treatment to their patients.

Since 1969 the International Atomic Energy Agency (IAEA), together with the World Health Organization (WHO), offer postal dose quality audit services to cancer hospitals around the world. The IAEA's Dosimetry Laboratory (DOL) regularly sends out solid state dosimeters to hospitals, where medical physicists irradiate them with a typical dose and send these dosimeters back to DOL for further measurement and analysis. In case a discrepancy is detected between the dose measured at DOL and the dose stated by the participating hospital DOL will conduct one more check, a follow-up, and if necessary, organize an on-site visit by an expert who will investigate a possible cause of the discrepancy and will provide recommendations to the hospital.

Audits to hospitals are organized monthly with nine irradiation runs (windows) planned annually. An irradiation run has a defined time period within which a group of radiophotoluminescent dosimeter (RPLD) sets called an RPLD batch is sent to hospitals in different countries. RPLD sets that are sent to the same country usually belong to the same irradiation run. Distribution of participating countries into irradiation runs is done based on several parameters and historical participation data. For example, if countries have the same audit coordinator, DOL will group them into one batch. Similar annual plans are used each year, which means that countries usually receive RPLD sets from DOL almost at the same time each year. Hospitals are required to return dosimeters within two weeks after irradiating them but in practice the turnover time might be longer for individual hospitals or countries. This results in the duration between the irradiation window and receiving the packages back at DOL differ and change over time, making it challenging to plan and evenly distribute the monthly workload at DOL. To improve the balance in the working rhythm of DOL and avoid queues to the dosimetry readers that may result in delays in reporting results to participants, an analysis of the patterns of returning dosimeter packages to the IAEA should be performed.

1.2 Goal

For this reason, the main objective of this project is to predict the date of receipt of RPLD packages from the hospitals back at DOL. This forecast will assist in establishing an algorithm to help schedule when a package is anticipated to be received back at DOL, based on the date of the irradiation window. Furthermore, it is anticipated that through this planning, the schedules throughout the year can be uniformly distributed. The ultimate aim of scheduling irradiation windows in such a way is that it would 1) alleviate the work intensity and pressure on DOL staff and 2) reduce potential delays in reporting back to participating hospitals. The optimally balanced system should have almost the same number of RPLD sets, which the laboratory receives from countries, each month.

1.3 Description of processes

Firstly, in November DOL creates plan with irradiation windows for the next year. In December it sends out invitations to participants asking whether they want to participate and how many dosimeters they need. The laboratory has a rule of maximum three sets per hospital. DOL has established rules for sending sets. So, if it sends RLPD sets to Chili, for example, it will send them two months before irradiation window, to which Chili is assigned. Having the response from hospitals and knowing to which irradiation window country is assigned by created plan, DOL sends out sets. After the hospital receives dosimeters with all instructions, medical physicists irradiate them and send them back to the laboratory for check. Physicists have to irradiate sets between start and end of irradiation window (batch). The time between the end of irradiation window and receive date is called "waiting time" in the scope of this project. When DOL receives sets, it checks them and sends back to the hospital results of an audit.

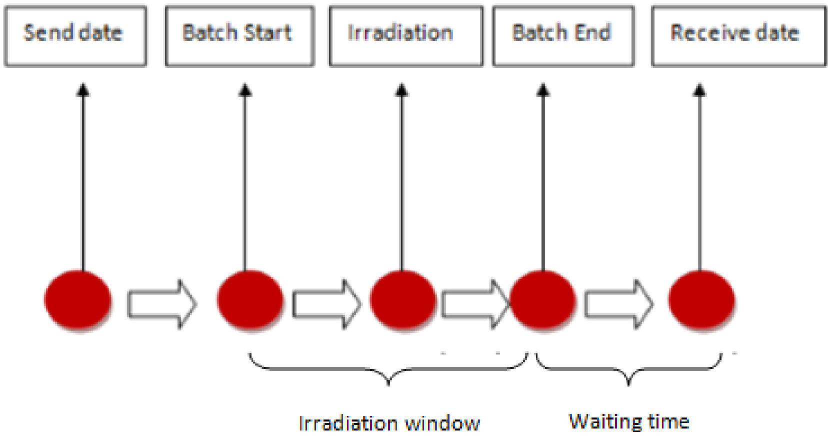


FIGURE 1.1: Description of processes

1.4 Approach

Project processes are divided into three main steps:
1) forecasting the time between the end of an irradiation window (end date when the hospital can irradiate RPLDs to be checked according to the annual audits agenda) and receipt dates,
2) forecasting number of RPLD sets that DOL sends to countries in each batch and
3) modeling assignment of countries to irradiation runs through the year.

The second step is needed because when DOL creates a plan in November, it does not know how many sets to expect from each country. The laboratory sends out invitations to participate in an audit after a plan is created. And only from responses to these invitations DOL finds out how many sets to send to each hospital. But using historical data the number of sets can be forecasted and used for creating balanced schedule.

Predicted waiting time and the number of sets allows us to understand when and how much sets we expect to receive from countries, based on irradiation window to which country is assigned.

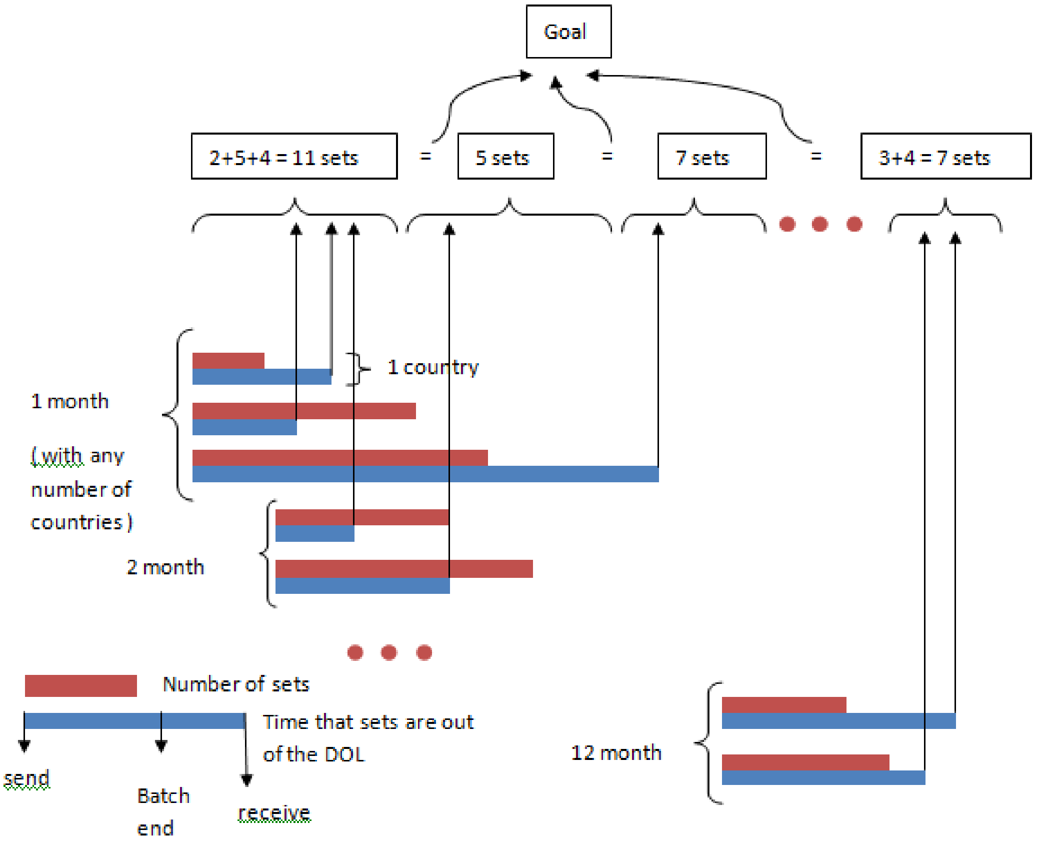


FIGURE 1.2: Visualisation of project's goal

1.5 Constraints

Having the duration between the end of irradiation window date and the receipt date, and expected number of sets per country, a schedule should have almost the same number of dosimeters received each month, taking into account the following constraints:

1) Holidays – DOL should avoid sending dosimeters to countries during their national holidays;

2) Methods of dosimeters' distribution. Countries which receive dosimeters through PAHO (distributor for Latin America and the Carribean), should be assigned to one of the specific three irradiation runs.

3) Non-operational/fixed months. DOL has scheduled ten irradiation windows when countries are supposed to irradiate sets. One irradiation window is fixed for SSDL. But in the scope of this project we analyze just hospitals, that means nine irradiation windows. In January and July the laboratory does not have planned irradiation windows, May is fixed for SSDL.

4) Already established annual audit schedule. The laboratory has agreements with each country for more than 20 years, and changes to established process should be really valuable to be made. So, the task is to find changes that would influence the most on balance of DOL workload.

Chapter 2

Related Works and Background Information

2.1 Forecasting

Forecasting – is predicting the future using all available information, including historical data and knowledge about any processes, events that might impact the forecasts.

Forecasting becomes more and more popular for data analysis and a wide variety of ways to create predictions are being improved with impressive speed. You can use simply the most recent observation as a forecast or think about something more complex, such as neural networks and economic systems of simultaneous equations. The choice always depends on the number and quality of the available dataset. Regardless of the circumstances or time constraints, forecasting is an important aid to effective and efficient planning.

In the book (Hyndman, 2018), it is recommended firstly to check predictability of an event, that depends on:

1. how well we understand the factors that contribute to it;
2. how much data are available;
3. whether the forecasts can affect the thing we are trying to forecast.

In our case, it is very difficult to define reasons, why one or another hospital tends to keep packages a particular period of time. We can not analyze seasonality, because each country used to receive packages almost in the same month each year. We do not have any information about processes in the hospital. For example, who is responsible for accepting applications, packages and irradiating them and what influences the time, when they do this. Also, we have limited data about sets' transportation. We know what type of transportation we use for each country (direct to hospital, through PAHO or representative of the country) but we have no idea how long it takes as soon as time depends not only on distance but also on responsibility of ambassador and any unexpected circumstances that can appear.

It would be much easier if we have any information about hospitals' schedules and how control over them is organized. We can just do some intuitive assumptions based on historical data about their previous experience participating in the audit and general information about the hospital (location, number of machines and human resources, participation in other programs organized by IAEA). Lack of more detailed information causes uncertainty in analysis and difficulties in making forecasts.

To sum up, without knowing whether we have data that can influence on RPLD package waiting time, we would like to try all three ways: descriptive model (using all possible features), time series forecasting (using only historical data about waiting time (WT) in the past) and mixed model, that combines two previous models.

A model with predictor variables, in the first case, might be of the form:

$$WT = f(size, location, machines, specialists, error),$$

second case:

$$WT_{t+1} = f(WT_t, WT_{t-1}, WT_{t-2}, WT_{t-3}, \ldots, error),$$

third case:

$$WT_{t+1} = f(WT_t, size, location, machines, specialists, error),$$

The 'error' for any variation caused by not included variables.

2.1.1 Hierarchical Time Series

Since we have a hierarchical structure of data (countries, hospitals, machines), it makes sense to try hierarchical time series forecasting (Hyndman, 2018). We would like to create a complex system, where forecasts of country's waiting time add up to give forecasts of hospital's waiting time.

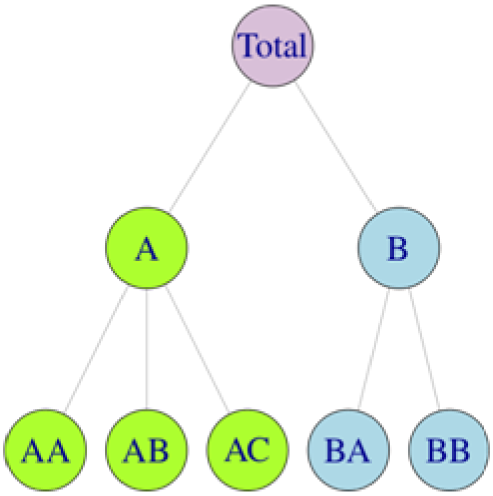


FIGURE 2.1: 2-level Hierarchical structure

'Total' is the most aggregate level of data (0-level), that is divided into smaller series (lower level) recursively. The t-th observation of Total series is denoted by y_t for t=1,. . . ,T. To denote the t-th observation of Total series is corresponding to node j at a lower level, we use y_j,t. For example, y_{AA}, t denotes the t-th observation of the series corresponding to node AA at level 2. The total number of series in this hierarchy is n=1+2+5=8 and the number of observations at bottom-level will sum to observations of series above for any time t.

$$y_t = y_{AA,t} + y_{AB,t} + y_{AC,t} + y_{BA,t} + y_{BB,t}$$

$$y_{A,t} = y_{AA,t} + y_{AB,t} + y_{AC,t}$$

These equations can be more efficiently represented using matrix notation.

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{AA,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} y_{AA,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix} \right) \tag{2.1}$$

The simplest way to generate coherent forecasts is to use the bottom-up approach, that means basically the creation of bottom-level forecasts and summing these to get forecasts of levels above. But there is little bottom-level data and it is quite noisy. So, that makes the top-down approach more attractive for use. This approach works only with strictly hierarchical aggregation structures and begins with generating Total series forecast y_t and then disaggregating these down the hierarchy. $p_1, \ldots, p_m$ – set of proportions which define how forecasts of Total series should be distributed to obtain bottom-level forecasts.

$$\tilde{y}_{AA,t} = p_1 \hat{y}_t, \qquad \tilde{y}_{AB,t} = p_2 \hat{y}_t$$

There are several ways to calculate p_j. One of them is called "average historical proportions". Each p_j is the average of all historical proportions of bottom-level series $y_{j,t}$ over the period t=1, . . . ,T.

$$p_j = \frac{1}{T} \sum_{t=1}^{T} \frac{y_{j,t}}{y_t}$$

In a book (Hyndman, 2018), you can find several more complex approaches to calculate proportions and even how to forecast them to solve the problem of information loss caused by aggregation and to take advantage of individual series characteristics. Sometimes a special combination of bottom-up and top-down methods is used, when there is an interest in the forecast on the intermediate level of hierarchy. Unfortunately, all methods do not take into account exciting correlations in the hierarchy structure, same as prediction intervals for the forecast. In a book (Hyndman, R.A. Ahmed, and Shang, 2011) it was proposed new optimal combination forecast method for HTS. It involves creating forecasts for all hierarchy series and then using of regression model to get a reconciliated forecast.

Unfortunately, currently we have too little data about hospitals to make hierarchical forecasts, but in future, it can become an effective approach.

2.1.2 ARIMA

Autoregressive Integrated Moving Average Model (ARIMA) – class of statistical models, that is used to better understand or predict time series data (Hyndman and Athanasopoulos, 2018). ARIMA is a generalization of the autoregressive moving average model. The AR part of ARIMA stands for 'autoregression'. This model describes time series in terms of dependency between observed values and some linear combination of previous observed values up to a defined maximum lag, denoted p, plus random error.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

The MA of ARIMA stands for 'moving average'. This model uses dependency between an observed value and some linear combination of previous random error terms (residuals) up to defined maximum lag, denoted q.

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

,where θ_q is the cofficient for the lagged error term in time t-q.

The I of ARIMA, in the middle, stands for 'integrated' and indicates, that data values are replaced, in some cases even several times, with differences between current and previous values. Using ARIMA allows modeling non-stationarity. ARIMA (p,d,q,) –standard notation, where parameters mean:

p - lag order
d - degree of differencing (to make time series more stationary)
q - order of moving average

ARIMA can be configured to do the function of AR, I, MA, ARMA models if specific parameter(s) will have the value of 0. ARIMA model in the formula can be written as:

$$y'_t = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where y'_t - differenced series, predictors include lagged values of y_t and lagged errors.

2.1.3 LSTM

Long Short Term Memory (LSTM) network – recurrent neural network (RNN) that is able to learn long-term dependencies (Hochreiter and Schmidhuber, 1997). Simple RNN is a chain of repeating modules with one tanh layer of the natural network. (Olah, 2015)

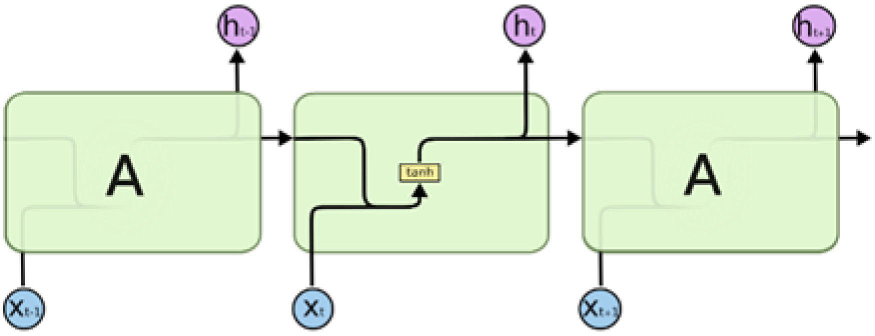


FIGURE 2.2: Recurrent Neural Network

LSTMS's repeating module consists of four interacting layers, that allow it to forget what we want to forget and keep just needed information.

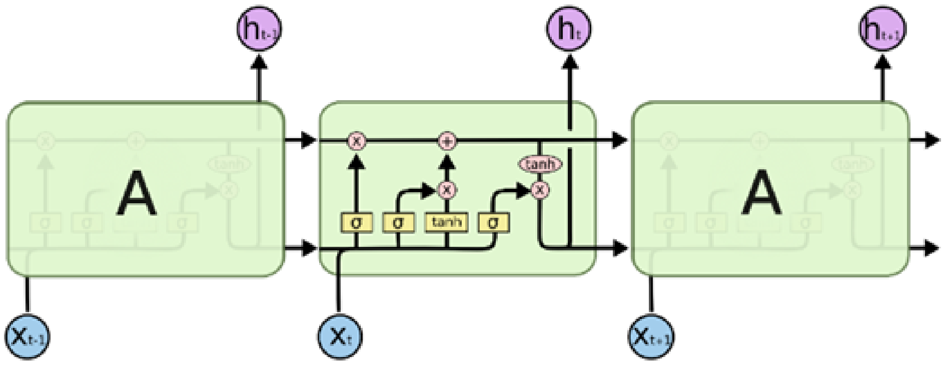


FIGURE 2.3: Long Short Term Memory network

Starting from the left, new value x is concatenated with previous output from cell h_{t-1} and fitted to sigmoid layer (also called "forget gate layer"), which outputs numbers between 0 and 1. Output f_t indicates how much data should be kept: 1 means – everything, 0 – nothing.
Next process – deciding what information should be stored in a cell state. Sigmoid layer (also called "input gate layer") returns values that will be updated, while tanh layer generates new candidate values for cell state.
Then we are ready to update the state. The old state should be multiplied by f_t to forget not needed information. Then we add result of the previous operation with multiplied candidate values by the value that indicates how much state values should be updated. Next sigmoid layer decides what cell state parts should be included in the output. We fit cell state to tanh and multiply it by results from sigmoid gate. The returned output will include just parts that it has been decided to be important.

2.1.4 Light GBM

Ensemble – is a combination of predictors that aim to reduce the difference between an actual value and predicted one. Such difference is caused by noise, variance and bias factors. Ensembling techniques are classified into Bagging and Boosting.

Bagging - is a collection of independent predictors combined by some averaging techniques. Each predictor uses different observations from the dataset obtained by the bootstrap process. It means that we randomly choose sub-samples (bootstraps) for each predictor. Since many uncorrelated learners in bagging ensemble work on final prediction, the variance will be reduced. Random forest is an example of bagging ensemble.

Boosting - is a sequential combination of predictors, where models learn from mistakes of previous models. Observations for each model are chosen not randomly but based on how often previous models made a mistake with this observation. Boosting ensemble is faster to compare with bagging one. Gradient boosting is an example of boosting ensemble.

