TIØ4285: Deliverable 1

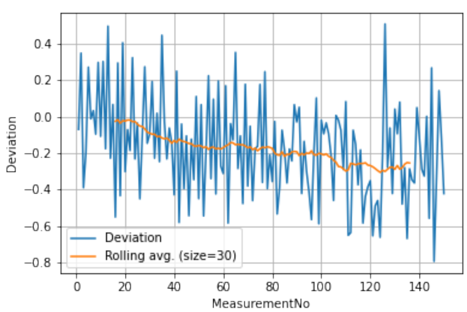
Group 73
Patrik Kjærran & Morgan Heggland & Martin Skogset

February 25, 2021

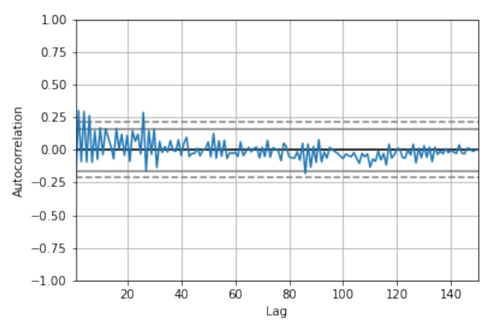
Task 1 Time series forecasting

1.a Suitable forecasting methods

In order to determine which forecasting methods are suitable, we need to examine the underlying properties of the time series data. A first natural step is to determine whether the time series is *stationary*, a term used to describe a series or process which statistical properties does not depend on timestep. This is important, as a number of forecasting methods assume a stationary process. We therefore perform and interpret the *Augmented Dickey-Fuller (ADF)* and *KPSS* tests from the `statsmodels`[1] library, a Python module with implementations for a series of statistical utilities.



(a) Original time series with a rolling average window of size 30.



(b) Autocorrelation plot.

Figure 1: Dataset with autocorrelation plot.

Using the significance of p=0.01 we can conclude that the series is *difference-stationary* as the Dickey-Fuller null hypothesis is rejected and there is no evidence that the KPSS null hypothesis can be rejected. This means that we can make the series stationary by *differencing*[2]. This simply means subtracting the series with a lagged version of

[1] https://www.statsmodels.org
[2] https://otexts.com/fpp2/stationarity.html#differencing

```
Results of Dickey-Fuller Test:
Test statistic               -4.185143
p-value                       0.000698
Lags used                     3.000000
   Critical Value (1%)       -3.475953
   Critical Value (5%)       -2.881548
   Critical Value (10%)      -2.577439
```

(a) Dickey-Fuller test results.

```
Results of KPSS Test:
Test statistic                1.526122
p-value                       0.010000
Lags used                     2.000000
   Critical Value (10%)       0.347000
   Critical Value (5%)        0.463000
   Critical Value (1%)        0.739000
```

(b) KPSS test results.

Figure 2: Tests to determine whether the time series is stationary.

itself, which stabilizes the mean and eliminates any linear trends (we can see a slight downward slope in the original data). As most forecasting methods assume a stationary process, we made efforts to stationarize the process in order to have a larger base of feasible forecasting methods to choose from. For this we will use a first-order differencing, which will have to be inverted when making forecasts.

Post-differencing, the ADF and KPSS tests return p-values of 1.17e-9 and 0.01, respectively. We now have a stationary time series. For further details on how the test results are interpreted, see the statsmodels documentation[3].

From the autocorrelation plot in Figure 1b, we can see that there is no significant autocorrelation. That is, the series is not significantly correlated with a lagged version of itself. This indicates that there is no significant *seasonality* in the time series, given the small sample size that we have observed. From the shape of the autocorrelation plot, we can see that a value is significantly more correlated to recent observations, indicating that exponential smoothing could be suitable.

Based on the above insight, a number of forecasting methods may be suitable. First of all, simple models like last-value forecasting and averaging forecasting could be considered. However, these are rather naive, where the former method uses only the last observation and the latter simply averages the dataset. As we have observed a slight correlation between recent observations, a more suitable choice would be to use a form of moving average method. Such methods utilize only a limited number of recent observations instead of averaging the entire dataset, thus being more appropriate in this case. Standard moving average forecasting could be a viable alternative, however this weighs all observations included equally. A more suitable method could be to use expontential smoothing which allows us to weigh the most recent observations higher, in line with the observed correlations. Methods for data with trends or seasonal data could be also considered, however the simplified time series produced and the small sample size implies that use of such methods would be less suitable nor necessary.

