02435 - Decision-making under Uncertainty

Assignment 2

Reda LAHLOU s192431

02435 - Decision-making under Uncertainty
April 7, 2020

1 Task 1 - Linear Time Series Analysis

The aim of this task is to fit a time series model to historical data of real estate prices in the investment area *ZIP2000*.

The model used will an **ARIMA** model (autoregressive integrated moving average). For a time series $(y_t)_{t\geq 0}$, an ARIMA model is formulated as follows:

$$(1 - \sum_{j=1}^{p} \Phi_j B^j)(1 - B)^d y_t = (1 - \sum_{j=1}^{q} \theta_j B^j) e_t \tag{1}$$

,where:

- p, q, d are respectively the parameter of the autoregressive process AR(p), the parameter of the moving average process MA(q) and the order to which the data are differentiated
- $(\Phi_j)_{j\geq 0}$ are the coefficients in the AR process
- $(\theta_j)_{j\geq 0}$ are the coefficients in the MA process
- $(B^j)_{j\geq 0}$ is the series of the lag operators defined as $B^j = \frac{y_{t-j}}{y_t}$
- $(e_t)_{t\geq 0}$ is the series of the error terms

The first step to fit an ARIMA is to make sure that the data are stationary, i.e. that its properties (mean and variance) do not change over time.

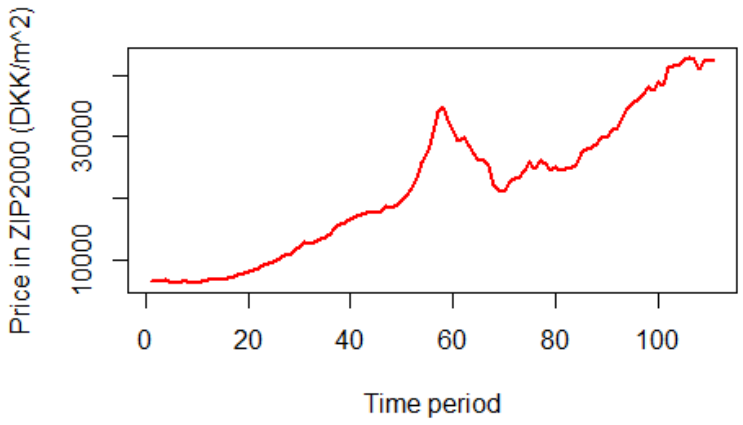

Figure 1: Historical investment prices in ZIP2000 over time

Figure 1 shows an increase in the mean of the data over time, as well as a non-constant variance. In order to stabilize the variance, the data are first transformed using a logarithmic transformation.

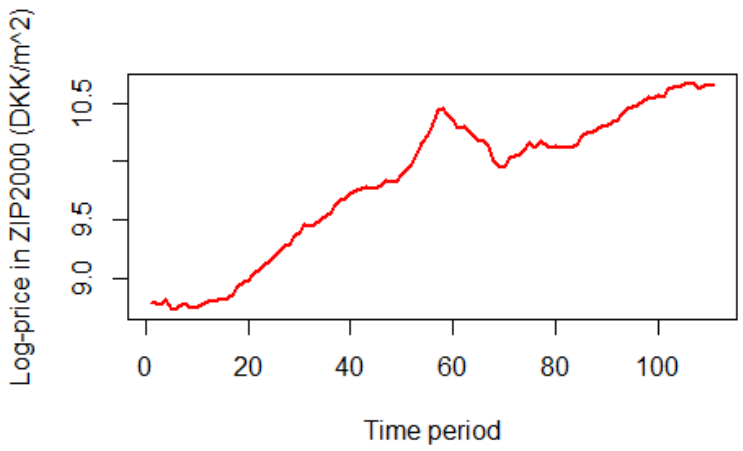

Figure 2: Historical log-prices in ZIP2000 over time

Some improvement can be seen on figure 2, particularly regarding the price spike around time unit 60, which seems flatter than on figure 1.