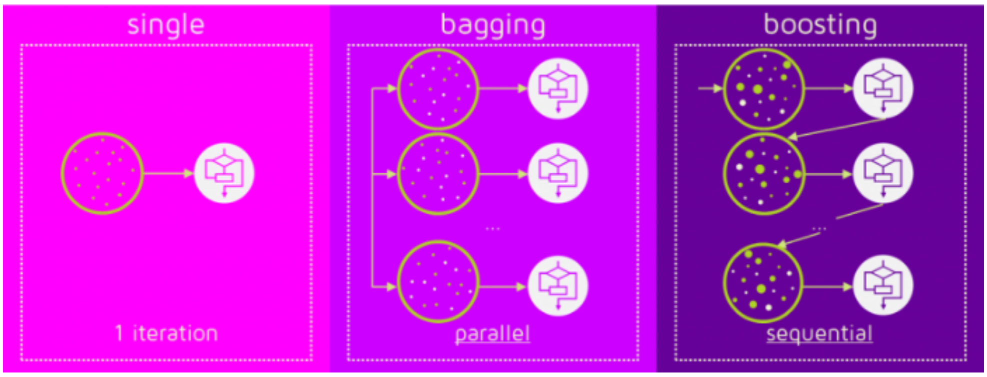


FIGURE 2.4: Difference between Bagging and Boosting

Gradient Boosting – is a machine learning algorithm for regression and classification that predict using an ensemble of weak predictors. (Friedman, 1999) The idea of algorithm is to repetitively leverage the patterns in residuals and strengthen model with weak predictions.(Grover, 2017) The process looks like:

1) we use simple models to predict values and find errors

2) while predicting with next models we focus on data points with the largest errors from previous models

3) we give weight to each predictor and combine them all

Light Gradient Boosting Model (GBM) - is a gradient boosting framework, where word "Light" stands for "fast", based on decision tree algorithm. (Ke et al., 2017) Training speed is faster, memory usage is lower and accuracy usually better to compare with XGBoost model. Light GBM grows tree leaf-wise, not level-wise as usually other tree models do. To grow the algorithm will choose leaf with maximum delta loss. (Mandot, 2017)

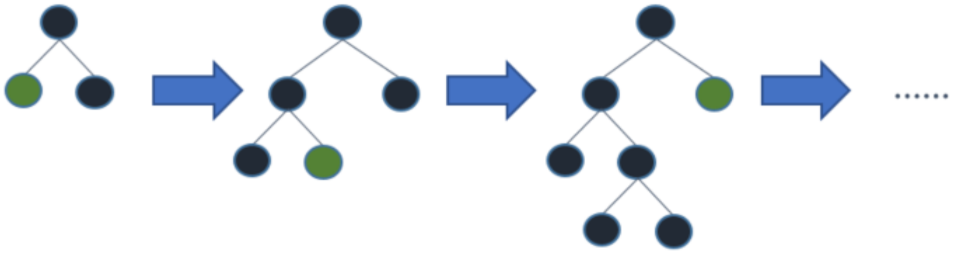


FIGURE 2.5: Leaf-wise tree growth

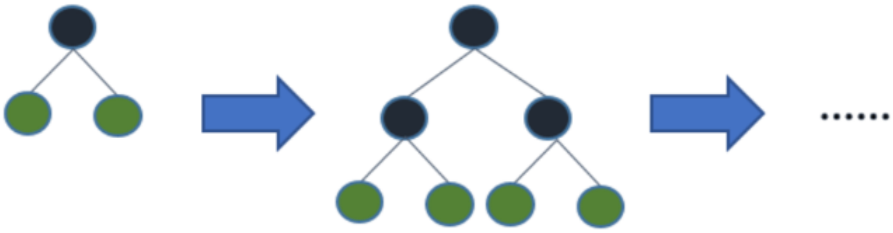


FIGURE 2.6: Level-wise tree growth

Growing leaf-wise allows for reducing more loss in comparison with level-wise growth.

2.1.5 Dynamic Regression

Dynamic regression model (DRM) is a model which is time-dependent and which includes the logged value of explanatory variables (Hyndman, 2018).

Auto-regressive distributed lag model (ARDL) - specific case of a class of dynamic regression models.

$$y_t = B_0 x_t + B_1 x_{t-1} + B_2 x_{t-2} + ... + B_k x_{t-k} + etc$$

where $B_0, B_1, B_2, ..., B_k$ – impulse response function of the mapping on $x_t$to y_t.

A lot of papers are written about comparing time series forecasting and dynamic regression model. For example, in the paper (Othman, Mohammed, and Ismaeel, 2013), researchers predict average monthly humidity using only one feature that has a big impact on humidity - average monthly relative rainfall. They were using both ARIMA and DRM for forecasting and concluded with best choice for their dataset - DRM.

ARIMA model can handle time series dynamics but does not take in a count other features of the process that can have a big impact on forecasts. But fortunately, ARIMA can be extended to include other information. Details about how to combine ARIMA and regression can be found in a book (Hyndman, 2018).

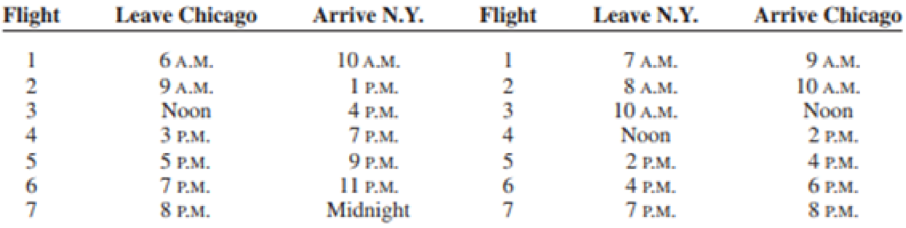
2.2 Scheduling

2.2.1 Linear Programming

For scheduling task, Linear Programming (LP) model may be used. LP is a method to achieve the best outcome in a mathematical model whose requirements are represented by linear relationships (Schrijver, 1998). LP model has the number of activities, some constraints, and objective that the model tries to achieve by changing decision variables. The typical objective of such model is about minimizing costs or maximizing profit.

LP is one of mathematical programming models, that distinguishes itself by possessing three main properties: proportionality, additivity, and divisibility. The proportionality means that by multiplying the value of any activity by a constant factor, we multiply its contribution to objective or constraints by this factor as well. The additivity property means that total contribution to constraint equals to the sum of activity contributions to that constraint. The divisibility property means that both integers and non-integers can be used as changing variables in the LP model.

There is a good example for scheduling task in a book (Winston and Albright, 1994). Braneast Airlines have to schedule daily flights between New York and Chicago. Every day Braneast's crew have two flights Chicago-New York and New York – Chicago with at least one hour of downtime. The airline company wants to schedule flights in such a way that it covers all flights and minimize downtime for crews. Braneast's crews are Chicago based or New York based and the company should figure out how many of each city based crews it needs.



Flight	Leave Chicago	Arrive N.Y.	Flight	Leave N.Y.	Arrive Chicago
1	6 A.M.	10 A.M.	1	7 A.M.	9 A.M.
2	9 A.M.	1 P.M.	2	8 A.M.	10 A.M.
3	Noon	4 P.M.	3	10 A.M.	Noon
4	3 P.M.	7 P.M.	4	Noon	2 P.M.
5	5 P.M.	9 P.M.	5	2 P.M.	4 P.M.
6	7 P.M.	11 P.M.	6	4 P.M.	6 P.M.
7	8 P.M.	Midnight	7	7 P.M.	8 P.M.

FIGURE 2.7: Flight schedule

Decision variables: 0–1 values depending on the assignment of crews to pairs of flights

Other output cells: Downtimes for crews

Objective: Total downtown

Constrains: Flow balance

It was decided to model this problem as a network with flows. The type of node indicates whether a flight is from Chicago ('C') or New York ('N'). Nodes are linked by arcs if there is an ability for the crew to make this flight, taking in a count at least one hour of downtime.

For example, if C2 (second Chicago based crew) leaves Chicago at 9 A.M., he arrives in New York at 1 P.M. Then he has to spend at least one hour of downtime to be able to take another flight. So, there are three possible flights for him from New York back to Chicago: 2 P.M., 4 P.M. and 7 P.M.

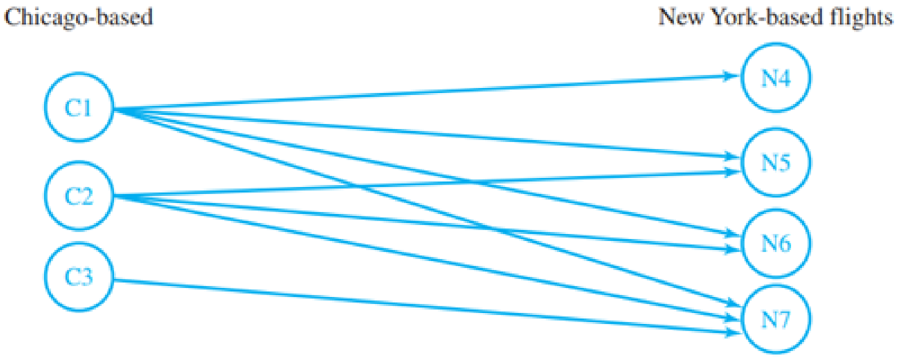


FIGURE 2.8: Network of flights

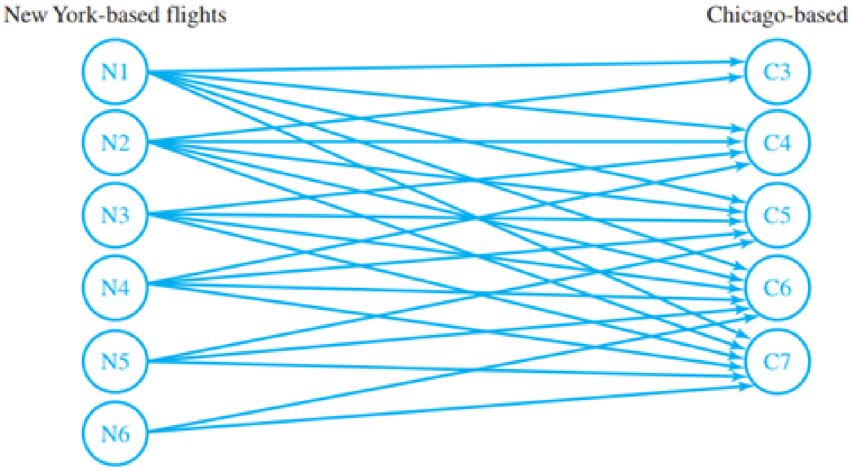


FIGURE 2.9: Network of flights 2

Using this concept excel file was created, and a solution was found by excel solver. The company should have two Chicago-based crews and five New York-based crews with a minimal total downtime of 26 hours.

2.2.2 LRM

LRM - heuristic algorithm for balanced multi-way number partitioning (BMNP), that splits the collection of numbers into subsets with almost same cardiality and subset sums. (Zhang, Mouratidis, and Pang, 2011)
LRM works with uniformly distributed numbers with odd subset cardinality b (for example b=3).
Firstly, collection of numbers is sorted in descending order and divided into b groups (p_1, p_2, p_3) with means (μ_1, μ_2, μ_3) respectively. To get perfect balanced partitioning, the sum of each final subset should be close to $\mu_1 + \mu_2 + \mu_3$. To form subset, we take the leftmost number v_L from one group, the rightmost number v_R from another one and compensating number from the middle of remaining group that is closest to $\sum_{i=1}^{i=3} \mu_i - v_L - v_R$. Leftmost number we take from group with the largest spread, rightmost from second largest spread group and middle from group with smallest spread.
For example, we have $p_1 = (12, 11, 10, 9), p_2 = (8, 7, 6, 5), p_3 = (4, 3, 2, 1)$ and $\mu_1 + \mu_2 + \mu_3 = 19.5$. Since all 3 groups have same spread of 3, we use p_1 for getting leftmost number v_L - 12, and p_2 for rightmost v_R - 5. The compensation number is chosen from p_3 - 2, to be the closest to 19.5 - 12 - 5 = 2.5. So, first subset will be {12,5,2}. Now, spread of $p_1 = (11, 10, 9)$, $p_2 = (8, 7, 6)$ and $p_3 = (4, 3, 1)$ is 2,2,3 respectively. It means that p_3 and p_1 will be chosen for L and R operations. Using previously described concept, second subset will have the form of {4,9,6}. Repeating the process, final optimal partitioning can be obtained - ({12,5,2},{4,9,6},{3,10,7},{1,11,8}), with a spread of only 1. LRM can be extended to odd values of b larger than 3.
The time complexity of LRM is O(nlogn).

Chapter 3

Description of data

3.1 Dosimetry audits dataset

ID	Name	Type	Values	Description
1	Audit Type	String	RT	Radiotherapy
2	Participation Type	String	a) Hospitals b) SSDL	Hospitals and Second Standard Laboratories (SSDL) are participating in audits. Only Hospitals are analyzed in the scope of this project.
3	Participation Category	String	a) CRP b) QUATRO c) Regulator Participation d) Special Request	a) Coordinator Research Project b) Quatro mission c) Planned participation d) Participation upon special request (highest priority)
4	Coordination Network	String	a) PAHO b) WHO c) Blank	a) Coordinator for Latin America b) World Health Organization c) Hospital does not have coordinator
5	Operator ID	Integer	3260 -17183	ID of hospital in DIRAC database
6	Batch ID	Integer	1-385	ID of Irradiation window
7	Batch No	Integer	1-303	Number of Irradiation window (descriptive id)
8	Batch Year	Integer	1969-2019	Year of Irradiation window
9	CCode	String	UKR(Ukraine), FIN(Finland), etc.	Code of country (Human Health division of IAEA Standard)
10	Country ISO3Code	String	UKR(Ukraine), FIN(Finland), etc.	Code of country (ISO3)
11	Communication Language	String	a) English b) Spanish c) Russian d) Blank	Language of communication in country

TABLE 3.1: Dosimetry audit dataset

12	Send To Data Sheets Option	String	a) DirectHospital b) NationalCoordinatorPAHO c) HospitalCCNationalCoordinator d) NationalCoordinator e) NationalCoordinatorCCPAHO f) Blank	Option that is used for sending datasheet and sets to hospitals: a) Directly to hospital b) Through PAHO c) Directly to hospital notifying national coordinator d) Through national coordinator e) Through national coordinator notifying PAHO
13	Return Data Sheets Option	String	a) OptionA b) OptionB c) OptionC d) Blank	a) Return of data sheets to the IAEA and dosimeter sets to the National Dosimetry Audit Coordinator b) Return of data sheets and dosimeter sets to the National Dosimetry Audit Co-ordinator c) Return of data sheets and dosimeter sets to the IAEA
14	Set Status	String	a) Accepted b) DataSheetsMissing c) ErrorDuringIrradiation d) ImprovedAfterFollowUp e) ImprovedNextParticipation f) LostInTransit g) NotReturned h) NotReturnedNextParticipation i) Other j) Persisting k) ResolvedByExpert l) ResolvedBySSDL m) Unexposed n) Blanks	Status of set: a) Results are considered to be successful b) Datasheet(s) missing or not filled in c) Error occurred during irradiation d) Successful results for follow up set e) Successful results for next participation (after follow up) f) Set is lost during transition g) Set is not returned to DOL h) Set is not returned to DOL during next participation i) Other j) Same results k) Set is checked by an expert (after follow up) l) Set is checked by SSDL (after follow up) m) Set is unexposed (returned but not irradiated) n) Results are in limits
15	Set Type	Integer	a)1 b) 2	a) First participation b) Follow up in case first results outside the limit
16	Set ID	Integer	1-17889	ID of set

17	Application End Date	Date	2013-2019	Last date when application for participation can be sent to hospitals
18	TLD Package Send Date	Date	2013-2014, 2016-2018	Date when datasheets, instructions are sent through email
19	Batch Start Date	Date	1998-2019	Date when Irradiation window starts
20	Batch End Date	Date	1998-2019	Date when irradiation window ends
21	Set Sent On	Date	1996-1997, 2001-2019	Date when physical package (dosimeter, holder, etc. . .) is sent
22	Irradiation Date	Date	1969-2019	Date when dosimeter is irradiated in hospital
23	Set Received On	Date	2001-2019	Date when DOL receives set from hospital
24	Reading Date	Date	1969-2019	Date when DOL reads dosimeter
25	Evaluation Date	Date	1969-2019	Date when results of dosimetry audit are evaluated
26	Certificate Date	Date	1969-2019	Date when certificate with results is prepared
27	Sign By Officer On	Date	1969-2019	Date when certificate is signed by officer
28	Sign By Section Head On	Date	1969-2019	Date when certificate is signed by Section Head
29	Dispatched On	Date	1969-2019	Date when certificate is sent back to the hospital
30	Archived On	Date	1969-2019	Date when process is closed

3.2 DIRAC dataset

ID	Name	Type	Values	Description
1	Longitude	Float	102.62, 19.82351, etc...	Longitude of hospital
2	Latitude	Float	102.62, 19.82351, etc...	Latitude of hospital
3	Region ID	Integer	0-15	ID of Region of hospital: 0 - IAEA 1 - North America 2 - Mexico and Central America 3 - Tropical South America 4 - Temperate South America 5 - Caribbean 6 - Western Europe 7 - Eastern Europe and Northern Asia 8 - North Africa 9 - Middle Africa 10 - Southern Africa 11 - Middle East 12 - South Asia 13 - East Asia 14 - Southeast Asia 15 - Southern and Western Pacific
4	RTCenters	Integer	1	Number of Radiotherapy centers with Brachytherapy in hospital
5	RTCenters WithRT	Integer	0, 1	Number of Radiotherapy centers in hospital
6	Linear Accelerator	Integer	1-10	Number of Linear Accelerators in hospital
7	Radionuclide Teletherapy	Integer	0-6	Number of Radionuclide Therapy machines in hospital
8	CT	Integer	0-4	Number of Computed Tomography in hospital
9	Simulators	Integer	0-4	Number of Simulators in hospital
10	RadOncologists	Float	0-50	Number of Radiotherapy Oncologists in hospital
11	TPS	Float	0-42	Number of Treatment Planning Systems in hospital
12	Physicists	Float	0-680	Number of Physics in hospital
13	Technicians	Float	0-94	Number of Technicians in hospital

TABLE 3.2: DIRAC dataset

3.3 World Bank dataset

ID	Name	Type	Values	Description
1	Income Group	String	a) H b) L c) LM d) UM	Income Group of country: a) High-income b) Low-income c) Low-middle income d) Upper-middle income
2	Population	Integer	65441-1386395000	Population of country
3	GDP	Float	44595558.6-12237700479375	Gross domestic product of country
4	GNIpc	Integer	70-89950	Gross National Income per Capita of country
5	Health Expenditures	Float	5.1875-2816.1263	Health and health-related expenditures of country
6	Life Expectancy	Float	43.746-84.278	An average time an organism is expected to live

TABLE 3.3: World Bank dataset

3.4 New own features

ID	Name	Type	Values	Description
1	Waiting Time	Float	-91-875	Time between "BatchEnd" (end of irradiation window) and "SetReceivedOn" (date when set is received back to DOL from the hospital)
2	MIN wt	Float	-91-62.38	Minimum "WaitingTime" for country from all history
3	MAX wt	Float	-6-875	Maximum "WaitingTime" for country from all history
4	LEN wt	Integer	1-752	Number of times when country participated in audit
5	TOTAL wt	Float	-103.61 - 76578.41	Sum of all "WaitingTime" for country
6	MEAN wt	Float	-25.9 -124.87	Mean of "WaitingTime" for country

TABLE 3.4: New own features

Chapter 4

Visualisation

4.1 Waiting time for all countries

Waiting Time - the time between the end of an irradiation window (BatchEndDate in the database) and receipt date was calculated. It is the target variable that will be predicted in this project. Bellow representation of all waiting times is shown, when BatchEndDate and SetReceivedOn are not empty. (starting from the 2001 year, 130 BatchID).