[3]https://www.statsmodels.org/stable/examples/notebooks/generated/stationarity_detrending_adf_kpss.html

1.b Creating a forecast

Based on the discussion in 1.a, we deem exponential smoothing to be the most suitable forecasting method for the time series at hand. Using this method, prediction results using exponential smoothing with different values of the smoothing constant α were produced. The results can be seen in Figure 3.

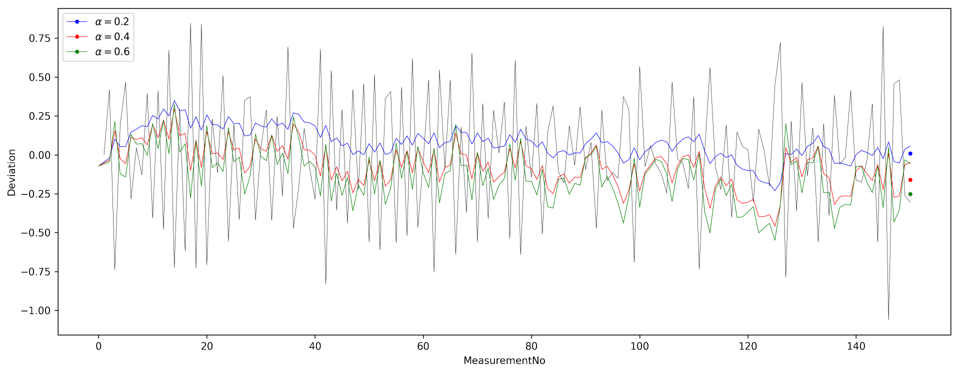


Figure 3: One-step forecasts using exponential moving average with smoothing constant $\alpha \in (0.2, 0.4, 0.6)$. Future forecasts are indicated with a dot. The colored lines represent forecasts using historic values up to the given point. The dark line represents the original data.

Choice of the smoothing factor α has considerable impact on the forecast as illustrated in Figure 3, thus impacting the quality as well. A higher alpha would imply a more volatile forecast, but one which quickly captures a change in the output of the process. On the contrary, a lower alpha would imply more lag but less volatility. From visual inspection, we consider $\alpha = 0.4$ to nicely balance these two effects. However, in order to quantitatively assess the quality of the forecast, one could evaluate the forecast using known values using a measure of fit, for example *Mean Squared Error (MSE)*.

We can see that all of the predictions are placed relatively close to the mean. In comparison, the process itself has large spikes in both positive and negative direction, indicating high variance. Hence, it seems like the forecasting method is unable to consider the real variance of the process. This could also be derived from the forecasting method, which only considers the averages of the historical timesteps. On the contrary, one would expect the amplitude of the forecasts to be lower than the underlying data. This is partially because the variance of historic steps partially cancel when averaging, producing a less volatile forecast. One could argue that the reduced variance is an advantage of the method as it produces less extreme forecasts, which in expectation should be more accurate. This is an inherent property of the exponential smoothing procedure (for a continous trend, an exponential smoothing forecast would always lag behind the actual trend).

1.c Predicting the next 3 measurements

A possible way to extend to multiple forecasts for exponential smoothing could be to use the forecasting model to predict the next three measurements by chaining the predictions. That is, after producing the forecast for the first future measurement, this forecast can be considered as the newest measurement and can thus be used to predict the next measurement. Repeating once more in this manner, we can produce a third measurement using the two first predictions.

This would cause basing predictions on other uncertain predictions in addition to the original data, but would cause the outputs to have an increasing degree of uncertainty.

A different, common way[4] to extend to multiple forecasts is to extrapolate using the last predicted value in the following manner:

$$\hat{y}_{T+h|T} = \hat{y}_{T+1|T}, \quad h = 2, 3, ...$$

where T denotes the timestep for the last known value and $\hat{y}_{t|T}$ the forecast for step t given values up to T. This is known as *flat forecasting*. Applied to the differenced timeseries, the resulting forecast would be a line with a constant slope, which for this particular series could provide reasonable forecasts for a short time horizon.