The next step is to remove the increasing trend in the mean. This can be done by differentiating the log-data:

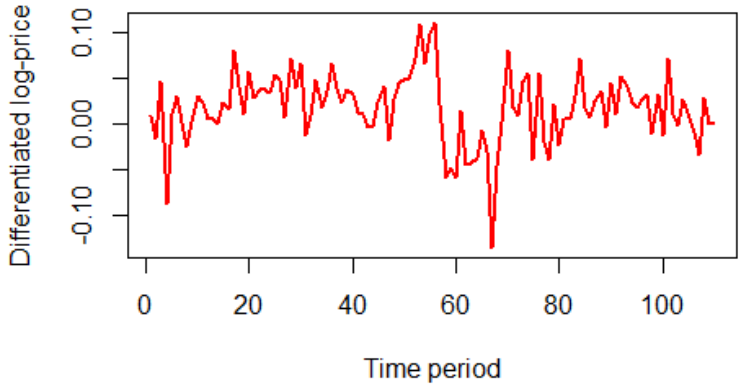

Figure 3: Historical differentiated log-prices in ZIP2000 over time

Although the increasing trend in the mean appears to have vanished on figure 3, there are still some time frames when the properties of the log-data do not seem stationary. Another differentiation is performed on the log-data:

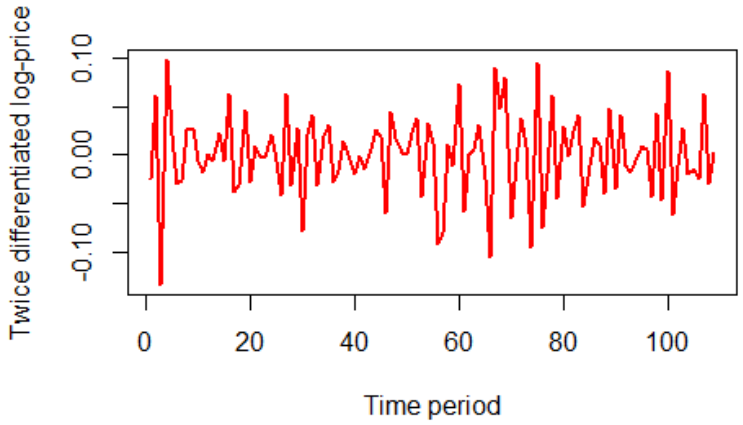

Figure 4: Historical twice differentiated log-prices in ZIP2000 over time

Figure 4 seems more stationary than 3. To validate this impression, the auto-correlation function (ACF) of both the transformed data are plotted on figure 5.

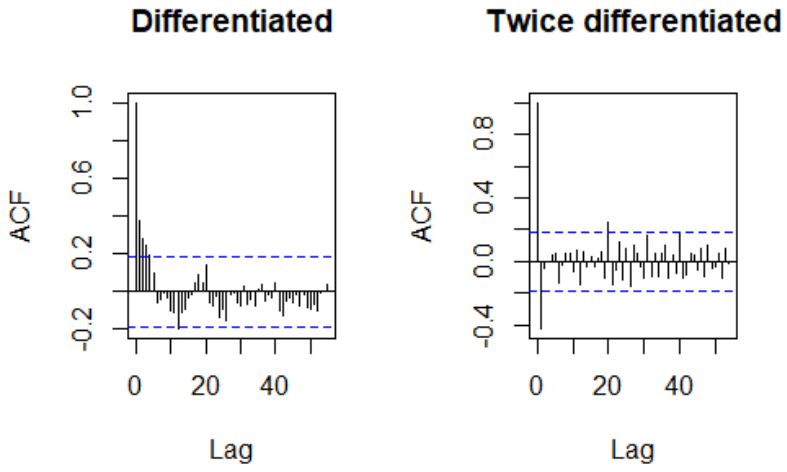

Figure 5: Comparison of ACF between twice differentiated log-prices and differentiated log-prices in ZIP2000 over time

For stationary data, the ACF is supposed to cut off near zero after a few lags. When differentiating twice instead of once, the ACF cuts off near zero after less lags, and it seems

harder to guess any clear patterns in the values taken by the ACF.

To sum up, it appears that differentiating twice the logarithm of the data leads to seemingly stationary data. This suggests a value d = 2 for the ARIMA model.