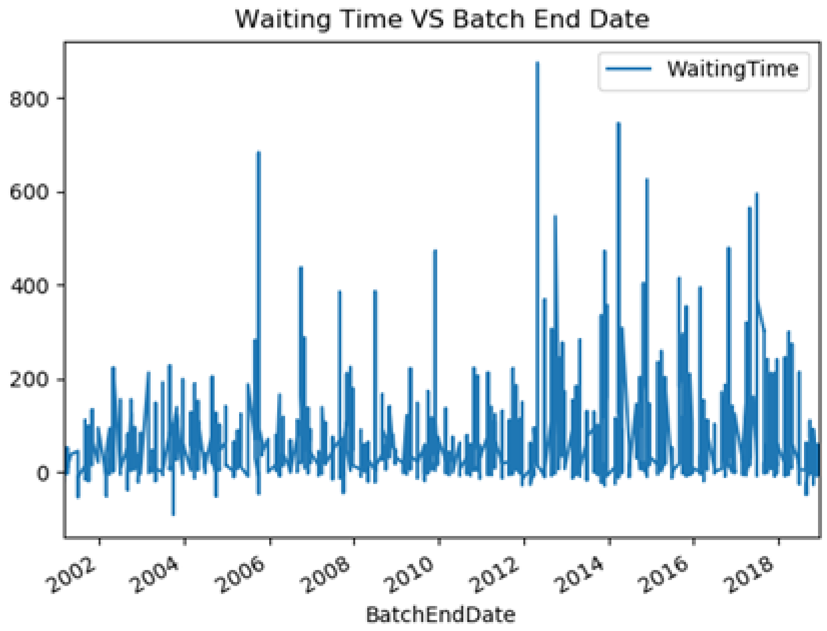


FIGURE 4.1: Waiting time VS Batch End Date - all countries

4.2 Waiting time per country

In order to understand better trends that countries have, there some visualizations with waiting times per one country. India and Ukraine were chosen for further experiments. They have close number of data but different ability to be predicted. (shown later in "Experiments" chapter)

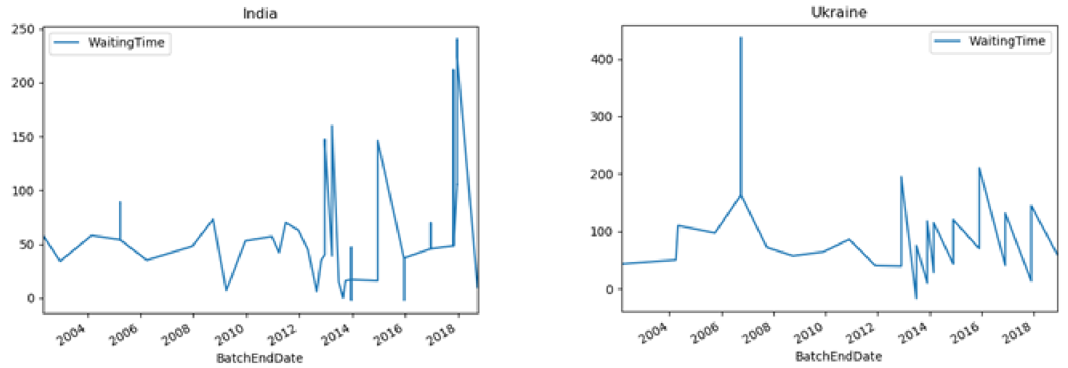
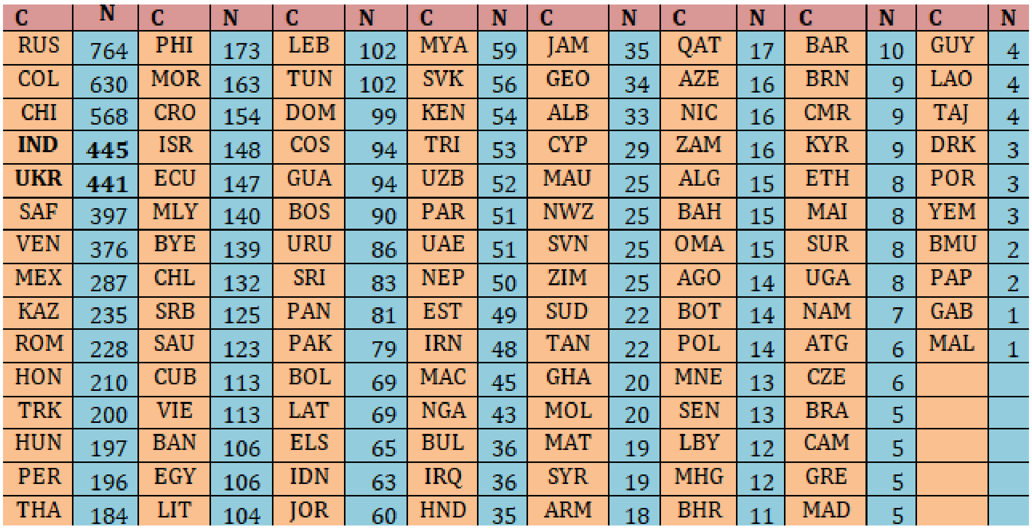


FIGURE 4.2: Waiting time VS Batch End Date - India & Ukraine



C	N	C	N	C	N	C	N	C	N	C	N	C	N	C	N
RUS	764	PHI	173	LEB	102	MYA	59	JAM	35	QAT	17	BAR	10	GUY	4
COL	630	MOR	163	TUN	102	SVK	56	GEO	34	AZE	16	BRN	9	LAO	4
CHI	568	CRO	154	DOM	99	KEN	54	ALB	33	NIC	16	CMR	9	TAJ	4
IND	**445**	ISR	148	COS	94	TRI	53	CYP	29	ZAM	16	KYR	9	DRK	3
UKR	**441**	ECU	147	GUA	94	UZB	52	MAU	25	ALG	15	ETH	8	POR	3
SAF	397	MLY	140	BOS	90	PAR	51	NWZ	25	BAH	15	MAI	8	YEM	3
VEN	376	BYE	139	URU	86	UAE	51	SVN	25	OMA	15	SUR	8	BMU	2
MEX	287	CHL	132	SRI	83	NEP	50	ZIM	25	AGO	14	UGA	8	PAP	2
KAZ	235	SRB	125	PAN	81	EST	49	SUD	22	BOT	14	NAM	7	GAB	1
ROM	228	SAU	123	PAK	79	IRN	48	TAN	22	POL	14	ATG	6	MAL	1
HON	210	CUB	113	BOL	69	MAC	45	GHA	20	MNE	13	CZE	6		
TRK	200	VIE	113	LAT	69	NGA	43	MOL	20	SEN	13	BRA	5		
HUN	197	BAN	106	ELS	65	BUL	36	MAT	19	LBY	12	CAM	5		
PER	196	EGY	106	IDN	63	IRQ	36	SYR	19	MHG	12	GRE	5		
THA	184	LIT	104	JOR	60	HND	35	ARM	18	BHR	11	MAD	5		

FIGURE 4.3: Number of data per country

4.3 Descriptive Statistics

Some descriptive statics about waiting time:

Min – -91 days (ROM 2003)

Max – 875 days (BAN 2012)

Median – 41 days

Mean – 54.1 days

Standard Deviation – 58.29 days

Mode – 34 and 45 days

Especially interesting is the minimum. According to the dataset, DOL received set three months before the end date for irradiation. Waiting time can be negative in case if DOL sends set to the hospital very early and hospital returns it back after irradiation before the end of irradiation window. But three months it is probably too long period to be true. There are situations, when the hospital asks to perform audit earlier, but documents are formed later. Maximum defines the case, when the set was in a the hospital more than two years.

4.4 Comparing DOL work intensity for last 3 years

Comparing DOL work intensity through distribution of receipts and number of sets per receipt for last 3 years.
Upper graphic shows number of sets received at one day of the month (all days of one month have one color).
Lower graphic shows distance in days between receive dates.

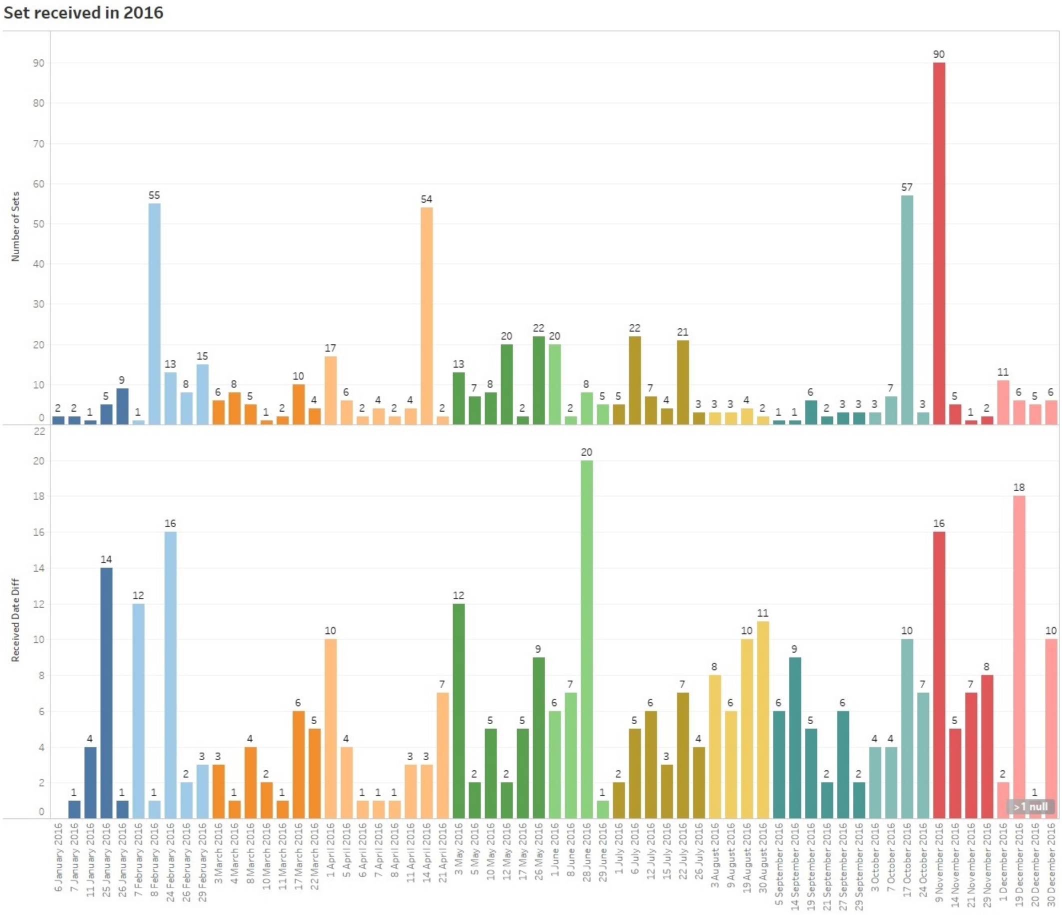


FIGURE 4.4: DOL work intensity in 2016

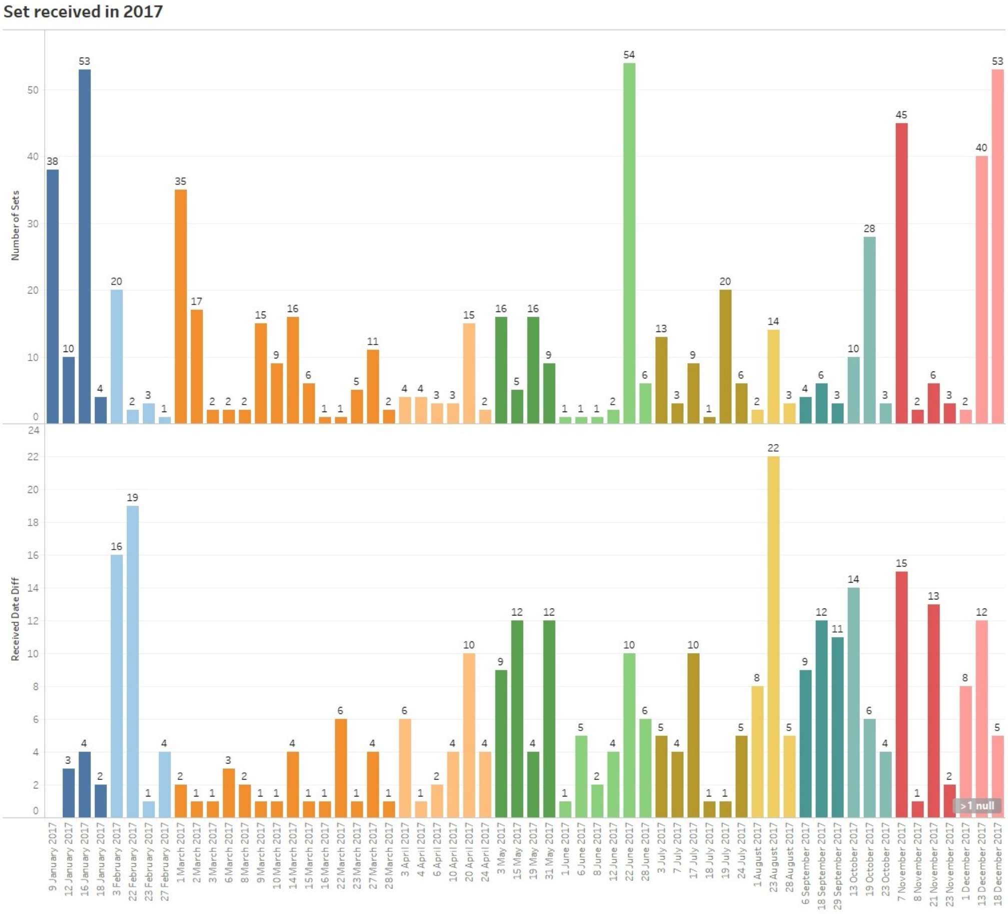


FIGURE 4.5: DOL work intensity in 2017

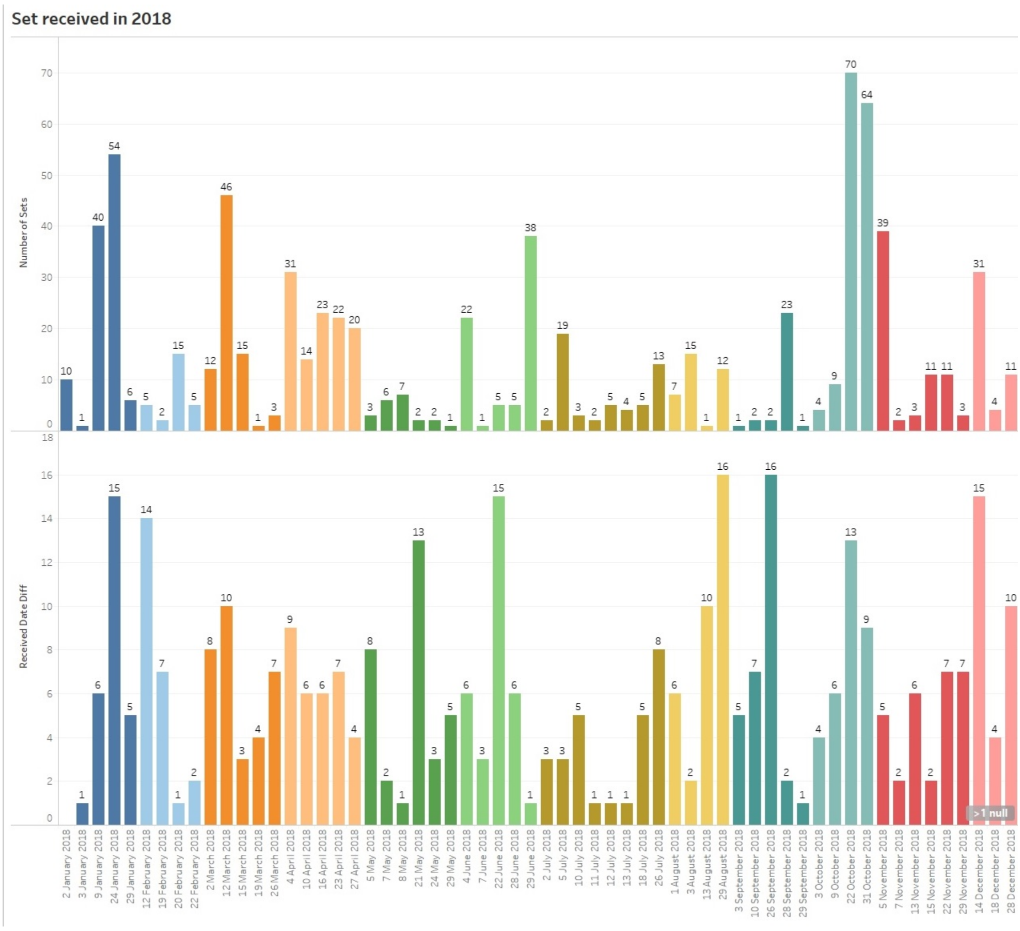


FIGURE 4.6: DOL work intensity in 2018

Figures above show that distribution changes and becomes more balanced, but still, improvements are needed. DOL work intensity increases significantly when it receives a lot of sets at one time. From figures, also can be seen, that there are periods when DOL receives sets more often. New plan should balance the number of sets received each month.

Chapter 5

Pre-processing

5.1 Cleaning

Only hospitals' data was analyzed in the scope of this project. Audits for SSDL were deleted. Available data starts from 1996 but only starting from 2001 year, some additional columns in the database, such as Receive date, Signed date, Archived date, etc., were created. So, the decision was made to use only data after 2001.
Thanks to first created visualization, I have found several strange outliers, which were reviewed and edited by an expert.

5.2 Missed data

While merging different databases, I have faced with some missed data. World Bank dataset about countries included only data till 2017, so data from last available year was duplicated for the next two.
Rows from Dosimetry audit dataset, where Batch end date or Receive date is empty, were deleted.

5.3 Transformation

For descriptive model all features were transformed to numeric.

> String values were converted to dummy variables (0/1).
>
> Dates were converted to 3 separate columns with year, month and day.

All three databases, Dosimetry audits, World Bank and DIRAC, were merged based on hospital or country values.
Also, new features were created using several statistics about waiting time per country. All used databases and new features are described in "Description of data" chapter.

5.4 Splitting data

Length of data (Receive and Batch end are not null) – 10018
Splitting data to test and train is used to evaluate predicting models. Random splitting will not be used to not destroy time component of data. Train and Test data were saved to different excel files.

Train data – 85.3% - (BatchYear<= 2016) – 8547 rows

Test data – 14.7% - (BatchYear> 2016) – 1470– rows

Note: another situation with a rolling forecast (ARIMA & LSTM)

Chapter 6

Evaluation

For algorithm evaluation mean squared error (MSE) and coefficient of determination (R^2) estimators are used.

Mean squared error (MSE) – mean of the squares of the errors, where error is the difference between the real value and predicted one.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

This measure of the estimator's quality is always non-negative and the expected value should be close to zero.

Coefficient of determination or R^2 – is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). R squared is a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. There are cases where the computational definition can yield negative values, depending on the definition used.

Chapter 7

Experiments

7.1 Predicting waiting time

7.1.1 Linear Regression

Firstly, one of descriptive models was chosen to use – Linear Regression.

Linear Regression is a linear approach to model relationships between the dependent variable and independent variables. The main idea of linear regression is to understand whether some descriptive variables can influence the outcome and which ones are more significant for predicting future outcomes (Yan, 2009). Regression equation can be represented with formula: y = c + b * x, Where y - prediction, c – constant, b – coefficient of regression, x – independent variable(s).

Dependant variable denoted as Y -“WaitingTime” (time between the end of irradiation window and receipt date) Independent variables denoted as X – features from Dosimetry Dataset.