[4] https://otexts.com/fpp2/ses.htmlflat-forecasts

Task 2 Distribution Planning Problem

2.a Type of planning problem

The distribution planning problem described represents a transshipment problem, a type of transportation problem where the transport may go through intermediate nodes, such as the distribution centers in this case. Given a set of demand for the customers, the problem revolves around deciding upon how to distribute the transport such that customer demand is satisfied while minimizing the transportation costs, and accounting for any capacity and throughput constraints. As demand is known before planning is done, the problem can be solved deterministically as a single stage problem for each planning period.

2.b Problem formulation

In order to formulate the problem as an optimization problem, we start the problem formulation by defining all relevant sets, indices, parameters and variables. Then, the actual problem consists of the objective function seeking to minimize transportation costs, and a set of constraints defining and restricting the flow of transport between factories, distribution centers and customers.

Sets

$F = \{1, ..., \lvert F\rvert\}$	Set of factories, indexed f
$D = \{1, ..., \lvert D\rvert\}$	Set of distribution centers, indexed d
$C = \{1, ..., \lvert C\rvert\}$	Set of customers, indexed c
$E^{FD} \subseteq F \times D$	Possible transportation between factories and distribution centers
$E^{DC} \subseteq D \times C$	Possible transportation between distribution centers and customers
$E^{FC} \subseteq F \times C$	Possible transportation between factories and customers

Parameters

C^{FD}_{fd}	Transportation cost from factory f to distribution center d
C^{FC}_{fc}	Transportation cost from factory f to customer c
C^{DC}_{dc}	Transportation cost from distribution center d to customer c
PC_f	Production capacity for factory f
T_d	Maximum throughput in distribution center d
D_c	Demand of customer c

Variables

$$
\begin{aligned}
x_{fd}, &\quad \text{Amount to transport from factory } f \text{ to distribution center } d \\
y_{fc}, &\quad \text{Amount to transport from factory } f \text{ to customer } c \\
z_{dc}, &\quad \text{Amount to transport from distribution center } d \text{ to customer } c
\end{aligned}
$$

Minimize:

$$
\sum_{(f,d)\in E^{FD}} C^{FD}_{fd} x_{fd} + \sum_{(f,c)\in E^{FC}} C^{FC}_{fc} y_{fc} + \sum_{(d,c)\in E^{DC}} C^{DC}_{dc} z_{dc}
$$

Subject to:

$$
\sum_{d:(f,d)\in E^{FD}} x_{fd} + \sum_{c:(f,c)\in E^{FC}} y_{fc} \leq PC_f \qquad f \in F \tag{1}
$$

$$
\sum_{f:(f,d)\in E^{FD}} x_{fd} = \sum_{c:(d,c)\in E^{DC}} z_{dc}, \qquad d \in D \tag{2}
$$

$$
\sum_{f:(f,d)\in E^{FD}} x_{fd} \leq T_d, \qquad d \in D \tag{3}
$$

$$
\sum_{d:(d,c)\in E^{DC}} z_{dc} + \sum_{f:(f,c)\in E^{FC}} y_{fc} = D_c, \qquad c \in C \tag{4}
$$

$$
x_{fd} \geq 0 \quad (f,d) \in E^{FD} \tag{5}
$$

$$
y_{fc} \geq 0 \quad (f,c) \in E^{FC} \tag{6}
$$

$$
z_{dc} \geq 0 \quad (d,c) \in E^{DC} \tag{7}
$$

Constraints (1) restrict flow of transport from a factory to be less than or equal to the capacity of the factory, (2) specify conservation of flow for the distribution centers, (3) limit the flow of transport through the distribution centers to be less than or equal to the throughput capacities, (4) ensure that customer demand is satisfied and (5)-(7) ensure non-negativity of the transported amounts.

After implementing the model in Mosel and solving the problem instance supplied in the excel-sheet, the solution seen in Table 1 was found. The model implementation can be found in `excercise_2b.mos`. The optimal objective value was **198 500**.