The next step is to find the correct parameters p and q to build the ARIMA model. Their values are related respectively to the PACF (partial auto-correlation function) and the ACF of the residuals. If these parameters are correctly stated, the ACF and PACF of the residuals should be close to 0 for every lag. Looking first at the ACF and PACF of the transformed data can give an idea of the values of p and q to choose:

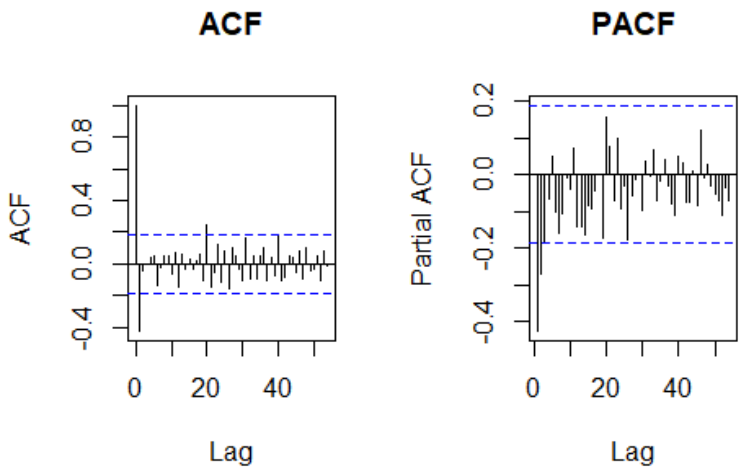

Figure 6: ACF and PACF of twice differentiated log-prices in ZIP2000 over time

The number of significant lags in the ACF and PACF of the transformed data gives an idea of the values to be taken by q and p. In order to avoid over-fitting the model, incremental changes are made until no improvement in the ACF and PACF of the residuals occurs. Figure 6 shows that there is at least one significant lag in both the ACF and the PACF of the transformed data. The initial guess for guess is thus to fit an ARIMA model on the log-data with parameters (p,d,q) = (1,2,1). The ACF and PACF of the residuals of this model are shown on figure 7:

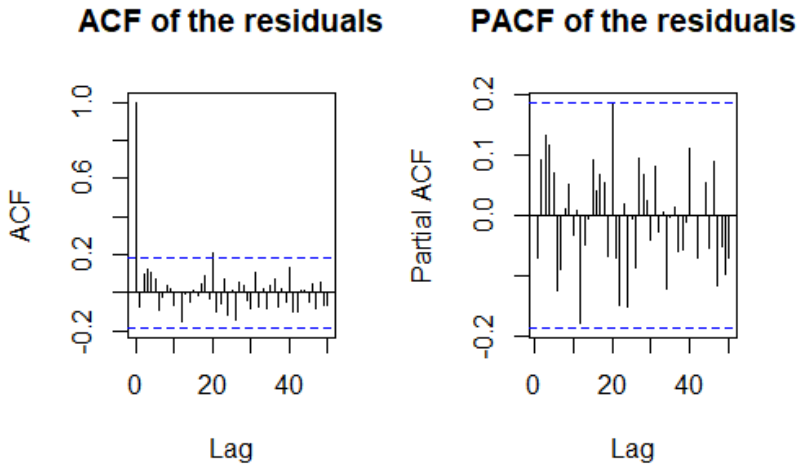

Figure 7: ACF and PACF of the residuals of ARIMA(1,2,1) model fitted log-prices in ZIP2000 over time

The plots on figure 7 look acceptable, as the values for the ACF and PACF are almost equal to 0 for every lag. There are however a few spikes, particularly around time unit 20. In order to check if the values taken for the ARIMA parameters are high enough, incremental changes are made to build the 3 following models: ARIMA(2,2,1), ARIMA(1,3,1) and ARIMA(1,2,2). The residuals of of the fits of these models on the log-prices can be seen on figure 8:

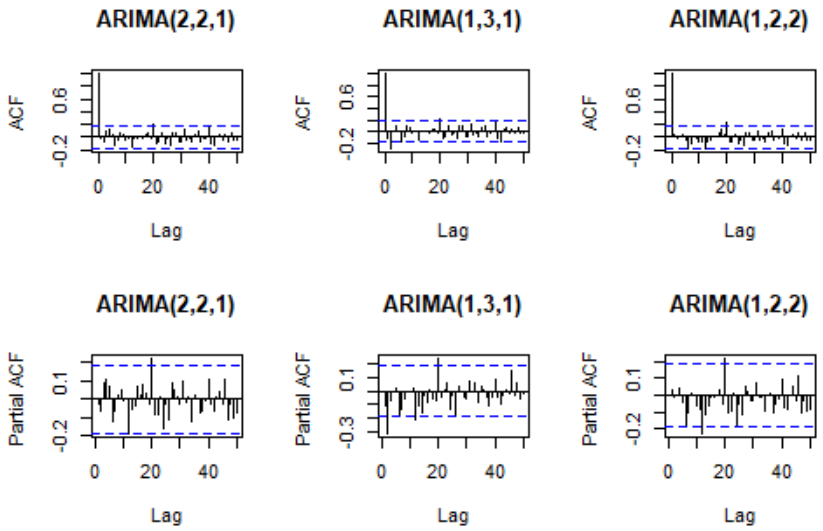

Figure 8: ACF and PACF of the residuals of ARIMA(2,2,1), ARIMA(1,3,1) and ARIMA(1,2,2) models fitted on log-prices in ZIP2000 over time

Neither of these models shows a clear improvement in the ACF and the PACF of the residuals. The final model chosen is thus ARIMA(1,2,1) fitted on the log-data.

In order to validate this model, its residuals are plotted in different ways on figure 9:

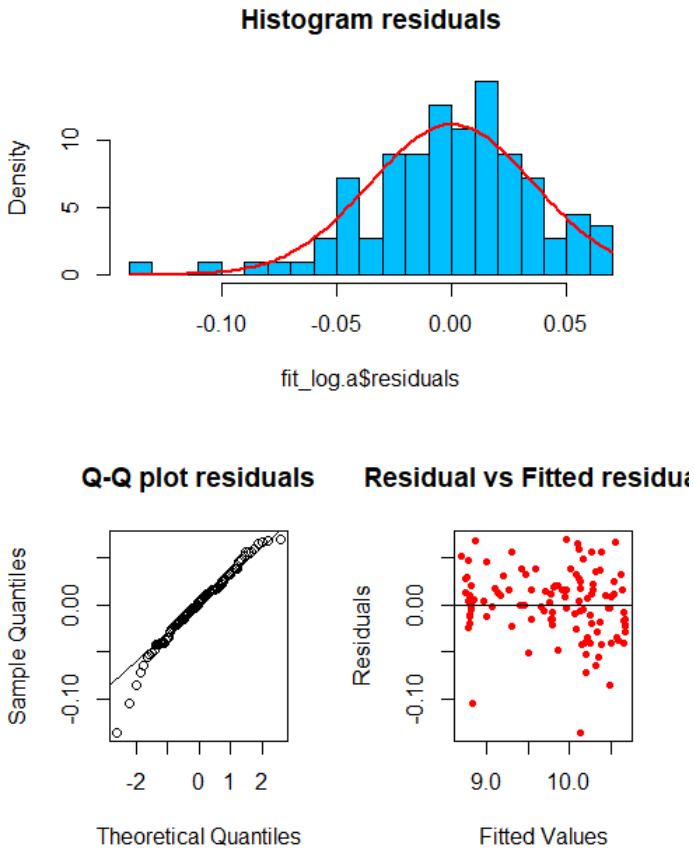

Figure 9: Histogram, QQ plot and spread of the residuals of ARIMA(1,2,1) model fitted on log-prices in ZIP2000 over time

A slight negative skewness can be observed in the distribution of the residuals. But overall, the residuals seem to follow quite accurately a normal distribution, with a constant mean of 0 and a constant variance. The model can thus be considered accurate enough to generate predictions from it.

2 Task 2 - Scenario generation

Using the time series model built in Task 1, it is now possible to generate scenarios for future price values. To generate a scenario, a random error term is sampled from a normal distribution $\mathcal{N}(0,\sigma)$ for each time step. The predicted values are then calculated using the ARIMA model stated in equation 1.