Green – predicted, yellow – real values

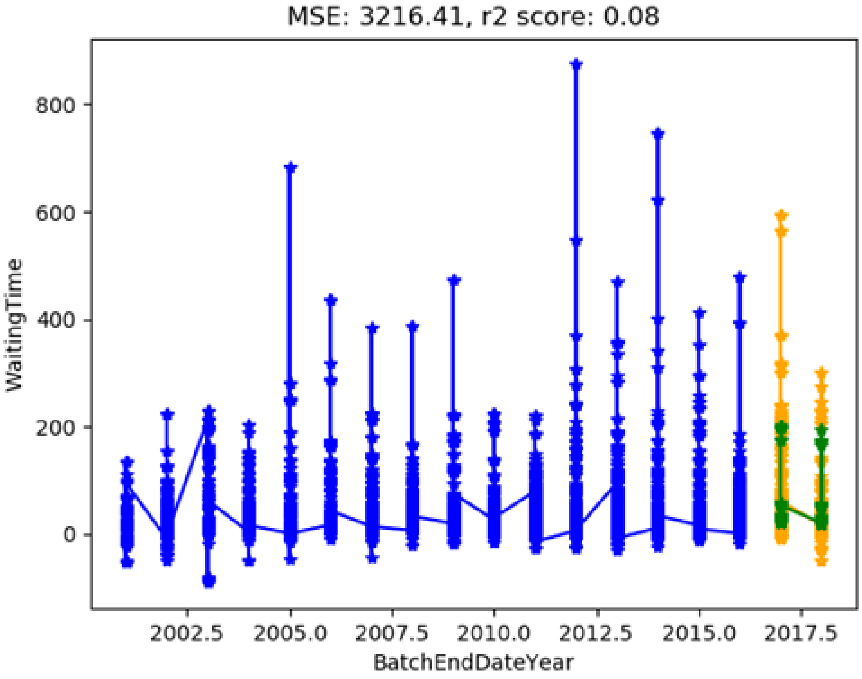


FIGURE 7.1: LR with Dosimetry Audits for All countries

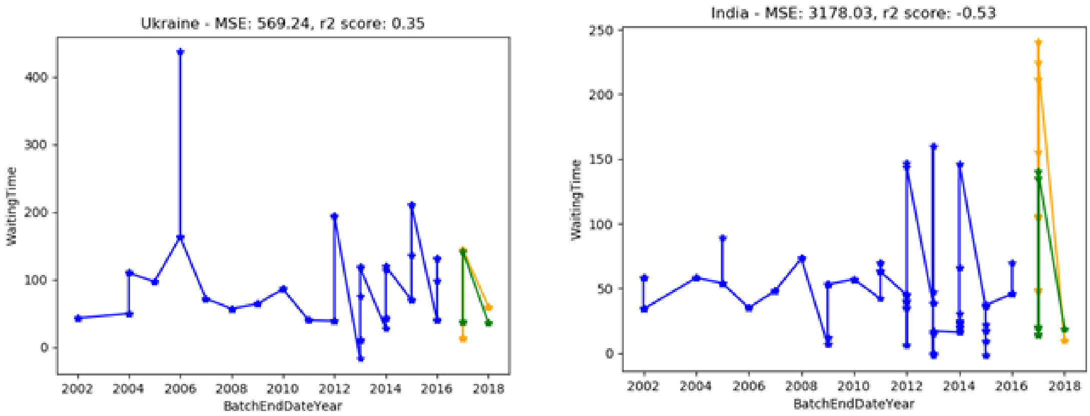


FIGURE 7.2: LR with Dosimetry Audits for Ukraine &India

Next World Bank and DIRAC data were used. Databases were merged and transformed in order to use Linear Regression. CommunicationLanguage, SendToDataSheetsOption, ReturnDataSheetsOption, IncomeGroup – converted to dummy variables (0/1) using get_dummies() function from pandas library.

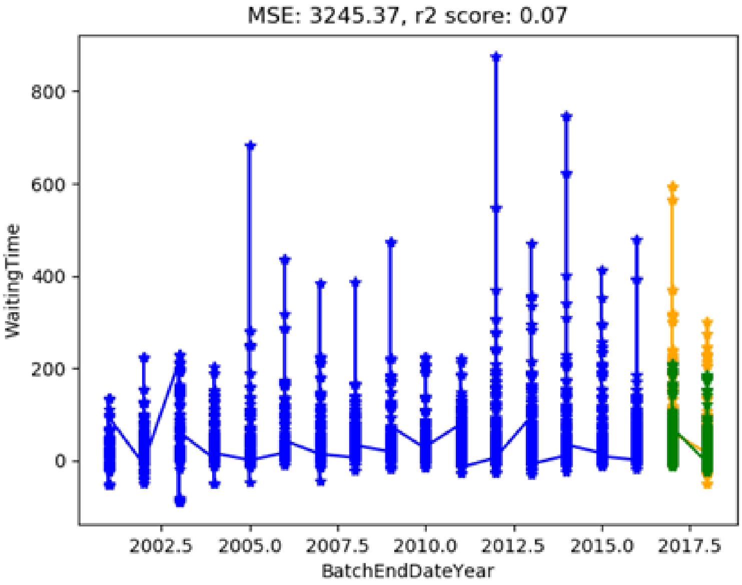


FIGURE 7.3: LR with Dosimetry Audits, DIRAC & WorldBank for All countries

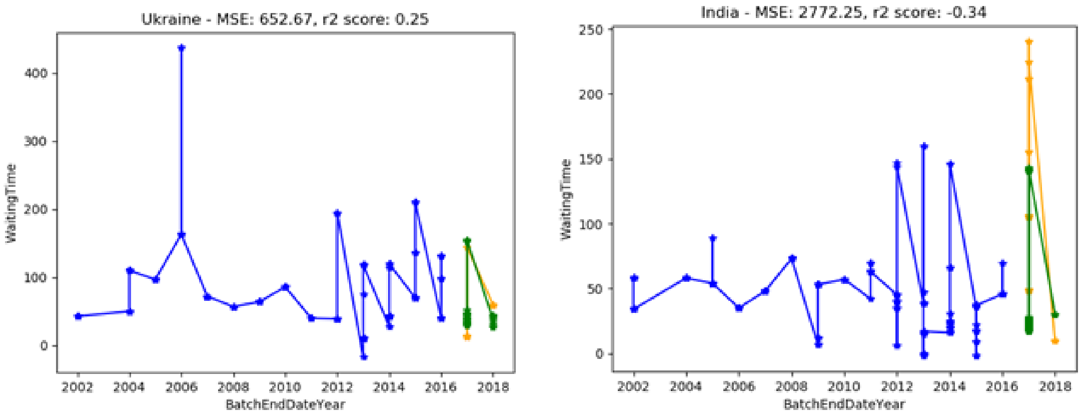


FIGURE 7.4: LR with Dosimetry Audits, DIRAC & WorldBank for Ukraine & India

There were created several additional features that describe waiting time for countries (described in "Data description" chapter above in table "New own features"). Results using additional features:

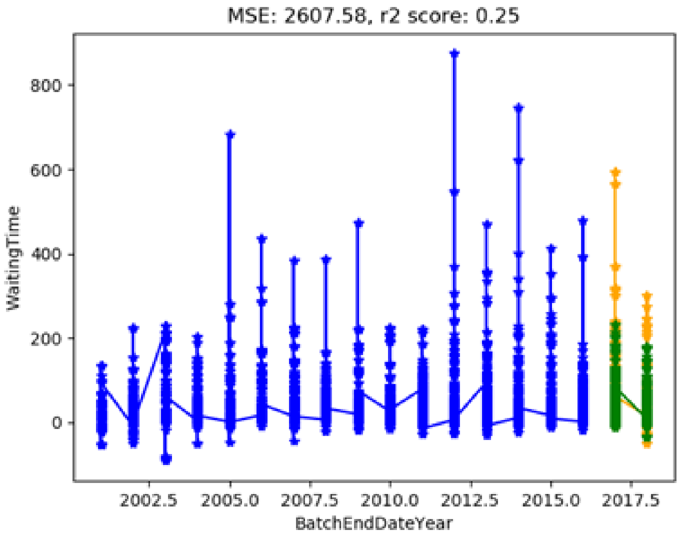


FIGURE 7.5: LR with Dosimetry Audits, DIRAC, WorldBank & New own features for All countries

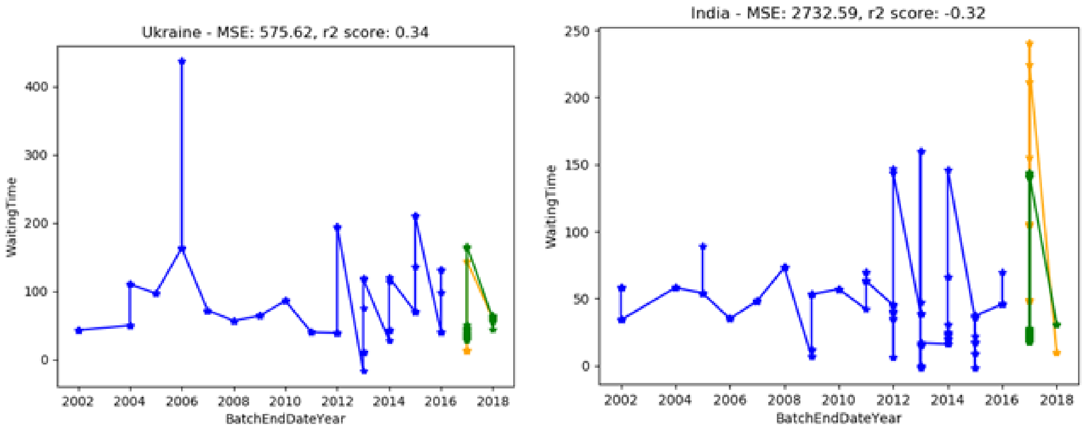


FIGURE 7.6: LR with Dosimetry Audits, DIRAC, WorldBank & New own features for Ukraine &India

From graphics, it can be observed that additional features improved R2Score: for all countries from 0.07 to 0.24 r2score, when for Ukraine from 0.25 to 0.34. The merged data set includes 43 features (columns). Some of them can be correlated and lead to worse results.

Principal component analysis (PCA) helps to get linearly uncorrelated descriptive variables for better results of prediction, using the orthogonal transformation of possibly correlated variables.

N_components = 35

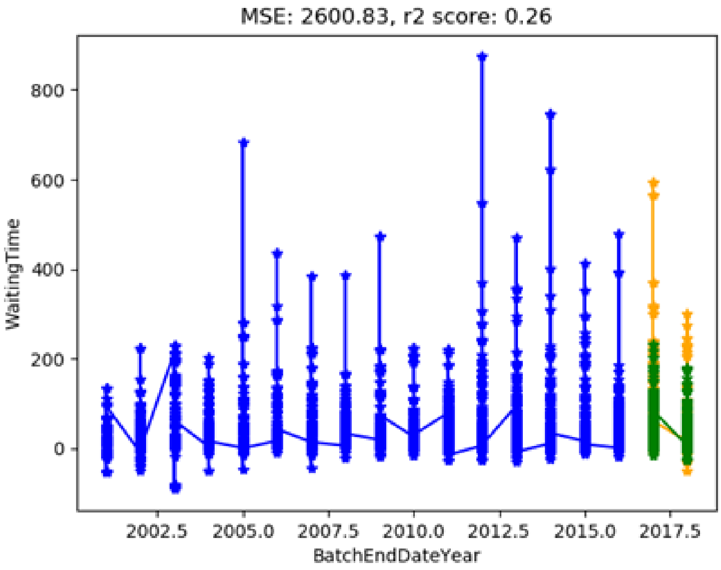


FIGURE 7.7: LR with PCA for All countries

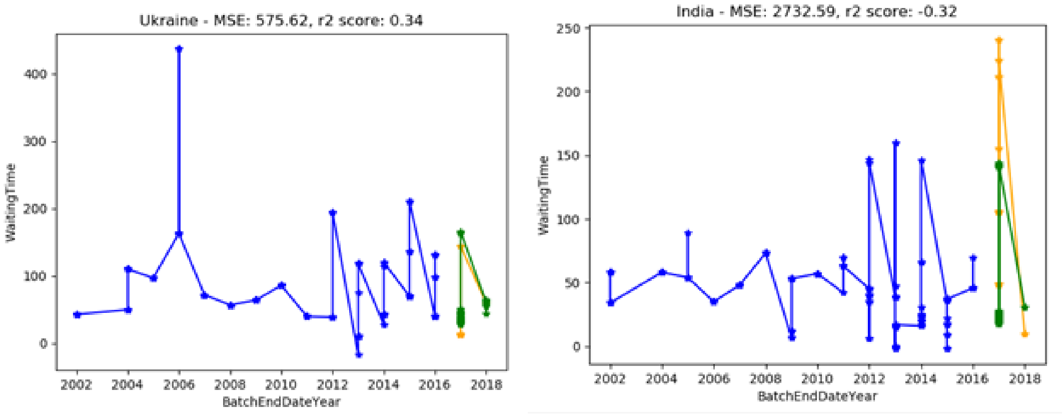


FIGURE 7.8: LR with PCA for Ukraine &India

7.1.2 Light GBM

LightGBM - fast gradient boosting model, explained in "Related Works and Background Information" chapter.

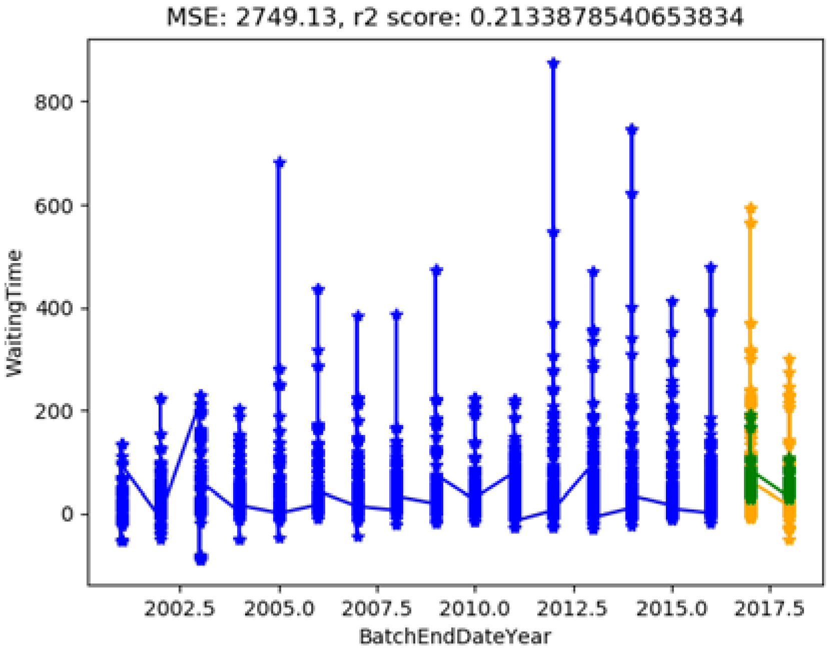


FIGURE 7.9: LightGBM for all countries

To tune parameter **GridSearch** from sklearn library and train dataset was used.

Tuned parameters: max_bin, learning_rate, num_leaves, subsample. Defined parameters: boosting_type=”gbdt”, objective=”regression”, n_jobs=3, metric=”mse”.

After tuning parameters, the best ones for available dataset were found:

’max_bin’: 275, ’num_leaves’: 5, ’learning_rate’: 0.02, ’subsample’: 0.03

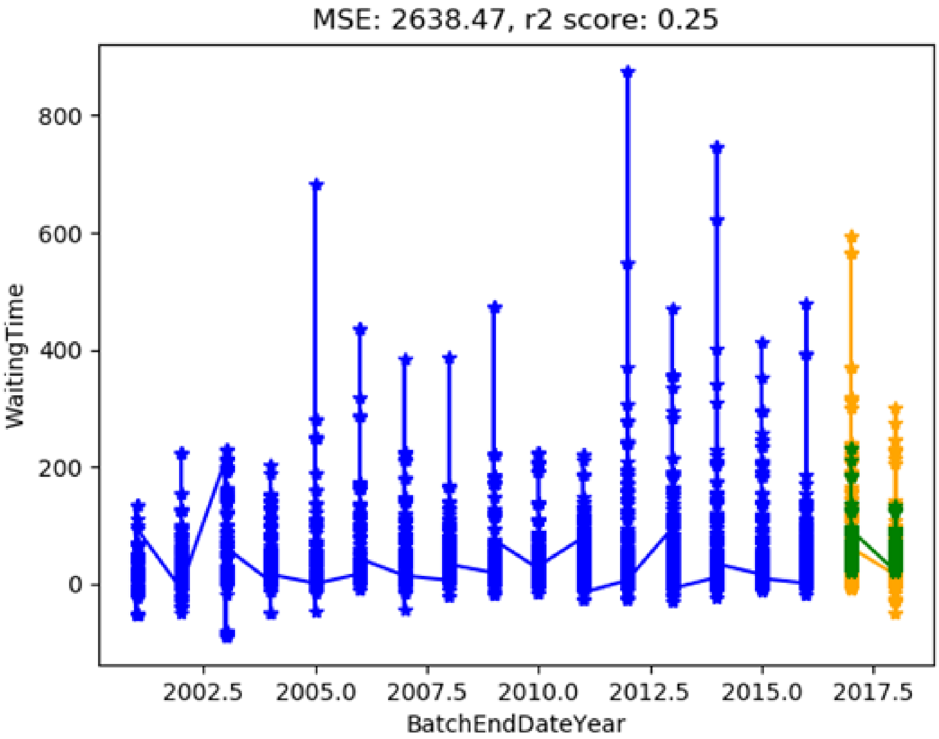


FIGURE 7.10: LightGBM with best parameters for all countries

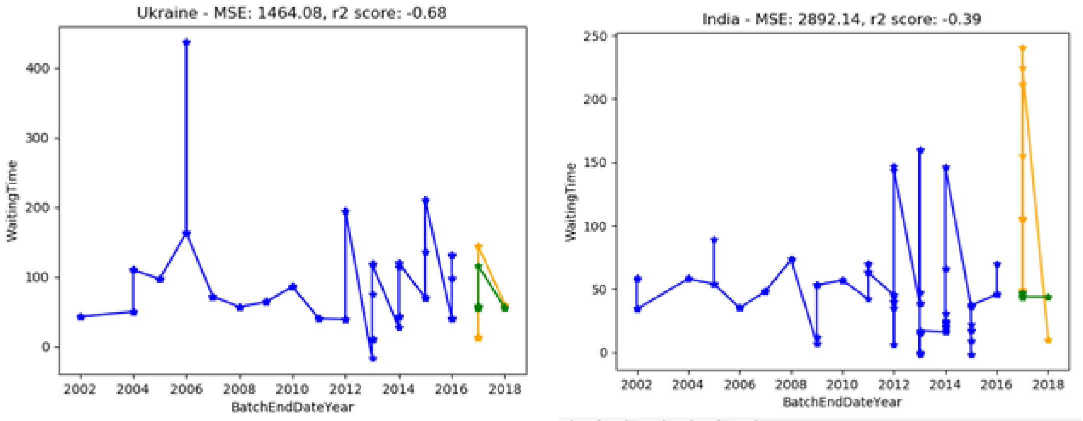


FIGURE 7.11: LightGBM with same parameters for Ukraine &India

7.1.3 ARIMA

Autoregressive Integrated Moving Average Model (ARIMA) is described in "Related Works and Background Information" chapter.
If time series has statistical properties, such as mean, variance, autocorrelation, constant over a time, it can be considered stationary. Only for stationary series, such statistics can be used for predicting future behavior.
Plotted BatchEndDateYear with WaitingTime (Receive date – Batch end date) for India (number of data=445)

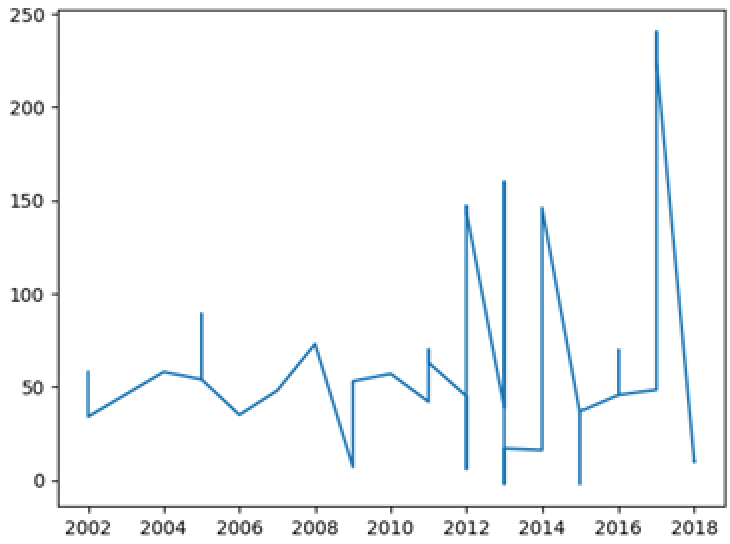


FIGURE 7.12: BatchEndDateYear with WaitingTime for India

This time series is not stationary and will require differencing to make it stationary, at least a difference order of 1. Correlating time series with itself with some amount of shift is what autocorrelation is. To calculate it we use original and shifted datasets, to find the coefficient of correlation between them.