To	From Liverpool	Brighton	Newcastle	Birmingham	London	Exeter
Newcastle	0	-				
Birmingham	0	50000				
London	0	55000				
Exeter	40000	0				
C1	50000	0	-	0	-	-
C2	-	-	0	10000	0	-
C3	0	-	0	0	0	40000
C4	0	-	0	35000	-	0
C5	-	-	-	5000	55000	0
C6	2000	-	0	-	0	0

Table 1: Optimal solution values for each of the decision variables in the problem instance provided.

2.c Change of Problem Type Assessment

In the new situation, demand is no longer known during the initial transportation planning phase. Thus, one has to plan under uncertainty, and when demand information becomes available, one has to make additional decisions based on this information. Consequently, this is now a two-stage stochastic recourse problem. The following explanation details our interpretation of the new situation.

We interpret the problem text such that any decision to distribute from a factory at the regular costs must be done before demand is known, so that only express transport is available from the factory after demand is known. During the first stage, one can then distribute products from factories to distribution centers or directly to customers at the regular transportation costs. Thus, we consider these decisions as first-stage decisions.

As deciding all transportation of products to customers before knowing demand risks creating an infeasible solution, all transport to customers except for regular direct transportation from factories are considered second-stage decisions. Thus, transportation from distribution centers is decided once demand is known in order to ensure cost-efficient feasible solutions. Once all distribution between distribution centers and customers is decided upon, any unsatisfied demand must be covered using express services from factories to customers. Therefore, this represents another second-stage decision.

2.d Model Changes due to Uncertainty

The introduction of uncertainty implies a number of changes to the model. First of all, we have to model the different possible outcomes for demand, which we will represent as scenarios. Each scenario is associated with a probability of that scenario occuring. This is realized in the model by introducing a set of scenarios, and parameters specifying the associated probabilities. Additionally, we extend the demand parameters to depend on both the customer and the scenario.

As described in 2.c, we now split decisions into whether they are taken during the first and second stage. Additionally, we introduce a new type of variable deciding the amount of express transport to utilize. We also discern the decision variables based on which scenario occurs. Through this formulation, solving the stochastic model allows the planners to determine a priori the optimal decisions to make in any given stage for any given scenario that may occur, taking uncertainty into account.

Our set of variables now include two types of first-stage variables x_{fds} and y_{fcs}, representing the amounts to transport from factories to distribution centers and from factories to customers respectively. Additionally, we introduce two types of second-stage variables z_{dcs} and w_{fcs}, representing the amounts to transport from distribution centers to customers and express transport from factories to customers respectively. This implies that before demands are known, the planners would implement decisions upon regular transportation from factories to distribution centers and to customers, and then once demand is known, they would implement decisions regarding transportation from distribution centers to customers and express transportation from factories to customers.

As we now have a new form of transportation, we introduce new parameters representing the express transportation costs. The objective function is updated to minimize the expected transportation costs, including the express transportation. That is, we minimize the transportation costs we decide for each scenario and weight that cost by the likelihood of the scenario occuring. The constraints are updated to include the amounts of express transport used. The distribution center does not necessarily need to be emptied in stage 2, and this is reflected in the change of the flow conservation constraint to a "$\geq$"-constraint. New non-anticipativity constraints are also introduced, to ensure that the same decisions are made when the same information is available. The constraints are described in further detail in 2.e.

2.e Stochastic Model Formulation

Sets

$F = \{1, ..., \lvert F \rvert\}$	set of factories, indexed f
$D = \{1, ..., \lvert D \rvert\}$	set of distribution centers, indexed d
$C = \{1, ..., \lvert C \rvert\}$	set of customers, indexed c
$E^{FD} \subseteq F \times D$	possible transportation between factories and distribution centers
$E^{DC} \subseteq D \times C$	possible transportation between distribution centers and customers
$E^{FC} \subseteq F \times C$	possible transportation between factories and customers
$S = \{1, ..., \lvert S \rvert\}$	set of scenarios, indexed s

Parameters

C^{FD}_{fd}	Transportation cost from factory f to distribution center d
C^{FC}_{fc}	Regualt transportation cost from factory f to customer c
C^{DC}_{dc}	Transportation cost from distribution center d to customer c
C^{X}_{fc}	Express transportation cost from factory f to customer c
PC_f	Production capacity for factory f
T_d	Maximum throughput / storage capacity in distribution center d
D_{cs}	Demand of customer c in scenario s
p_s	Probability for scenario s

Variables

– First-stage variables

x_{fds}, Amount to transport from factory f to distribution center d in scenario s.

y_{fcs}, Amount to transport from factory f to customer c in scenario s.