The aim here is to predict the prices in area ZIP2000 for the next period. 100 scenarios are generated using the R function *simulate*:

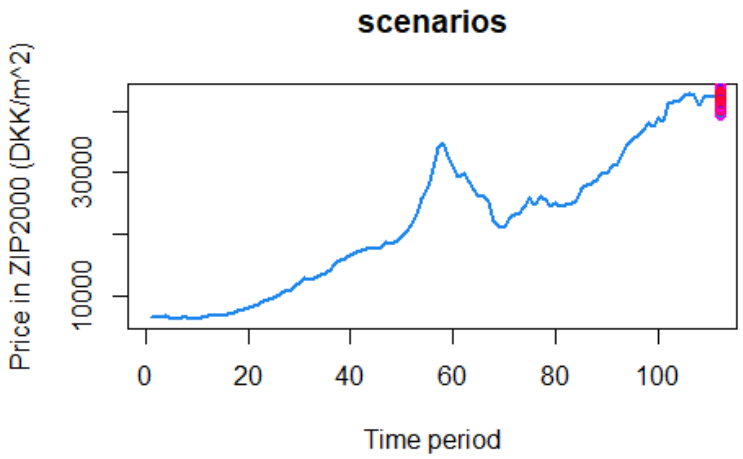

Figure 10: 100 scenarios to predict the investment price in ZIP2000 in the next time period

These scenarios are spread around the value of the price for the last time unit, and indicate some quite different possibilities - increasing, decreasing or stable prices. Figure 11 shows a zoom on the last time periods:

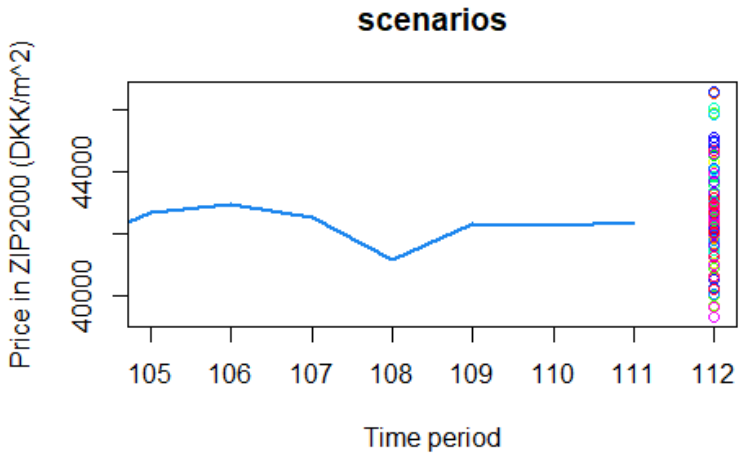

Figure 11: 100 scenarios to predict the investment price in ZIP2000 in the next time period (zoom)

It is often useful to perform some scenario reduction, in order to group the relatively high number of scenarios into a smaller number of representative scenarios. Indeed, these scenarios may be used in stochastic programs, whose computational time can limit the possibility of using large sets of scenarios.

Thus, 10 representative scenarios are built using the clustering method known as *partitioning around medoids* (pam). A pam algorithm selects a specified number of representative scenarios among the initial set of scenarios, dividing the set into clusters of similar scenarios. Figure 12 shows the 10 scenarios selected with this method using the R function *pam*:

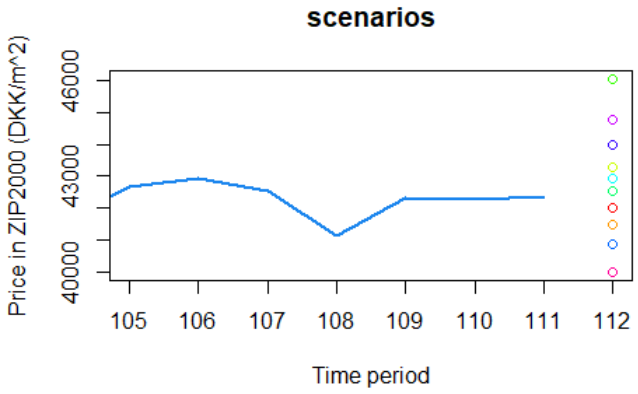

Figure 12: 10 representative scenarios to predict the investment price in ZIP2000 in the next time period (zoom)

While the 100 scenarios shown on figure 11 each have the same probability of occurring, it is not the case for the 10 scenarios shown on figure 12. Indeed, they each represent a cluster of scenarios, and each cluster does not hold the same number of scenarios. The probability of each of these 10 scenarios can thus be calculated by multiplying the probability of one scenario (here, 0.01) by the number of scenarios the cluster it represents contains.