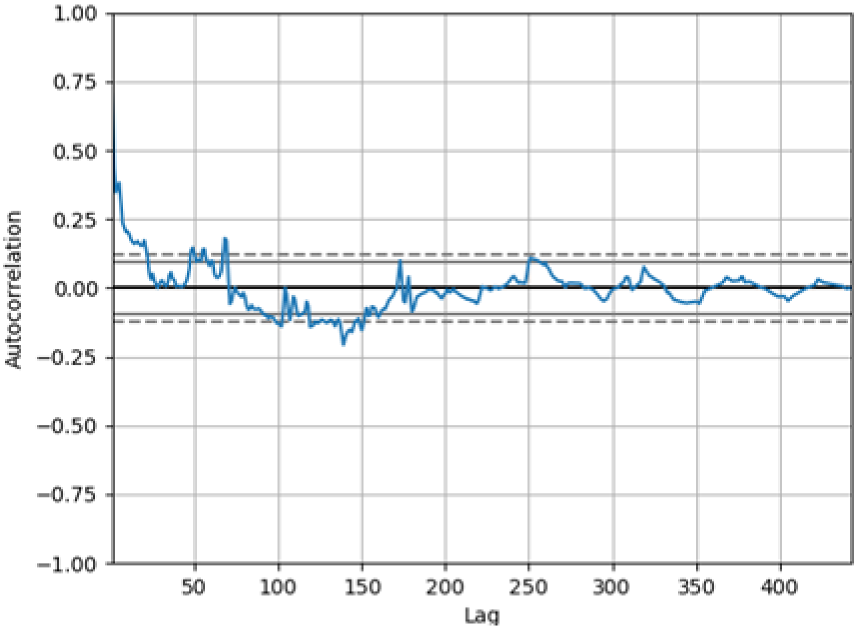
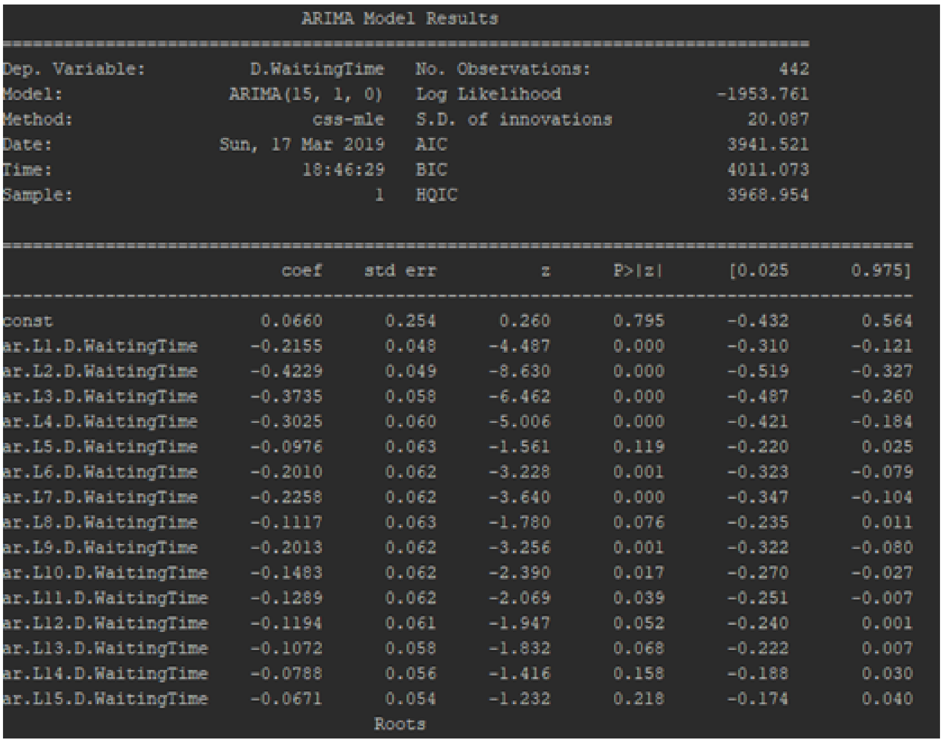


FIGURE 7.13: Autocorrelation for India

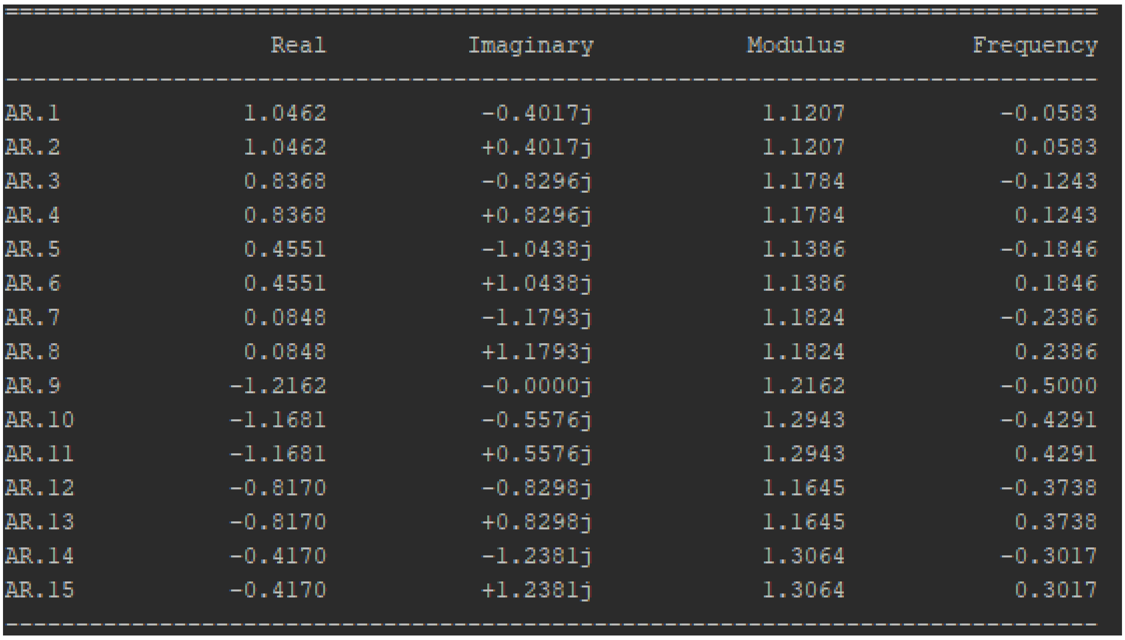
There is a positive correlation with the first approximately 75 lags that is perhaps significant for the first 15-25 lags.
A good starting point for the AR parameter of the model may be 15.
ARIMA(15,1,0) model sets the lag value to 15 for autoregression, uses a difference order of 1 to make the time series stationary, and uses a moving average model of 0.



```
                             ARIMA Model Results
==============================================================================
Dep. Variable:          D.WaitingTime   No. Observations:                  442
Model:                 ARIMA(15, 1, 0)   Log Likelihood               -1953.761
Method:                       css-mle   S.D. of innovations             20.087
Date:                Sun, 17 Mar 2019   AIC                           3941.521
Time:                        18:46:29   BIC                           4011.073
Sample:                             1   HQIC                          3968.954

========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                    0.0660      0.254      0.260      0.795      -0.432       0.564
ar.L1.D.WaitingTime     -0.2155      0.048     -4.487      0.000      -0.310      -0.121
ar.L2.D.WaitingTime     -0.4229      0.049     -8.630      0.000      -0.519      -0.327
ar.L3.D.WaitingTime     -0.3735      0.058     -6.462      0.000      -0.487      -0.260
ar.L4.D.WaitingTime     -0.3025      0.060     -5.006      0.000      -0.421      -0.184
ar.L5.D.WaitingTime     -0.0976      0.063     -1.561      0.119      -0.220       0.025
ar.L6.D.WaitingTime     -0.2010      0.062     -3.228      0.001      -0.323      -0.079
ar.L7.D.WaitingTime     -0.2258      0.062     -3.640      0.000      -0.347      -0.104
ar.L8.D.WaitingTime     -0.1117      0.063     -1.780      0.076      -0.235       0.011
ar.L9.D.WaitingTime     -0.2013      0.062     -3.256      0.001      -0.322      -0.080
ar.L10.D.WaitingTime    -0.1483      0.062     -2.390      0.017      -0.270      -0.027
ar.L11.D.WaitingTime    -0.1289      0.062     -2.069      0.039      -0.251      -0.007
ar.L12.D.WaitingTime    -0.1194      0.061     -1.947      0.052      -0.240       0.001
ar.L13.D.WaitingTime    -0.1072      0.058     -1.832      0.068      -0.222       0.007
ar.L14.D.WaitingTime    -0.0788      0.056     -1.416      0.158      -0.188       0.030
ar.L15.D.WaitingTime    -0.0671      0.054     -1.232      0.218      -0.174       0.040
                                    Roots
```

FIGURE 7.14: ARIMA model results for India



```
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.0462           -0.4017j            1.1207           -0.0583
AR.2            1.0462           +0.4017j            1.1207            0.0583
AR.3            0.8368           -0.8296j            1.1784           -0.1243
AR.4            0.8368           +0.8296j            1.1784            0.1243
AR.5            0.4551           -1.0438j            1.1386           -0.1846
AR.6            0.4551           +1.0438j            1.1386            0.1846
AR.7            0.0848           -1.1793j            1.1824           -0.2386
AR.8            0.0848           +1.1793j            1.1824            0.2386
AR.9           -1.2162           -0.0000j            1.2162           -0.5000
AR.10          -1.1681           -0.5576j            1.2943           -0.4291
AR.11          -1.1681           +0.5576j            1.2943            0.4291
AR.12          -0.8170           -0.8298j            1.1645           -0.3738
AR.13          -0.8170           +0.8298j            1.1645            0.3738
AR.14          -0.4170           -1.2381j            1.3064           -0.3017
AR.15          -0.4170           +1.2381j            1.3064            0.3017
-----------------------------------------------------------------------------
```

FIGURE 7.15: ARIMA model results for India 2

Line plot of the residual errors (Residual = Observed – Predicted)

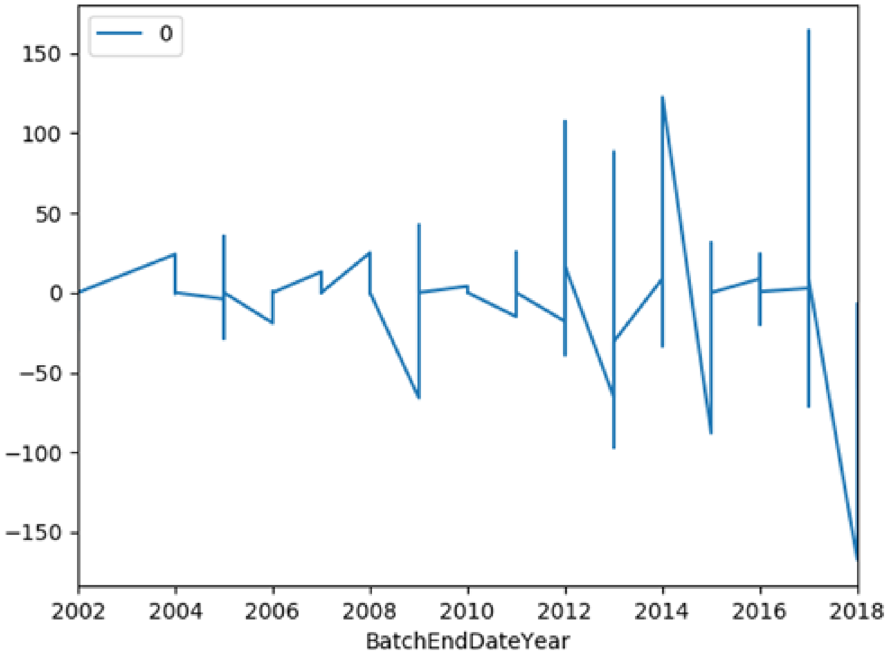


FIGURE 7.16: Residual errors

Density plot of the residual error values (Gaussian errors centered on zero)

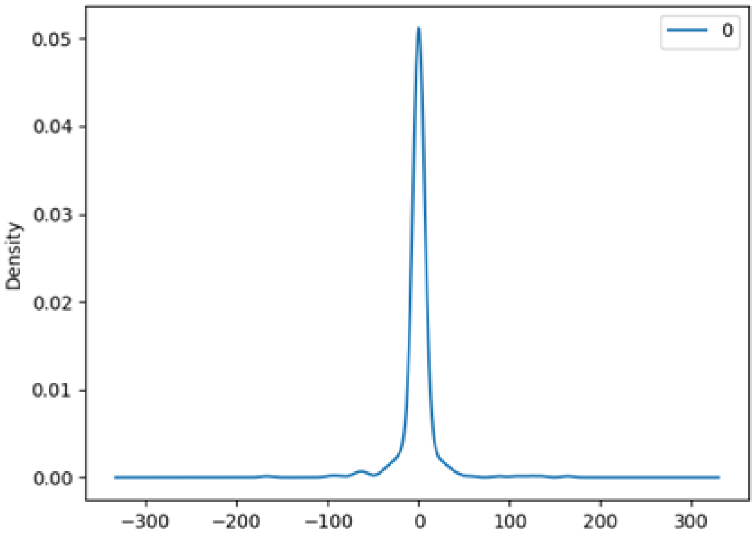
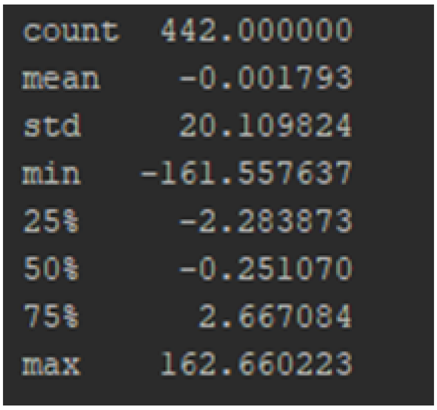


FIGURE 7.17: Density plot



```
count   442.000000
mean     -0.001793
std      20.109824
min    -161.557637
25%      -2.283873
50%      -0.251070
75%       2.667084
max     162.660223
```

FIGURE 7.18: Description of residuals

One-step forecast was chosen to use. Data is split to train and test sets. ARIMA model is fitted with train set to generate prediction for each element on the test set. Such forecast has a dependence on previous observations for differencing and AR model. New ARIMA model is recreated after every receiving new observation.

ARIMA(15, 1, 0) with Test MSE: 1160.322 and Test r2score:0.382684835536657, on 80% train
red = predictions, blue = true test data

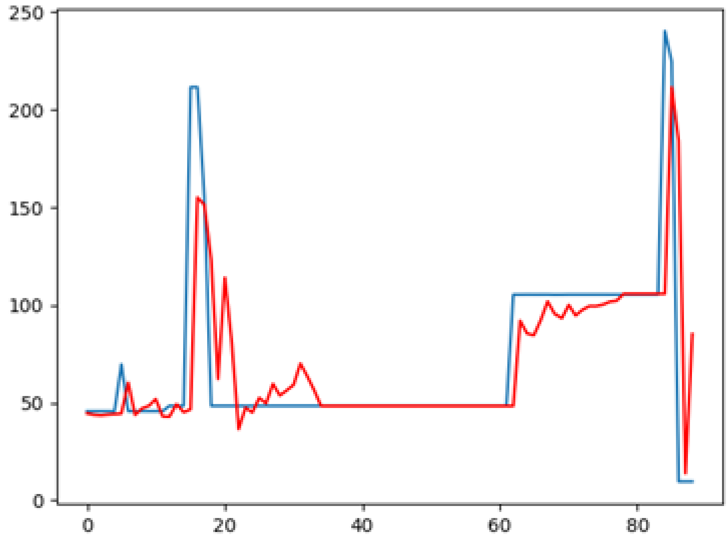


FIGURE 7.19: India - ARIMA(15, 1, 0)

Tuning Parameters using GridSearch

Parameters that were tuned:

p_values = [0, 1, 2, 4, 6, 10]

d_values = range (0, 3)

q_values = range (0, 3)

Results:

Best Parameters – ARIMA(1, 1, 2) with MSE=847.648 and Test r2score: 0.4985152749736379 on 66% train
ARIMA (1, 1, 2) with Test MSE: 1078.122 and Test r2score: 0.42641671300132955 on 80% train

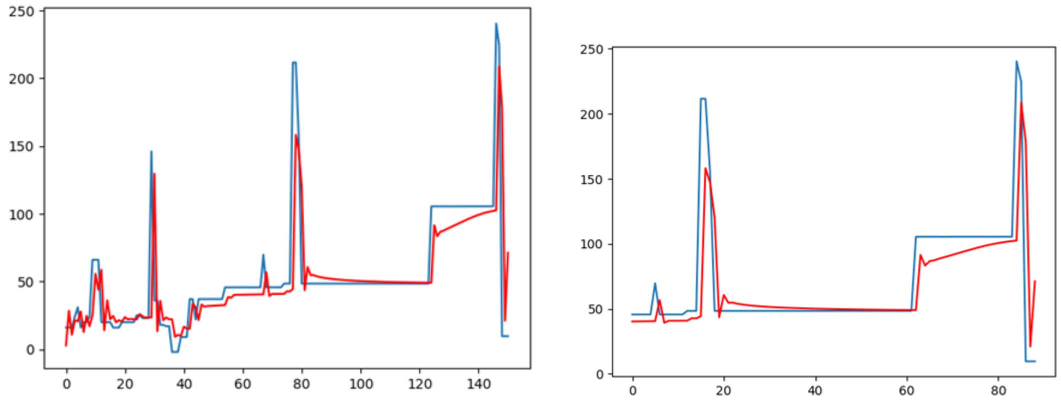


FIGURE 7.20: India - ARIMA(1, 1, 2) on 66% and 80% train

Arima for all countries

ARIMA (15,1,0) with Test MSE: 1745.930 and Test r2score: 0.415252718711274 on 80% train

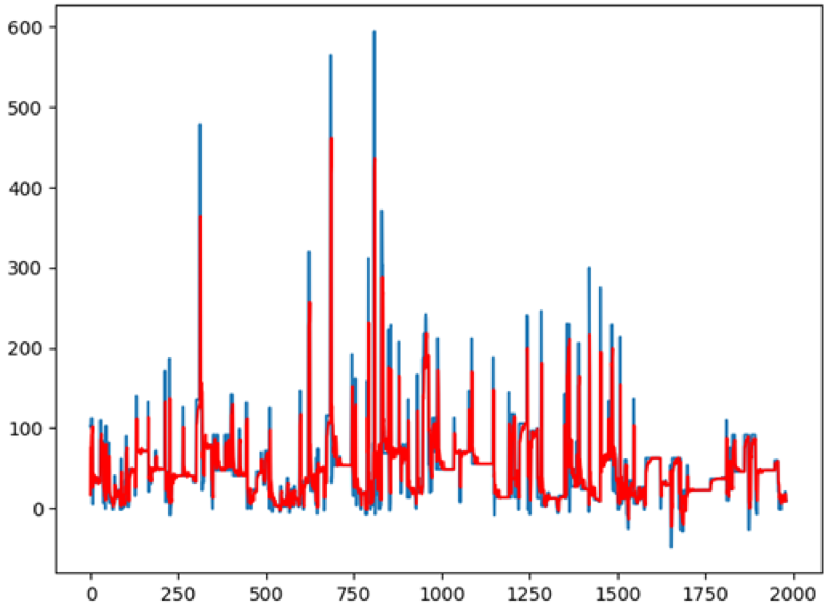


FIGURE 7.21: All countries - ARIMA(15,1,0) on 80% train

Best Parameters found by Grid Search (4, 0, 1) with Test MSE: 1685.807 and Test r2score: 0.43538895221550666 on 80%