– Second-stage variables

z_{dcs}, Amount to transport from distribution center d to customer c in scenario s.

w_{fcs}, Amount to transport with express from factory f to customer c in scenario s.

Minimize:

$$\sum_{s \in S} p_s [\sum_{(f,d) \in E^{FD}} C^{FD}_{fd} x_{fds} + \sum_{(d,c) \in E^{DC}} C^{DC}_{dc} z_{dcs} + \sum_{(f,c) \in E^{FC}} (C^{FC}_{fc} y_{fcs} + C^{X}_{fc} w_{fcs})]$$

Subject to:

$$\sum_{d:(f,d) \in E^{FD}} x_{fds} + \sum_{c:(f,c) \in E^{FC}} (y_{fcs} + w_{fcs}) \leq PC_f \qquad f \in F, s \in S \tag{8}$$

$$\sum_{f:(f,d) \in E^{FD}} x_{fds} \geq \sum_{c:(d,c) \in E^{DC}} z_{dcs}, \qquad d \in D, s \in S \tag{9}$$

$$\sum_{f:(f,d) \in E^{FD}} x_{fds} \leq T_d, \qquad d \in D, s \in S \tag{10}$$

$$\sum_{d:(d,c) \in E^{DC}} z_{dcs} + \sum_{f:(f,d) \in E^{FD}} (y_{fcs} + w_{fcs}) = D_{cs}, \qquad c \in C, s \in S \tag{11}$$

$$x_{fd} = x_{fds}, \quad f \in F, d \in D, s \in S \tag{12}$$

$$y_{fc} = y_{fcs}, \quad f \in F, c \in C, s \in S \tag{13}$$

$$x_{fds} \geq 0 \quad (f,d) \in E^{FD}, s \in S \tag{14}$$

$$y_{fcs}, w_{fcs} \geq 0 \quad (f,c) \in E^{FC}, s \in S \tag{15}$$

$$z_{dcs} \geq 0 \quad (d,c) \in E^{DC}, s \in S \tag{16}$$

Constraints (8)-(11) resemble constraints (1)-(4) from subtask 2.b. Constraints (8) and (11) are however updated to account for the fact that w_{fcs} also represent transport from factories to customers. Constraint (9) is changed from equality to "$\geq$", enforcing that the distribution centers does not necessarily need to be emptied in stage 2. This leads to increased flexibility for the planners in stage 1, since they can use the available storage capacity in the distribution centers to reduce the probabilities for need of express delivery in stage 2. Constraints (12) and (13) are non-anticipativity constraints, ensuring that decicions made in the first stage are equal, across all scenarios. This is based on the requirements of stochastic recourse problems, where all decicions made with the same information available must be equal. Lastly, all variables are indexed based on the scenario they represent.

2.f Solving the stochastic problem instances

The formulated problem was implemented in Mosel and solved using the two different set of scenarios for demand data. The model implementation can be found in `excercise_2f.mos`. The resulting objective values were **202 793** for the uncorrelated demand and **210 987.12** for the correlated demand. Further results for the different decision variables can be found in Table 2 and Table 3.

As we can see, the optimal costs for the uncorrelated demands are slightly lower. This may be partly explained by the fact that when demand is correlated, demand of all customers increase or decrease in tandem, but one cannot predict which outcome will be the case. This implies that the likelihood of either transporting too much or too little to the different distribution centers is increased compared to during uncorrelated demand. Thus, costs increase as one either has to increase the inventory buffer or pay extra for express transport.