Price	Probability
42006	0,11
41472	0,11
43265	0,12
46056	0,05
42558	0,13
42931	0,11
40891	0,09
43976	0,12
44780	0,09
39991	0,07

Table 1: Values and probabilities of prices given by 10 representative scenarios

3 Task 3 - Stochastic programming formulation

1) Using a set of 13 scenarios for the real estate value in the 4 zones *ZIP2000*, *ZIP2800*, *ZIP7400* and *ZIP8900* in the next time period, a two-stage stochastic program can be formulated to determine the optimal surface to buy initially in each area. In order to account for risk-management, a formulation including the conditional value-at-risk (CVaR) is considered. This allows to consider the values below the (1-α) quantile (the problem considered is a maximization problem).

Sets

$\mathscr{I}$	$\{ZIP2000, ZIP2800, ZIP7400, ZIP8900\}$	Area
$\mathscr{S}$	$[\![1, 12]\!]$	Scenario

Parameters

p_i^{init}	Initial price for buying one m^2 in area i [DKK/m^2]
$P_{i,s}$	Future price in area i according to scenario s [DKK/m^2]
B	Initial budget
π_s	Probability of scenario s
β	Weighing factor of the risk measure
α	Parameter of the quantile

Variables

b_i	$\mathbb{N}^+$	Surface bought in area i [m^2]	First-stage
η	$\mathbb{R}$	Value-at-risk	First-stage
δ_s	$\mathbb{R}$	$\max(0,\ \eta - \sum_{i\in\mathscr{I}} b_i(P_{i,s} - p_i^{init}))$	Second-stage

$$\max Z = (1-\beta)(\sum_{i\in\mathscr{I}, s\in\mathscr{S}} b_i \pi_s P_{i,s} - \sum_{i\in\mathscr{I}} b_i p_i^{init}) + \beta(\eta - \frac{1}{1-\alpha}\sum_{s\in\mathscr{S}} \pi_s \delta_s) \tag{2a}$$

$$\text{s.t.} \sum_{i \in \mathscr{I}} b_i p_i^{init} = B \tag{2b}$$

$$\eta - \sum_{i \in \mathscr{I}} b_i (P_{i,s} - p_i^{init}) \leq \delta_s \quad \forall s \in \mathscr{S} \tag{2c}$$

$$b_i \geq 0 \quad \forall i \in \mathscr{I} \tag{2d}$$

$$b_i \in \mathbb{N} \quad \forall i \in \mathscr{I} \tag{2e}$$

$$\delta_s \geq 0 \quad \forall s \in \mathscr{S} \tag{2f}$$

$$\tag{2g}$$

- Equation 2a is the objective function. The aim is to maximize both the expected gain related to the investment and the expected value of the objective values below the (1-α) quantile of the distribution.
- Constraint 2b states that the entirety of the initial budget has to be spent.
- Constraint 2c is necessary to account for the CVaR.

The results of the model for $\alpha = 0.9$ and $\beta = 0.2$ are given on figure 13:

```
Gain: 945708.827
CVaR: -4778259.362
Objective: -199084.811
Area: zip2000 3.0
Area: zip2800 326.0
Area: zip7400 13.0
Area: zip8900 981.0
```

Figure 13: Results of the stochastic model

2) Another formulation of the model is stated below: the expected value model. This time, no risk-management is considered. The variables are optimized considering the expected values of the uncertain parameters (this is no longer a two-stage stochastic program).