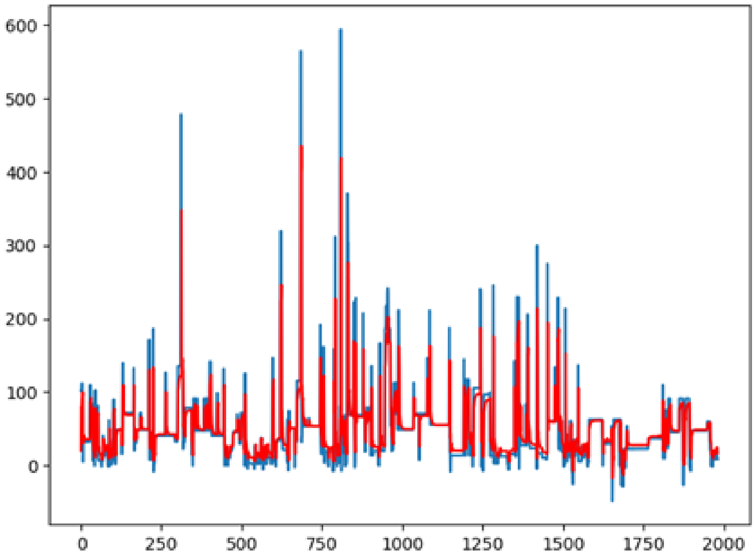
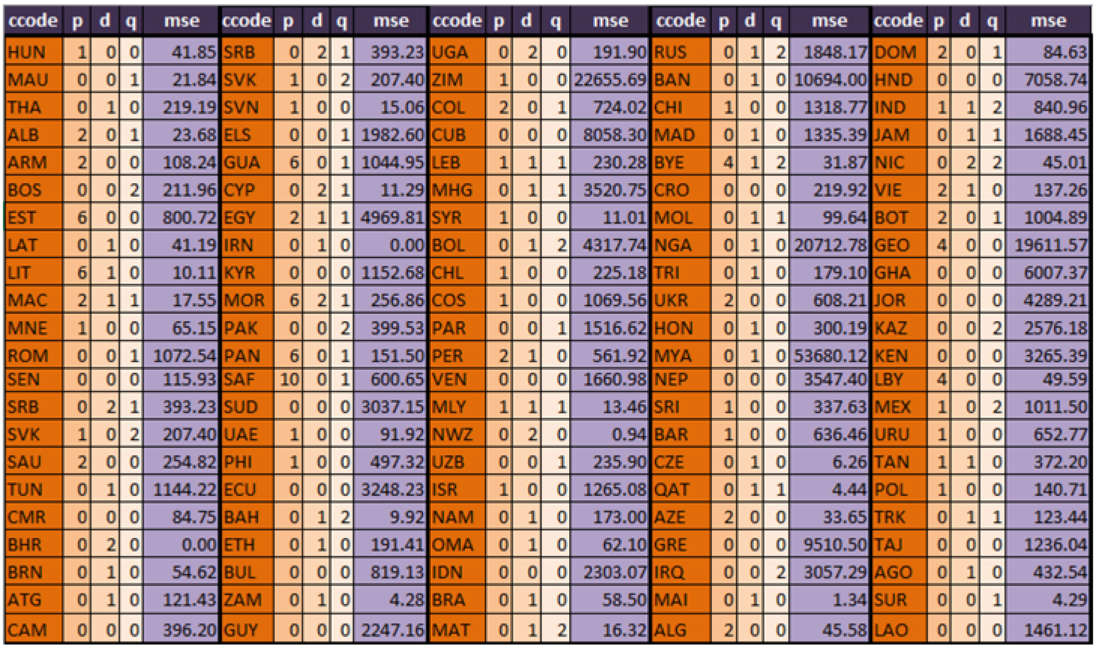


FIGURE 7.22: All countries - ARIMA(4,1,0) on 80% train

Tuning parameters for each country returned best parameters (p, d, q)



ccode	p	d	q	mse	ccode	p	d	q	mse	ccode	p	d	q	mse	ccode	p	d	q	mse	ccode	p	d	q	mse
HUN	1	0	0	41.85	SRB	0	2	1	393.23	UGA	0	2	0	191.90	RUS	0	1	2	1848.17	DOM	2	0	1	84.63
MAU	0	0	1	21.84	SVK	1	0	2	207.40	ZIM	1	0	0	22655.69	BAN	0	1	0	10694.00	HND	0	0	0	7058.74
THA	0	1	0	219.19	SVN	1	0	0	15.06	COL	2	0	1	724.02	CHI	1	0	0	1318.77	IND	1	1	2	840.96
ALB	2	0	1	23.68	ELS	0	0	1	1982.60	CUB	0	0	0	8058.30	MAD	0	1	0	1335.39	JAM	0	1	1	1688.45
ARM	2	0	0	108.24	GUA	6	0	1	1044.95	LEB	1	1	1	230.28	BYE	4	1	2	31.87	NIC	0	2	2	45.01
BOS	0	0	2	211.96	CYP	0	2	1	11.29	MHG	0	1	1	3520.75	CRO	0	0	0	219.92	VIE	2	1	0	137.26
EST	6	0	0	800.72	EGY	2	1	1	4969.81	SYR	1	0	0	11.01	MOL	0	1	1	99.64	BOT	2	0	1	1004.89
LAT	0	1	0	41.19	IRN	0	1	0	0.00	BOL	0	1	2	4317.74	NGA	0	1	0	20712.78	GEO	4	0	0	19611.57
LIT	6	1	0	10.11	KYR	0	0	0	1152.68	CHL	1	0	0	225.18	TRI	0	1	0	179.10	GHA	0	0	0	6007.37
MAC	2	1	1	17.55	MOR	6	2	1	256.86	COS	1	0	0	1069.56	UKR	2	0	0	608.21	JOR	0	0	0	4289.21
MNE	1	0	0	65.15	PAK	0	0	2	399.53	PAR	0	0	1	1516.62	HON	0	1	0	300.19	KAZ	0	0	2	2576.18
ROM	0	0	1	1072.54	PAN	6	0	1	151.50	PER	2	1	0	561.92	MYA	0	1	0	53680.12	KEN	0	0	0	3265.39
SEN	0	0	0	115.93	SAF	10	0	1	600.65	VEN	0	0	0	1660.98	NEP	0	0	0	3547.40	LBY	4	0	0	49.59
SRB	0	2	1	393.23	SUD	0	0	0	3037.15	MLY	1	1	1	13.46	SRI	1	0	0	337.63	MEX	1	0	2	1011.50
SVK	1	0	2	207.40	UAE	1	0	0	91.92	NWZ	0	2	0	0.94	BAR	1	0	0	636.46	URU	1	0	0	652.77
SAU	2	0	0	254.82	PHI	1	0	0	497.32	UZB	0	0	1	235.90	CZE	0	1	0	6.26	TAN	1	1	0	372.20
TUN	0	1	0	1144.22	ECU	0	0	0	3248.23	ISR	1	0	0	1265.08	QAT	0	1	1	4.44	POL	1	0	0	140.71
CMR	0	0	0	84.75	BAH	0	1	2	9.92	NAM	0	1	0	173.00	AZE	2	0	0	33.65	TRK	0	1	1	123.44
BHR	0	2	0	0.00	ETH	0	1	0	191.41	OMA	0	1	0	62.10	GRE	0	0	0	9510.50	TAJ	0	0	0	1236.04
BRN	0	1	0	54.62	BUL	0	0	0	819.13	IDN	0	0	0	2303.07	IRQ	0	0	2	3057.29	AGO	0	1	0	432.54
ATG	0	1	0	121.43	ZAM	0	1	0	4.28	BRA	0	1	0	58.50	MAI	0	1	0	1.34	SUR	0	0	1	4.29
CAM	0	0	0	396.20	GUY	0	0	0	2247.16	MAT	0	1	2	16.32	ALG	2	0	0	45.58	LAO	0	0	0	1461.12

FIGURE 7.23: Best ARIMA parameters for each country

Using these parameters forecasts were created and saved to excel for comparing later with results from LSTM.

7.1.4 LSTM

A rolling-forecast method was used (same as with arima). At one time model make a forecast for one time step from test data. Then model is evaluated by comparing actual and predicted values. After that, the actual value becomes available for model

to do prediction for the next time step.
Pre-processing data included scaling, differencing and transformation.

1. Features are scaled by MinMaxScaler to have values between -1 and 1 that corresponds to default hyperbolic tangent activation function of LSTM.

2. To make series more stationary, differencing was applied (as with arima).

3. Also, data was transformed, so that actual value from the previous time step can be used for predicting at the current time step.

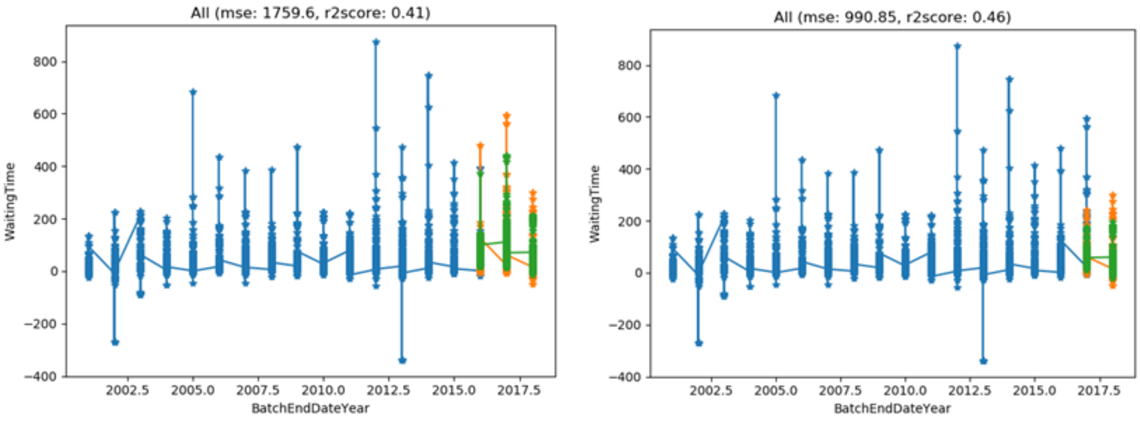


FIGURE 7.24: LSTM - All countries on 80% and 90% train

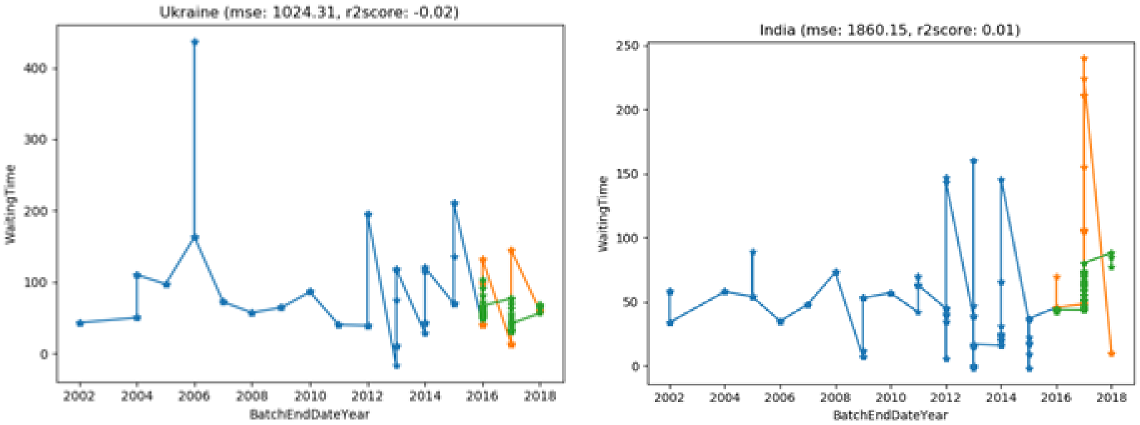


FIGURE 7.25: LSTM - Ukraine and India on 80% train

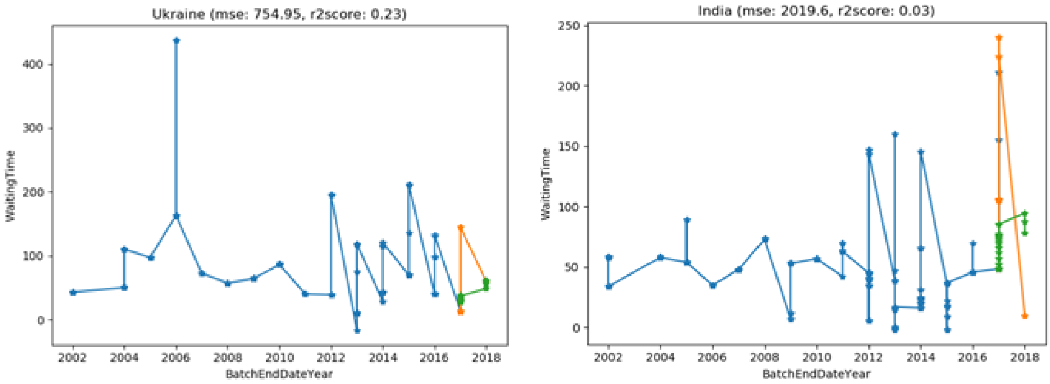


FIGURE 7.26: LSTM - Ukraine and India on 90% train

Tuning parameters

Each experiment is run 10 times because of random initial configuration for each training that can lead to different LSTM results.

1. Tuning the number of training epochs. (On India)
A line plot with train and test RMSE scores for each epoch was created.
Test(blue) and Train(yellow) RMSE

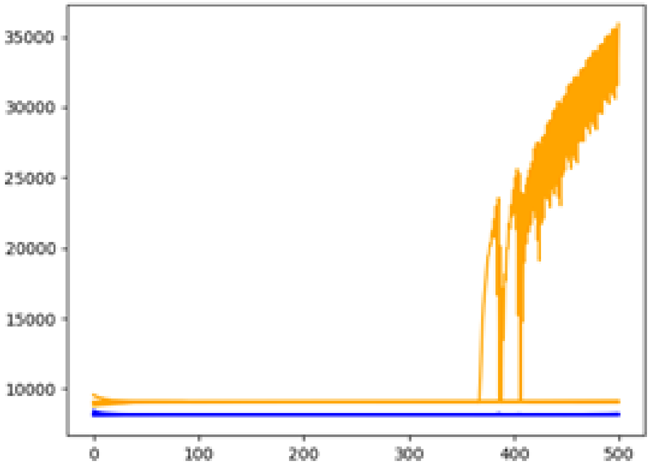


FIGURE 7.27: India (80% train): (1-neuron, 1-batch size, 500-epochs)

The model does not learn, and loss is not decreasing with the number of epochs. India’s data did not give good results. In the case with Ukraine, better results are expected.

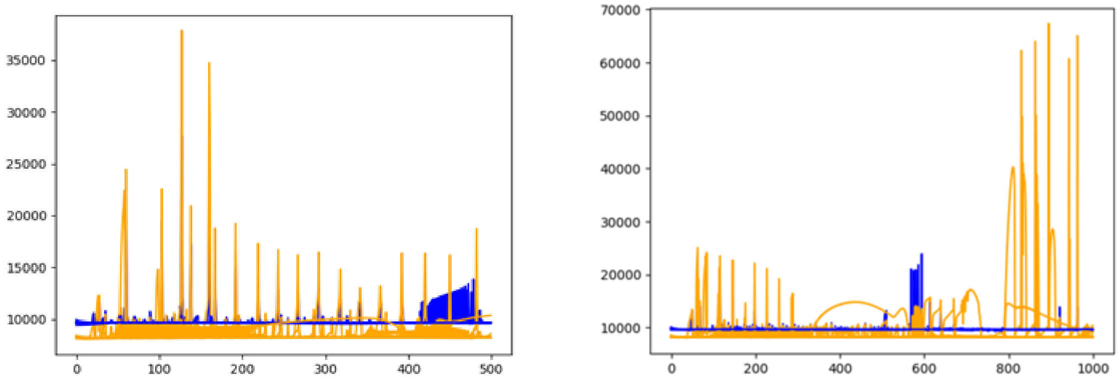
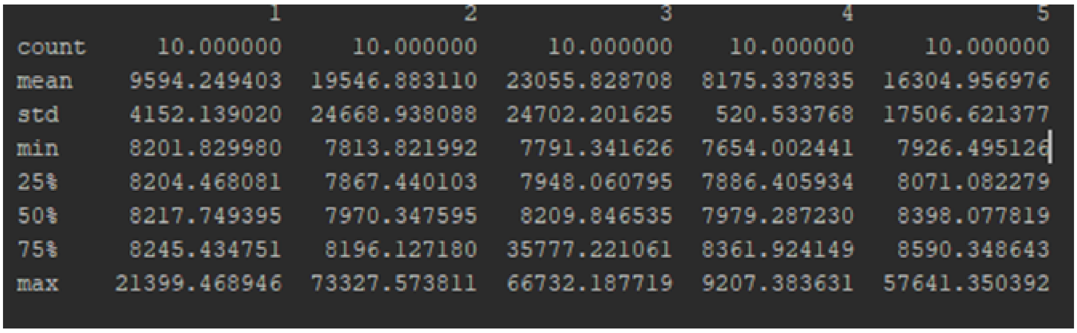


FIGURE 7.28: Ukraine (80% train): (1-neuron, 1-batch size, 500 & 1000-epochs)

The figure shows that the optimal number of epochs for Ukraine will be 350 (with minimum RMSE). Till 350 epoch, model's ability to learn increases (error decreases) but after we can see the increasing trend in error. This is a sign of overfitting.

2. Tuning the number of neurons

The number of neurons affects learning capacity for the network. More neurons can help to learn better problem but can cause overfitting as well. To choose the best number of neurons, 10 repeats for each number from [1,2,3,4,5] were run and described for Ukraine.



```
                 1             2             3            4             5
count    10.000000     10.000000     10.000000    10.000000     10.000000
mean   9594.249403  19546.883110  23055.828708  8175.337835  16304.956976
std    4152.139020  24668.938088  24702.201625   520.533768  17506.621377
min    8201.829980   7813.821992   7791.341626  7654.002441   7926.495126
25%    8204.468081   7867.440103   7948.060795  7886.405934   8071.082279
50%    8217.749395   7970.347595   8209.846535  7979.287230   8398.077819
75%    8245.434751   8196.127180  35777.221061  8361.924149   8590.348643
max   21399.468946  73327.573811  66732.187719  9207.383631  57641.350392
```

FIGURE 7.29: Description for Ukraine: ([1,2,3,4,5]-neuron, 1-batch size, 350-epochs)

Figure shows that 4 neurons is the best option without any doubts. (all RMSE statistics are the lowest)

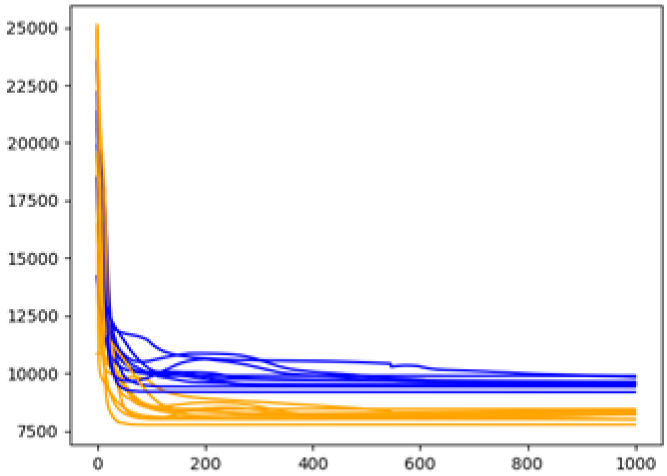


FIGURE 7.30: Ukraine (80% train): (4-neuron, 1-batch size, 350-epochs)

3. Tuning the batch size

Batch size defines how often weights for the network should be updated. In Keras, it depends on the size of train and test data.
Before the batch size of 1 was used. Weights were updated after each epoch.

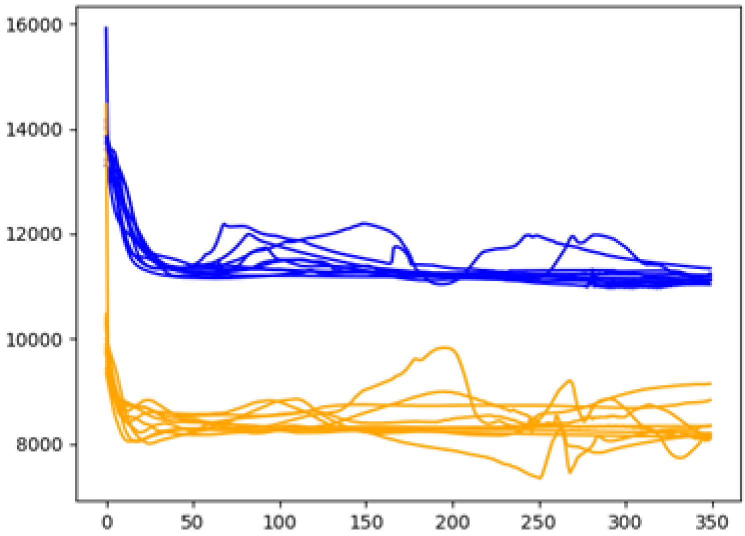


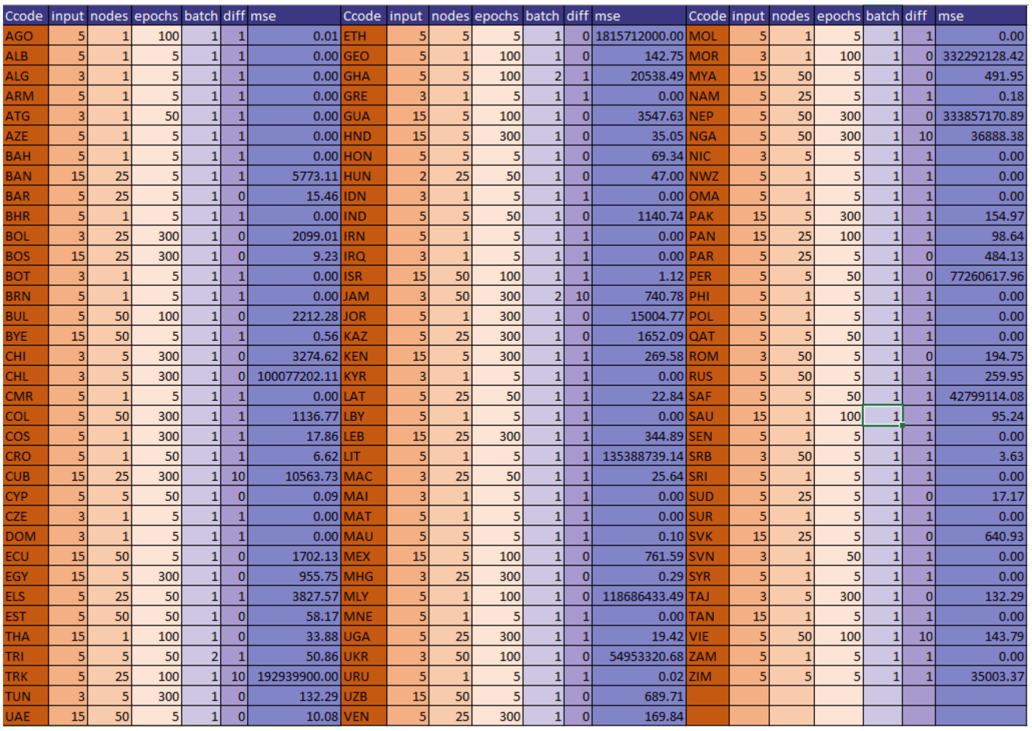
FIGURE 7.31: Ukraine (460 train, 88 test): (4-neuron, 2-batch size, 350 - epochs)

The plot shows more variability in the RMSE over time and it seems that RMSE will not decrease with more epochs.
Results: best parameters for Ukraine: (4-neurons, 1-batch size, 350-epochs)

Tuning parameters for all countries

GridSearch was used to tune parameters for each country. Parameters that were tuned:
n_input - The number of previous inputs to use as input for the model,
n_nodes - The number of nodes for the hidden layer,
n_epochs - The number of training epochs,
n_batch - The number of samples that each batch should include,
n_diff - The order of difference.