In addition to this, the difference in objective value can be partly explained by the fact that the sum of total demand across the different scenarios is slightly lower in the uncorrelated case. This implies that overall, a slightly lower amount of goods must be transported and paid for, reducing the overall cost.

	To	From Liverpool	Brighton	Newcastle	Birmingham	London	Exeter
	Newcastle	11200	-				
	Birmingham	0	50000				
	London	0	55800				
	Exeter	40000	0				
Scenario 1	**C1**	38000/6600	0/0	-	5400	-	-
	C2	-	-	0	10000	0	-
	C3	0/0	-	4200	0	0	35800
	C4	0/0	-	400	34600	-	0
	C5	-	-	-	0	55800	4200
	C6	13400/0	-	6600	-	0	0
Scenario 2	**C1**	38000/0	0/0	-	0	-	-
	C2	-	-	0	10000	0	-
	C3	0/0	-	6350	0	0	27650
	C4	0/0	-	0	28350	-	0
	C5	-	-	-	11650	55800	12350
	C6	13400/0	-	2400	-	0	0
Scenario 3	**C1**	38000/7150	0/0	-	5350	-	-
	C2	-	-	0	5700	0	-
	C3	0/0	-	11200	0	0	40000
	C4	0/0	-	0	18550	-	0
	C5	-	-	-	20400	55800	0
	C6	13400/0	-	0	-	0	0
Scenario 4	**C1**	38000/100	0/0	-	3400	-	-
	C2	-	-	0	11000	0	-
	C3	0/0	-	0	0	0	38400
	C4	0/0	-	2000	35600	-	1600
	C5	-	-	-	0	55800	0
	C6	13400/0	-	9200	-	0	0
Scenario 5	**C1**	38000/1650	0/0	-	13850	-	-
	C2	-	-	0	10200	0	-
	C3	0/0	-	0	0	0	37600
	C4	0/0	-	0	25950	-	2400
	C5	-	-	-	0	49800	0
	C6	13400/0	-	11200	-	1600	0

Table 2: Optimal solution values for each of the decision variables for the uncorrelated demand data. For the distribution from factories to customers, values for y_{fcs}- and w_{fcs}-variables are represented in the same cell for each scenario, displayed as y/w.

	To	From Liverpool	Brighton	Newcastle	Birmingham	London	Exeter
	Newcastle	6900	-				
	Birmingham	0	50000				
	London	0	63400				
	Exeter	40000	0				
Scenario 1	**C1**	41500/0	0/0	-	8500	-	-
	C2	-	-	0	10000	0	-
	C3	0/0	-	0	0	0	40000
	C4	0/0	-	3500	31500	-	0
	C5	-	-	-	0	60000	0
	C6	16600/0	-	3400	-	0	0
Scenario 2	**C1**	41500/13500	0/0	-	0	-	-
	C2	-	-	0	11000	0	-
	C3	0/0	-	6100	0	0	37900
	C4	0/0	-	0	38500	-	0
	C5	-	-	-	5000	63400	2100
	C6	16600/4600	-	800	-	0	0
Scenario 3	**C1**	41500/0	0/0	-	0	-	-
	C2	-	-	0	8300	0	-
	C3	0/0	-	0	0	0	33200
	C4	0/0	-	0	29050	-	0
	C5	-	-	-	0	43000	6800
	C6	16600/0	-	0	-	0	0
Scenario 4	**C1**	41500/0	0/0	-	3860	-	-
	C2	-	-	0	9072	0	-
	C3	0/0	-	0	0	0	36288
	C4	0/0	-	0	31752	-	0
	C5	-	-	-	0	50720	3712
	C6	16600/0	-	1544	-	0	0
Scenario 5	**C1**	41500/12500	0/0	-	0	-	-
	C2	-	-	0	10800	0	-
	C3	0/0	-	3200	0	0	40000
	C4	0/0	-	0	37800	-	0
	C5	-	-	-	1400	63400	0
	C6	16600/1300	-	3700	-	0	0

Table 3: Optimal solution values for each of the decision variables for the correlated demand data. For the distribution from factories to customers, values for y_{fcs}- and w_{fcs}-variables are represented in the same cell for each scenario, displayed as y/w.