Sets

$\mathscr{I}$	$\{ZIP2000, ZIP2800, ZIP7400, ZIP8900\}$	Area
$\mathscr{S}$	$[\![1, 12]\!]$	Scenario

Parameters

p_i^{init}	Initial price for buying one m^2 in area i [DKK/m^2]
P_i^{exp}	Expected value of the future price in area i according to scenario s [DKK/m^2]
B	Initial budget

Variables

b_i	$\mathbb{N}^+$	Surface bought in area i [m^2]

$$\max Z = \sum_{i \in \mathscr{I}} b_i (P_i^{exp} - p_i^{init}) \tag{3a}$$

$$\text{s.t.} \sum_{i \in \mathscr{I}} b_i p_i^{init} = B \tag{3b}$$

$$b_i \geq 0 \quad \forall i \in \mathscr{I} \tag{3c}$$

$$b_i \in \mathbb{N} \quad \forall i \in \mathscr{I} \tag{3d}$$

$$\tag{3e}$$

- Equation 3a is the objective function. The aim is to maximize the expected gain related to the investment.
- Constraint 3b states that the entirety of the initial budget has to be spent.

The results of the model are given on figure 14:

```
Gain: 1166467.159
Objective: 1166467.159
Area: zip2000 8.0
Area: zip2800 724.0
Area: zip7400 28.0
Area: zip8900 25.0
```

Figure 14: Results of the expected value model

4 Task 4 - Out-of-sample test

A comparison between the stochastic model and the expected value model can be done by performing an out-of-sample test. The principle of this method is to compare the values given by the deterministic model when the decision variables are fixed to the values given by the 2 models, and this for a lot of scenarios. It is called *out-of-sample* because the scenarios used to perform the test are not necessarily the ones used in the stochastic model.

The overall process can be described as follows:

- Solve the stochastic model to get the value of the decision variable b_i^S
- Solve the expected value model to get the value of the variable b_i^{EXP}
- For each scenario:
  - Solve the deterministic problem with fixed surfaces bought b_i^S
  - Solve the deterministic problem with fixed surfaces bought b_i^{EXP}

The result of this test is plotted on figure 15. Since the scenarios all have the same frequency, the values have been plotted in increasing order of the expected gain.

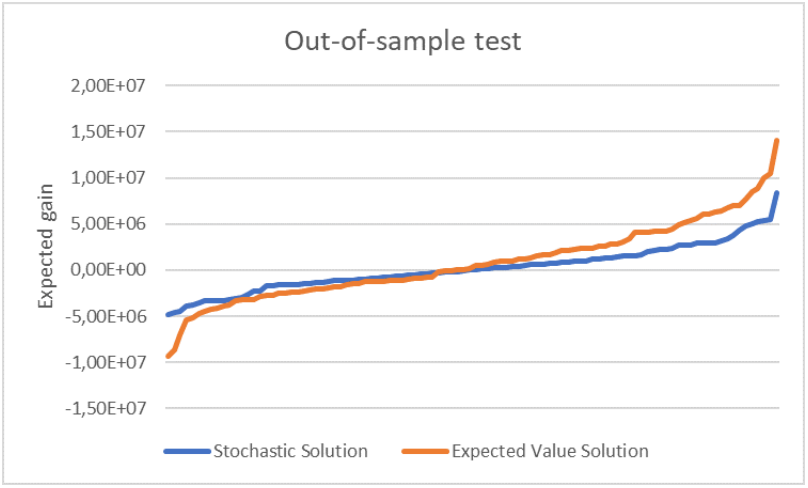

Figure 15: Out-of-sample test

A risk averse investor would prefer the stochastic solution over the expected value solution. Indeed, it can be seen on figure 15 that the expected gains given by the stochastic solution never reach extremely low values, even in unfavorable scenarios - at least compared to to expected values solution. Its drawback is that it does not give gains as high as the expected value solution in case of favorable scenarios. But a risk averse investor wants to make sure that its losses will not be too big in case of bad luck, so he will certainly choose

the stochastic solution.

Out-of-sample tests provide crucial information, that are not given by other comparison indicators such as the EVPI and VSS (expected value of perfect information and value of stochastic solution). Indeed, these values are calculated only on the basis of the scenarios used in the stochastic program. While they allow to take decisions regarding the first-stage variables values to choose, they do not necessarily cover the whole spectrum of the possible scenarios - which is most of the time continuous, and not discrete. This is why out-of-sample tests are so important to investors: they tell them what will happen if they choose one solution or another, depending on the infinite number of scenarios that can occur.