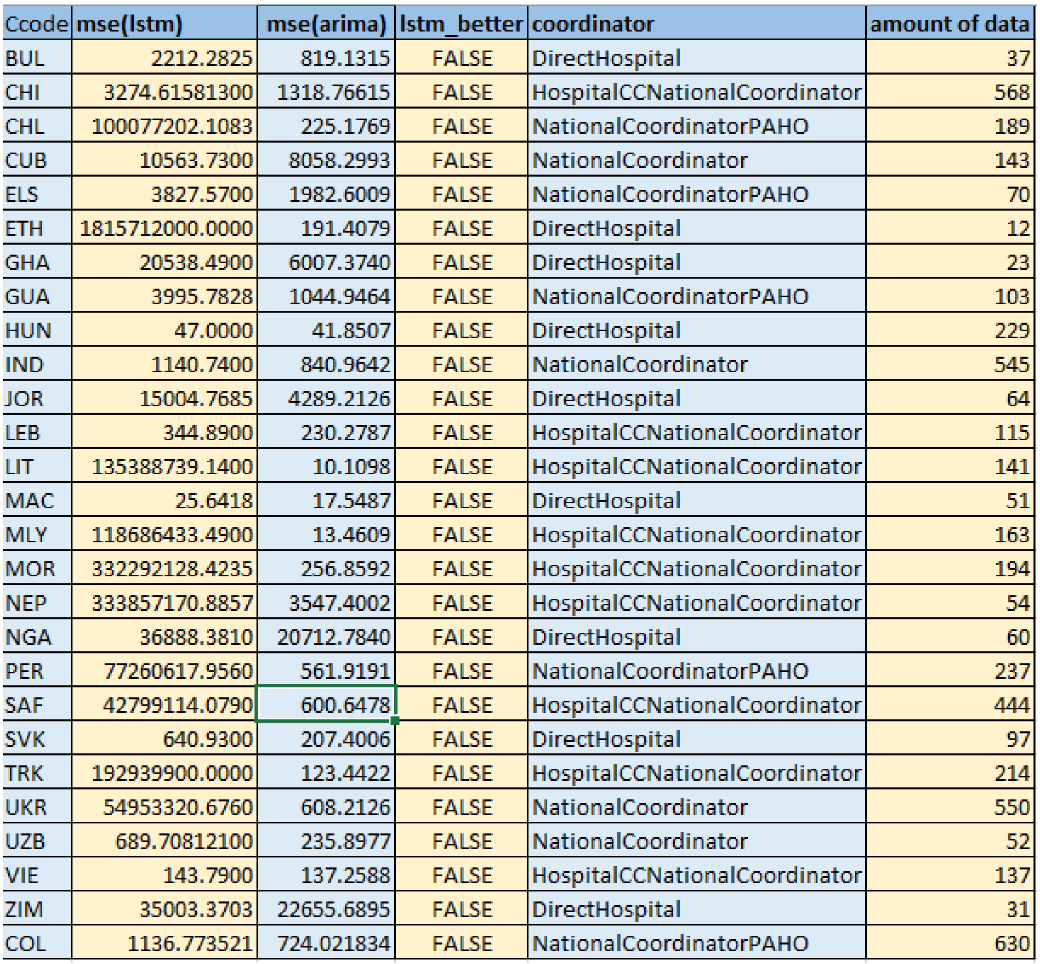
To tune 115 countries Azure Machine Learning Studio was used. Results with the best parameters for each country are presented below.



Ccode	input	nodes	epochs	batch	diff	mse	Ccode	input	nodes	epochs	batch	diff	mse	Ccode	input	nodes	epochs	batch	diff	mse
AGO	5	1	100	1	1	0.01	ETH	5	5	5	1	0	1815712000.00	MOL	5	1	5	1	1	0.00
ALB	5	1	5	1	1	0.00	GEO	5	1	100	1	0	142.75	MOR	3	1	100	1	0	332292128.42
ALG	3	1	5	1	1	0.00	GHA	5	5	100	2	1	20538.49	MYA	15	50	5	1	0	491.95
ARM	5	1	5	1	1	0.00	GRE	3	1	5	1	1	0.00	NAM	5	25	5	1	1	0.18
ATG	3	1	50	1	1	0.00	GUA	15	5	100	1	0	3547.63	NEP	5	50	300	1	0	333857170.89
AZE	5	1	5	1	1	0.00	HND	15	5	300	1	0	35.05	NGA	5	50	300	1	10	36888.38
BAH	5	1	5	1	1	0.00	HON	5	5	5	1	0	69.34	NIC	3	5	5	1	1	0.00
BAN	15	25	5	1	1	5773.11	HUN	2	25	50	1	0	47.00	NWZ	5	1	5	1	1	0.00
BAR	5	25	5	1	0	15.46	IDN	3	1	5	1	1	0.00	OMA	5	1	5	1	1	0.00
BHR	5	1	5	1	1	0.00	IND	5	5	50	1	0	1140.74	PAK	15	5	300	1	1	154.97
BOL	3	25	300	1	0	2099.01	IRN	5	1	5	1	1	0.00	PAN	15	25	100	1	1	98.64
BOS	15	25	300	1	0	9.23	IRQ	3	1	5	1	1	0.00	PAR	5	25	5	1	0	484.13
BOT	3	1	5	1	1	0.00	ISR	15	50	100	1	1	1.12	PER	5	5	50	1	0	77260617.96
BRN	5	1	5	1	1	0.00	JAM	3	50	300	2	10	740.78	PHI	5	1	5	1	1	0.00
BUL	5	50	100	1	0	2212.28	JOR	5	1	300	1	0	15004.77	POL	5	1	5	1	1	0.00
BYE	15	50	5	1	1	0.56	KAZ	5	25	300	1	0	1652.09	QAT	5	5	50	1	1	0.00
CHI	3	5	300	1	0	3274.62	KEN	15	5	300	1	1	269.58	ROM	3	50	5	1	0	194.75
CHL	3	5	300	1	0	100077202.11	KYR	3	1	5	1	1	0.00	RUS	5	50	5	1	1	259.95
CMR	5	1	5	1	1	0.00	LAT	5	25	50	1	1	22.84	SAF	5	5	50	1	1	42799114.08
COL	5	50	300	1	1	1136.77	LBY	5	1	5	1	1	0.00	SAU	15	1	100	1	1	95.24
COS	5	1	300	1	1	17.86	LEB	15	25	300	1	1	344.89	SEN	5	1	5	1	1	0.00
CRO	5	1	50	1	1	6.62	LIT	5	1	5	1	1	135388739.14	SRB	3	50	5	1	1	3.63
CUB	15	25	300	1	10	10563.73	MAC	3	25	50	1	1	25.64	SRI	5	1	5	1	1	0.00
CYP	5	5	50	1	0	0.09	MAI	3	1	5	1	1	0.00	SUD	5	25	5	1	0	17.17
CZE	3	1	5	1	1	0.00	MAT	5	1	5	1	1	0.00	SUR	5	1	5	1	1	0.00
DOM	3	1	5	1	1	0.00	MAU	5	5	5	1	1	0.10	SVK	15	25	5	1	0	640.93
ECU	15	50	5	1	0	1702.13	MEX	15	5	100	1	0	761.59	SVN	3	1	50	1	1	0.00
EGY	15	5	300	1	0	955.75	MHG	3	25	300	1	0	0.29	SYR	5	1	5	1	1	0.00
ELS	5	25	50	1	1	3827.57	MLY	5	1	100	1	0	118686433.49	TAJ	3	5	300	1	0	132.29
EST	5	50	50	1	0	58.17	MNE	5	1	5	1	1	0.00	TAN	15	1	5	1	1	0.00
THA	15	1	100	1	0	33.88	UGA	5	25	300	1	1	19.42	VIE	5	50	100	1	10	143.79
TRI	5	5	50	2	1	50.86	UKR	3	50	100	1	0	54953320.68	ZAM	5	1	5	1	1	0.00
TRK	5	25	100	1	10	192939900.00	URU	5	1	5	1	1	0.02	ZIM	5	5	5	1	1	35003.37
TUN	3	5	300	1	0	132.29	UZB	15	50	5	1	0	689.71							
UAE	15	50	5	1	0	10.08	VEN	5	25	300	1	0	169.84							

FIGURE 7.32: LSTM tuned parameters per country

To choose the best algorithm for predicting, LSTM and ARIMA results were compared. Since Linear Regression and Light GBM did not have good results, they were not taken for comparison. Below you can find a table with results for countries with better predictions generated by ARIMA algorithm. There are 27 such countries, other countries have better predictions created by LSTM or have 3 and less data point. It would be interesting to find any dependence of better algorithm on some features of data/country. For this reason, the number of data and distribution method were included in the table. But unfortunately, direct dependence was not found.



Ccode	mse(lstm)	mse(arima)	lstm_better	coordinator	amount of data
BUL	2212.2825	819.1315	FALSE	DirectHospital	37
CHI	3274.61581300	1318.76615	FALSE	HospitalCCNationalCoordinator	568
CHL	100077202.1083	225.1769	FALSE	NationalCoordinatorPAHO	189
CUB	10563.7300	8058.2993	FALSE	NationalCoordinator	143
ELS	3827.5700	1982.6009	FALSE	NationalCoordinatorPAHO	70
ETH	1815712000.0000	191.4079	FALSE	DirectHospital	12
GHA	20538.4900	6007.3740	FALSE	DirectHospital	23
GUA	3995.7828	1044.9464	FALSE	NationalCoordinatorPAHO	103
HUN	47.0000	41.8507	FALSE	DirectHospital	229
IND	1140.7400	840.9642	FALSE	NationalCoordinator	545
JOR	15004.7685	4289.2126	FALSE	DirectHospital	64
LEB	344.8900	230.2787	FALSE	HospitalCCNationalCoordinator	115
LIT	135388739.1400	10.1098	FALSE	HospitalCCNationalCoordinator	141
MAC	25.6418	17.5487	FALSE	DirectHospital	51
MLY	118686433.4900	13.4609	FALSE	HospitalCCNationalCoordinator	163
MOR	332292128.4235	256.8592	FALSE	HospitalCCNationalCoordinator	194
NEP	333857170.8857	3547.4002	FALSE	HospitalCCNationalCoordinator	54
NGA	36888.3810	20712.7840	FALSE	DirectHospital	60
PER	77260617.9560	561.9191	FALSE	NationalCoordinatorPAHO	237
SAF	42799114.0790	600.6478	FALSE	HospitalCCNationalCoordinator	444
SVK	640.9300	207.4006	FALSE	DirectHospital	97
TRK	192939900.0000	123.4422	FALSE	HospitalCCNationalCoordinator	214
UKR	54953320.6760	608.2126	FALSE	NationalCoordinator	550
UZB	689.70812100	235.8977	FALSE	NationalCoordinator	52
VIE	143.7900	137.2588	FALSE	HospitalCCNationalCoordinator	137
ZIM	35003.3703	22655.6895	FALSE	DirectHospital	31
COL	1136.773521	724.021834	FALSE	NationalCoordinatorPAHO	630

FIGURE 7.33: ARIMA VS LSTM MSE comparison

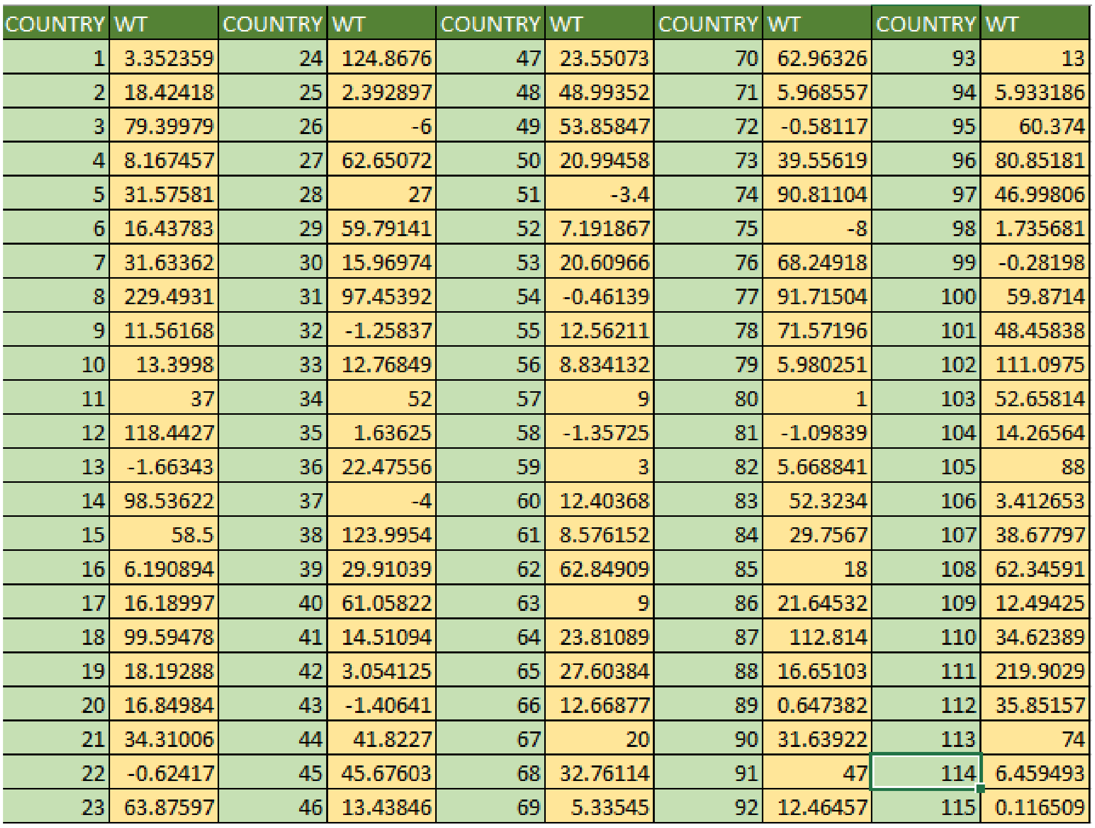
In general, LSTM works better than ARIMA for an available dataset. But to have the best results for each country, both algorithms will be used.

LSTM was tuned just for countries with more than 5 data points while ARIMA was tuned only for countries with more than 3 data points. So for those countries that have less or equal 5 but more than 3 data points, ARIMA forecast was used as final. For those countries that have 3 or less data points, the last value from historical data was used as forecast.

For 76 countries - LSTM forecast was used.

For 32 countries - ARIMA forecast was used.

For 7 countries - the last historical value was used.



COUNTRY	WT	COUNTRY	WT	COUNTRY	WT	COUNTRY	WT	COUNTRY	WT
1	3.352359	24	124.8676	47	23.55073	70	62.96326	93	13
2	18.42418	25	2.392897	48	48.99352	71	5.968557	94	5.933186
3	79.39979	26	-6	49	53.85847	72	-0.58117	95	60.374
4	8.167457	27	62.65072	50	20.99458	73	39.55619	96	80.85181
5	31.57581	28	27	51	-3.4	74	90.81104	97	46.99806
6	16.43783	29	59.79141	52	7.191867	75	-8	98	1.735681
7	31.63362	30	15.96974	53	20.60966	76	68.24918	99	-0.28198
8	229.4931	31	97.45392	54	-0.46139	77	91.71504	100	59.8714
9	11.56168	32	-1.25837	55	12.56211	78	71.57196	101	48.45838
10	13.3998	33	12.76849	56	8.834132	79	5.980251	102	111.0975
11	37	34	52	57	9	80	1	103	52.65814
12	118.4427	35	1.63625	58	-1.35725	81	-1.09839	104	14.26564
13	-1.66343	36	22.47556	59	3	82	5.668841	105	88
14	98.53622	37	-4	60	12.40368	83	52.3234	106	3.412653
15	58.5	38	123.9954	61	8.576152	84	29.7567	107	38.67797
16	6.190894	39	29.91039	62	62.84909	85	18	108	62.34591
17	16.18997	40	61.05822	63	9	86	21.64532	109	12.49425
18	99.59478	41	14.51094	64	23.81089	87	112.814	110	34.62389
19	18.19288	42	3.054125	65	27.60384	88	16.65103	111	219.9029
20	16.84984	43	-1.40641	66	12.66877	89	0.647382	112	35.85157
21	34.31006	44	41.8227	67	20	90	31.63922	113	74
22	-0.62417	45	45.67603	68	32.76114	91	47	114	6.459493
23	63.87597	46	13.43846	69	5.33545	92	12.46457	115	0.116509

FIGURE 7.34: Waiting time predictions

Because of IAEA security rules, countries are anonymized in this table.

7.2 Predicting number of sets

7.2.1 Exponential Smoothing

To predict the number of sets that we expect to send to each country exponential smoothing was used.

Simple (Brown's) Exponential Smoothing - method of forecasting that uses weighted averages of past observations, following an exponential decay, when old observations have less weight and importance on forecast then more recent ones. It does not take in a count seasonal and trend components. (Brown, 1963)

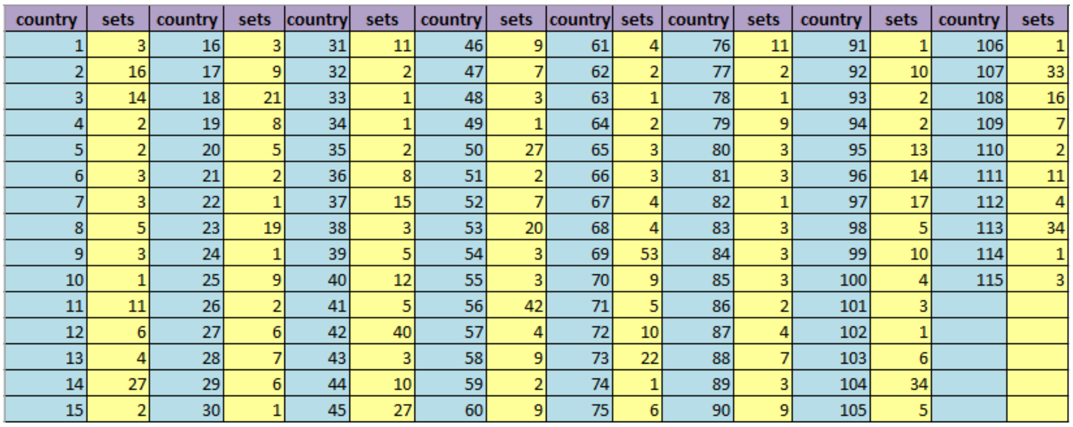
Dataset was prepared for predicting the number of sets to have the form: country, year, the number of sets.
Since there are a few data points and seasonality does not matter, exponential smoothing was used.

Prediction at time t is equal to a component called level, that is a weighted average of the previous level and the current observation with a smoothing parameter denoted as "alpha".

$$p_t = l_t$$

$$l_t = \alpha y_t + (1-\alpha) l_{t-1}$$

Results:



country	sets	country	sets	country	sets	country	sets	country	sets	country	sets	country	sets	country	sets
1	3	16	3	31	11	46	9	61	4	76	11	91	1	106	1
2	16	17	9	32	2	47	7	62	2	77	2	92	10	107	33
3	14	18	21	33	1	48	3	63	1	78	1	93	2	108	16
4	2	19	8	34	1	49	1	64	2	79	9	94	2	109	7
5	2	20	5	35	2	50	27	65	3	80	3	95	13	110	2
6	3	21	2	36	8	51	2	66	3	81	3	96	14	111	11
7	3	22	1	37	15	52	7	67	4	82	1	97	17	112	4
8	5	23	19	38	3	53	20	68	4	83	3	98	5	113	34
9	3	24	1	39	5	54	3	69	53	84	3	99	10	114	1
10	1	25	9	40	12	55	3	70	9	85	3	100	4	115	3
11	11	26	2	41	5	56	42	71	5	86	2	101	3		
12	6	27	6	42	40	57	4	72	10	87	4	102	1		
13	4	28	7	43	3	58	9	73	22	88	7	103	6		
14	27	29	6	44	10	59	2	74	1	89	3	104	34		
15	2	30	1	45	27	60	9	75	6	90	9	105	5		

FIGURE 7.35: Predicted number of sets per country

Because of IAEA security rules, countries are anonymized in this table.

7.3 Scheduling

The goal of the project is to balance DOL work intensity. The main part of work is carried out after DOL receives sets. It has to read, check dosimeters, prepare report and sign papers. After this process finished, DOL can send results of audit back to the hospital. The DOL work would be more balanced if it receives almost the same number of sets each month. We can not separate sets of one country to different months. So, the task is to group the predicted amount of sets per country in such a way, that sum of received sets is almost the same for each month.

7.3.1 LRM

LRM - first method, that was used. The problem was found, because LRM creates groups with the same number of countries making sums of sets as close as possible but not the same. For our task, it does not matter how much countries are included in a group, but the sum of sets is important.

7.3.2 Bin Packing

Bin packing algorithm can be used for grouping countries in a fixed number of bins (12) with almost the same number of sets and any number of countries in one bin. This algorithm considered to be one of NP-hard problems and simple version of it uses a greedy approach. The algorithm tries to put an element into the first bin that can accommodate it. In case, the bin does not have enough space anymore, it opens a new bin and put the element into it.

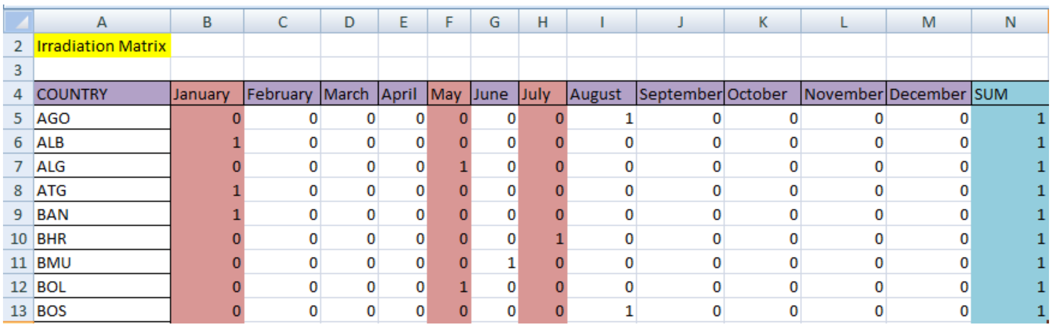
As a result we got 12 groups of sets - [[53, 7, 6, 4, 3, 2, 2, 1], [42, 10, 8, 6, 3, 3, 3, 2, 1], [40, 11, 9, 5, 4, 3, 3, 1, 1, 1], [34, 13, 9, 7, 5, 3, 3, 2, 1], [34, 12, 9, 7, 5, 3, 3, 2, 1, 1], [33, 14, 9, 7, 5, 3, 3, 2, 1], [27, 16, 10, 7, 5, 4, 3, 2, 2, 1], [27, 15, 10, 8, 5, 4, 3, 2, 2, 1], [27, 14, 10, 9, 5, 4, 3, 2, 2, 1], [22, 16, 11, 9, 6, 4, 3, 3, 2, 1], [21, 17, 11, 9, 6, 4, 3, 3, 2, 1], [20, 19, 11, 9, 6, 4, 3, 2, 2, 1]] with sums per group [78, 78, 78, 77, 77, 77, 77, 77, 77, 77, 77, 77] and number of countries per group [8, 9, 10, 9, 10, 9, 10, 10, 10, 10, 10, 10] accordingly.

Now we can choose any country with the corresponding number of sets and use predicted waiting time to assign each country to the particular irradiation window.

But this method does not allow us to get the full solution at one time, still manual work is needed.
The third option is to use linear programming.

7.3.3 Linear programming

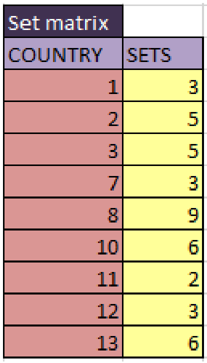
Four matrices were created.
First matrix - binary Irradiation matrix. Vertically we put countries, horizontally twelve months, values are filled with 1 (in this month hospitals from this country have to irradiate sets) or 0 (hospitals from this country do not have to irradiate sets in this month). Limitation - to each country we have to send sets just once during a year, the sum of values per row should be equal to 1. All values in the matrix are binary.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
2	Irradiation Matrix													
3														
4	COUNTRY	January	February	March	April	May	June	July	August	September	October	November	December	SUM
5	AGO	0	0	0	0	0	0	0	1	0	0	0	0	1
6	ALB	1	0	0	0	0	0	0	0	0	0	0	0	1
7	ALG	0	0	0	0	1	0	0	0	0	0	0	0	1
8	ATG	1	0	0	0	0	0	0	0	0	0	0	0	1
9	BAN	1	0	0	0	0	0	0	0	0	0	0	0	1
10	BHR	0	0	0	0	0	0	1	0	0	0	0	0	1
11	BMU	0	0	0	0	0	1	0	0	0	0	0	0	1
12	BOL	0	0	0	0	1	0	0	0	0	0	0	0	1
13	BOS	0	0	0	0	0	0	0	1	0	0	0	0	1

FIGURE 7.36: Irradiation matrix

The second matrix has the number of sets per country (anonymized picture because of IAEA security rules).

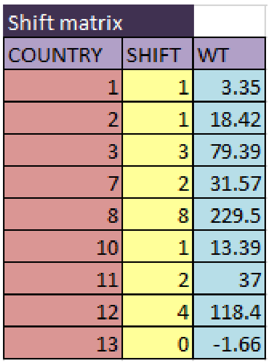


Set matrix	
COUNTRY	SETS
1	3
2	5
3	5
7	3
8	9
10	6
11	2
12	3
13	6

FIGURE 7.37: Set matrix

The third matrix has shift per country.
Waiting time - it is time between the end of an irradiation window and receive

date. Irradiation window starts normally on 15th each month and ends at the end of month. So, if country irradiates sets in February, for example, and waiting time is less then 30, we expect to receive sets in March. In such case, shift equals to one month. Using the same logic third matrix is filled with shift values using predicted waiting time per country.

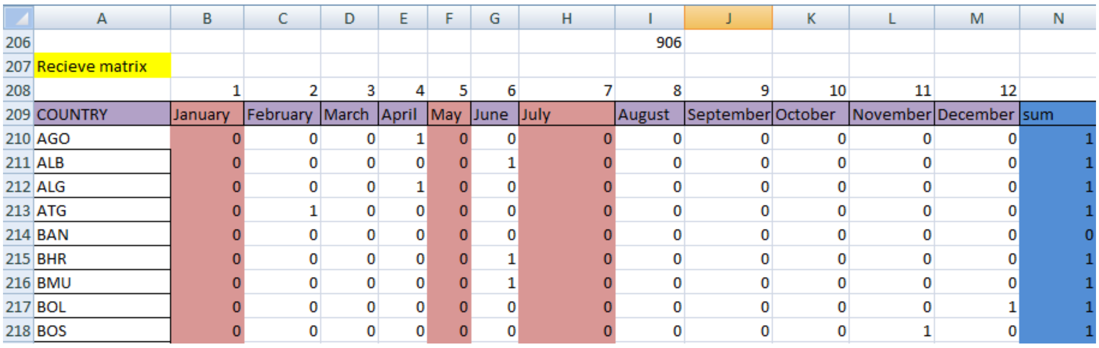


Shift matrix		
COUNTRY	SHIFT	WT
1	1	3.35
2	1	18.42
3	3	79.39
7	2	31.57
8	8	229.5
10	1	13.39
11	2	37
12	4	118.4
13	0	-1.66

FIGURE 7.38: Shift matrix

The picture is anonymized because of IAEA security rules.

Forth matrix is created by shifting values from the first matrix to next months depending on the value of shift from the third matrix. The last row equals to the sum of products of a column from forth matrix and value column from the second matrix. This last row has the number of sets that we send per month. Limitations for this matrix are the same as for the first matrix. Values should be binary, and sum per row should equal to 1.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
206									906					
207	Recieve matrix													
208		1	2	3	4	5	6	7	8	9	10	11	12	
209	COUNTRY	January	February	March	April	May	June	July	August	September	October	November	December	sum
210	AGO	0	0	0	1	0	0	0	0	0	0	0	0	1
211	ALB	0	0	0	0	0	1	0	0	0	0	0	0	1
212	ALG	0	0	0	1	0	0	0	0	0	0	0	0	1
213	ATG	0	1	0	0	0	0	0	0	0	0	0	0	1
214	BAN	0	0	0	0	0	0	0	0	0	0	0	0	0
215	BHR	0	0	0	0	0	1	0	0	0	0	0	0	1
216	BMU	0	0	0	0	0	1	0	0	0	0	0	0	1
217	BOL	0	0	0	0	0	0	0	0	0	0	0	1	1
218	BOS	0	0	0	0	0	0	0	0	0	0	1	0	1

FIGURE 7.39: Receive matrix

Target function equals to the sum of squared differences between perfect and actual sum per each month.

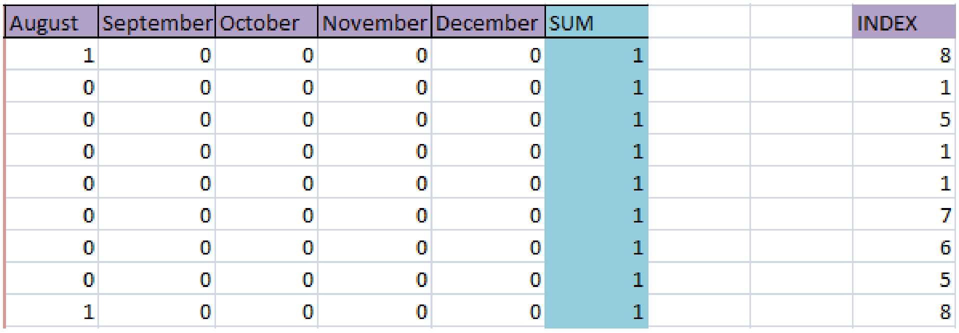
Perfect sum of sets per month equals the sum of all sets divided on the number of months.

To receive almost the same number of sets each month, this target function should be minimized.

This model allows to include more constraints, such as national holidays and non-operational/fixed months.

Shift matrix can dynamically change values depending on the month. So, if we plan to irradiate sets in particular country in May while there are 4 national holidays, we add this number of days to waiting time and modify shift value. Additional table

with national holidays per each country was created.
We have two months January and July without planned irradiation runs and one fixed irradiation run on May just for SSDL. But we can receive and check sets during these months. Additional limitations can be specified by limiting the sum of column for these months in the first matrix to be 0.
Excel solver is run but the task is too large. To reduce the number of changing cells new additional column for the first matrix was created. Values in this column indicate an index of month where 1 is supposed to be placed. So, if 1 (sets will be irradiated in May for this country) is put to May and index of January is 1, then new column value for this country will be 5. In such a way, we got 115 changing cells instead of 115*12 (we have 115 countries and 12 months). New limitations - values in this column should be integers less than 13 and larger then 0.

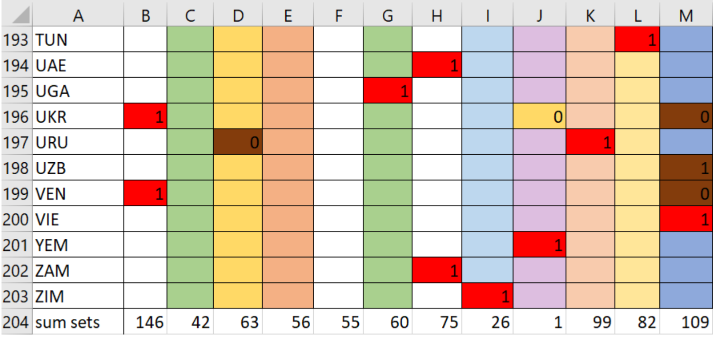


August	September	October	November	December	SUM			INDEX
1	0	0	0	0	1			8
0	0	0	0	0	1			1
0	0	0	0	0	1			5
0	0	0	0	0	1			1
0	0	0	0	0	1			1
0	0	0	0	0	1			7
0	0	0	0	0	1			6
0	0	0	0	0	1			5
1	0	0	0	0	1			8

FIGURE 7.40: Index matrix

The described model gave us a full solution with balanced number of sets that we expect to receive each month and already calculated groups of countries that should be assigned to particular irradiation windows. Such model allowed us to take in a count all constraints and is flexible for potential modifications in future.
The perfect schedule was created, but our last constrains are - distribution method and already established schedule.
Countries from Latin America get their sets from PAHO, the coordinator with whom DOL has an agreement. Such countries can be assigned just to three PAHO irradiation windows.
DOL has an agreement with each country for more then 20 years, so we can make changes and agree with country on new month for audit but only if it really makes a big change for DOL work intensity.

To see what changes can influence the most, similar to described above model was created but with values from actual schedule in 2017 (number of countries - 97). Number of sets that is expected to receive each month using predicted waiting time and sets can be found in the table below.

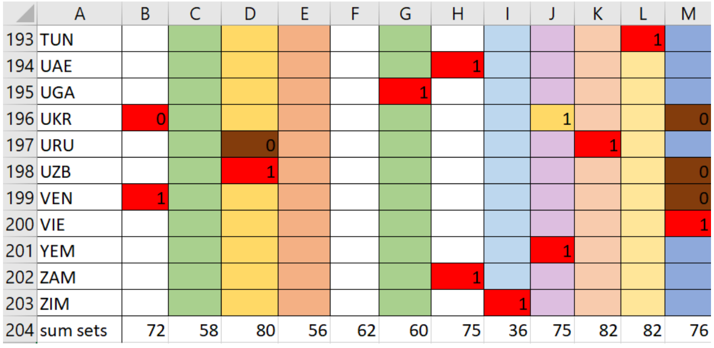


	A	B	C	D	E	F	G	H	I	J	K	L	M
193	TUN											1	
194	UAE							1					
195	UGA						1						
196	UKR	1								0			0
197	URU			0							1		
198	UZB												1
199	VEN	1											0
200	VIE												1
201	YEM									1			
202	ZAM							1					
203	ZIM								1				
204	sum sets	146	42	63	56	55	60	75	26	1	99	82	109

FIGURE 7.41: Balanced receive 2017

Since values in cells are changing dynamically, thanks to excel formulas, we can make changes in the first (irradiation) matrix and see results in the forth (receive) matrix.

While manually changing schedule, the main idea is to not have countries with a lot of sets together in one receive month. For example, in January we expect to receive 146 sets (when perfect is 67) because of Russia (53), Ukraine (40), Venezuela (19) and Chili (34). DOL sends sets to Russia and Ukraine for irradiation in November (but December is last month in the year with making all year reports in a hurry and long holidays). So most of sets are returning back in January.
Since 53+19=72(RUS+VEN), 40+34=74(UKR+CHI), 40+19=59(UKR+VEN) and (72-67=5) < (74-67=7) < (67-59=8), better choice will be to move Ukraine and Chili to another month. August, for example, has 26 sets, that makes combination 40+26=66 very close to perfect. The shift for Ukraine equals 2 months, so we need to send sets to Ukraine in May to receive them in August. But May is fixed for SSDL. But we can send sets also in June because September is extremely empty now with just 1 set expected.
It would be preferred to receive sets from Chili in August, but Chili has a shift of 1 and July is not operational. So we move Chili from December irradiation window to August to receive sets in September. Now we expect 75 sets in September.
By trying different combinations of changes in such a way, a new more balanced schedule was created by changing Chili, Ukraine, Ecuador, Panama and Egypt assignments to irradiation windows.



	A	B	C	D	E	F	G	H	I	J	K	L	M
193	TUN											1	
194	UAE							1					
195	UGA						1						
196	UKR	0								1			0
197	URU			0							1		
198	UZB			1									0
199	VEN	1											0
200	VIE												1
201	YEM									1			
202	ZAM							1					
203	ZIM								1				
204	sum sets	72	58	80	56	62	60	75	36	75	82	82	76

FIGURE 7.42: Balanced receive 2017

Chapter 8

Summary

To conclude, the perfect schedule for distribution of dosimeter sets by IAEA/WHO postal dose quality audit for 115 countries and 12 months was created. It takes in a count such constraints as national holidays and non-operational/fixed months.

The process was divided into three main stages: predicting waiting time (time between the end of an irradiation window and receive date), predicting the number of sets and scheduling (assignment of countries to particular irradiation windows).

Linear Regression, Light GBM, LSTM, and ARIMA were used to find the best algorithm for each country to predict waiting time. As a result for 76 countries LSTM prediction was used, for 32 - ARIMA and for 7 - last historical value (for countries with to less historical data to analyze).

Exponential Smoothing was used to predict the number of sets per country.

For third stage (scheduling) LRM, Bin packing and Linear Programming was used. The complete solution, that allowed to take in a count constraints, was produced my liner programming approach.

Since new perfect schedule requires plenty of changes to be made by DOL, that has agreements with each country more than 20 years and can not change everything at one time; the model was used to analyze what changes can influence the most on balance of DOL work intensity. Found changes can be applied next year for distribution of dosimeter sets by IAEA/WHO postal dose quality audit.

Bibliography

Brown, Robert Goodell (1963). *Smoothing Forecasting and Prediction of Discrete Time Series*.

Friedman, J. H. (1999). "Stochastic Gradient Boosting". In:

Grover, Prince (2017). *Gradient Boosting from scratch*. URL: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory neural computation". In:

Hyndman, R.J., G. Athanasopoulos R.A. Ahmed, and H.L. Shang (2011). *Optimal combination forecasts for hierarchical time series*. Chap. Computational Statistics and Data Analysis, 2580—2589.

Hyndman, Rob J (2018). *Forecasting: Principles and practice*.

Hyndman, Rob J and George Athanasopoulos (2018). *Forecasting: Principles and Practice*. Chap. ARIMA models.

Ke, Guolin et al. (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In:

Mandot, Pushkar (2017). *What is LightGBM, How to implement it? How to fine tune the parameters?* URL: https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc.

Olah, Christopher (2015). *Understanding LSTM Networks*. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Othman, Sameera A., Sizar A. Mohammed, and Shelan S. Ismaeel (2013). "ON forecasting by Dynamic Regression models". In: *J. of university of anbar for pure science*.

Schrijver, Alexander (1998). *Theory of Linear and Integer Programming*.

Winston, Wayne Leslie and S. Christian Albright (1994). *Practical Management Science*. Chap. CREW SCHEDULING AT BRANEAST AIRLINES, pp. 260–265.

Yan, Xin (2009). *Linear Regression Analysis: Theory and Computing*.

Zhang, Jilian, Kyriakos Mouratidis, and Hwee Hwa Pang (2011). "Heuristic Algorithms for Balanced Multi-way Number Partitioning". In: