Abstract

The analysis of non-stationary data streams requires a continuous adaption of the model to the relevant most recent data. This requires that changes in the data stream must be distinguished from noise. Many approaches are based on heuristic adaptation schemes. We analyse simple regression models to understand the joint effects of noise and concept drift and derive the optimal sliding window size for the regression models. Our theoretical analysis and simulations show that a near optimal window size can be crucial. Our models can be used as benchmarks for other models to see how they cope with noise and drift.

1 Introduction

Collecting data has become a routine in many areas nowadays. But this also means that data analysis and machine learning must take into account that the data set is constantly growing in the form of a data stream and that the underlying data generating process might not be static but changes over time. Examples of real world applications involving data from a non-stationary process are for instance prediction of stock markets, weather, change of the protein structure through mutation, buying behaviour of
customers of an on-line store and many others.

A principle problem of handling data from a non-stationary process is the distinction between random noise and true changes in the data stream. A very fast adaptation of a model to the data can lead to the undesired effect of tracking the noise instead of the changes in the data. However, a very slow model adaptation means that the model is too static and will not fit recent data very well when changes have occurred. Hence the effective treatment of non-stationarity is an important problem in machine learning.

For non-stationary data generating processes one distinguishes between concept drift and concept shift. Concept drift – in statistics also called structural drift – refers to continuous changes of the data generating process, whereas concept shift – in statistics also called structural break – means sudden changes in the data generating process. Here, we focus exclusively on concept drift.

Various strategies are proposed to handle the problem of change detection in data streams. The change of the data distribution and changes in terms of concept drift can be detected as proposed, for instance, by Tschumitschew & Klawonn (2010a). Adaptive learning algorithms for data with concept drift are for example proposed by Lindstrom et al. (2010); Ikonomovska et al. (2009, 2011); Chu et al. (2004). All these approaches focus on the question how to deal with changes in the data stream. However, it is also very important to know how does drift and noise in the data together affect the quality of prediction. Furthermore it might be important to have a “good” strategy for choosing which (and how many) data instances should be used for prediction, since it is not always optimal to use all previous data.

In this work we carry out a theoretical analysis for two data generating processes: a constant and a linear model with drift and noise. For these two models we consider a simple prediction task, the prediction of the next value which can be understood as a regression problem. As we assume to cope with streaming data, evolving systems should be used. For such problems, evolving systems have been proposed by Angelov & Filev (2004); Angelov & Lughofer (2008); Beringer & Hüllemeier (2007); Klawonn & Angelov (2006); Macias-Hernandez et al. (2007). In this work we use an extremely simplified evolving system based on a windowing technique. Under the assumption that the data generating process is known, we are interested in finding the optimal window size as a function of the process parameters. In such a way we can analyse the behaviour of the optimal window size depending on the parameters of the data generating process. These simple theoretical models for the data generating process could be also used as benchmarks for evolving systems Tschumitschew & Klawonn (2010b) to judge the performance of an evolving system for which the underlying data generating process is not known.

The aim of this paper is not to introduce a new algorithm suitable for regression problems within data streams. Our purpose is to better understand and quantify the joint effects of noise and drift. Although we consider very simple models, our analysis already shows that a heuristic choice of the parameters – which is very usual when the data generating process is not known – can lead to very bad results. In addition, our simple scenarios for generating data can be used as benchmarks for more sophisticated algorithms in the framework of data streams.

The paper is organised as follows. We will briefly review existing techniques for the selection of sliding window size in Section 2. In Section 3 we carry out a theoretical
analysis for a constant model with drift and noise. This can be considered as a regression problem with a constant function as the regression function. Section 4 introduces a simple linear model with drift and noise, which can be interpreted as linear regression. First we compute the optimal window size separately with respect to the prediction of the slope and intercept and finally for the actual target variable. Experimental results are discussed in Section 5. For the linear model with drift and noise different parameter settings have been used and empirical error functions are analysed. In all these consideration it is assumed that the parameters of the data generating process change over time, but that the meta-model is stationary, i.e. that the drift is random, but will not be extremely high in one period and very low in another period. Consequences for non-stationary meta-models are discussed in Section 6. We conclude our paper with remarks on future work and how machine learning could benefit from our results.

2 Related work

The majority of existing machine learning techniques for mining data streams uses a sliding time window of fixed size, another small part tries to adapt the size of a sliding window based for instance on the quality of prediction. Gather et al. (2006) apply robust regression techniques to the data streams. They analyse and compare four different techniques, such as repeated median, least median of squares, least trimmed squares and deepest regression. All these approaches are applied to sliding time windows of fixed length. Though the influence of the window size on the prediction is discussed in this work, nevertheless only heuristic methods for the choice of the widow size are used. Similar work is carried out in Davies et al. (2004), also here the robust methods for on-line regression are applied to the time windows with fixed size without taking the possible non-stationarity and noisiness of the data into account. In Nadungodage et al. (2011) not only the data from the current time window but also the previous data are used for prediction. As we will show in Section 5 this could have even more dramatic effects on the quality of prediction.

In the above approaches the size of the sliding window is selected without respect to the possible changes of the data concept whereas the techniques presented below try to select the window size based on some additional information like for instance prediction accuracy or relevance of the data instances. Klinkenberg (2004) automatically adjusts the window size in such a way that the estimated generalization error is minimized. Windows of different sizes are therefore used, the window with the highest prediction accuracy is finally chosen. This approach is computationally expensive. Therefore efficient computations are needed. An approach for supervised learning under the assumption of concept drift is presented by Beringer & Hüllemeier (2007). Here consistency, temporal and spatial relevance of the data is taken into account. Furthermore, a statistical test based on the prediction error is used to detect abrupt concept changes. When they occur, a certain amount of the instances is deleted randomly from the time window according to a distribution which is spatially uniform but temporally skewed. The number of instances to be removed is estimated based on the increase of the prediction error. The main idea of Zliobaite (2011) is to select the instances for the training data set based on the combination between space and time distance. The size
of the training data set is defined with the help of \( k \)-fold cross validation. Between \( N \) classifiers using the training sets with different sizes the classifier with the best classification accuracy is chosen. This approach requires high computational costs, since for each new data point the validation process should be repeated \( k \)-times for every training set (\( N \) different sets). Another problem is that the authors use the whole data set the for search of the relevant training instances. This is however impossible due to limited memory and computational time capacity.

In the next sections we will demonstrate that the choice of the window size is one of the crucial points in data stream mining in the presence of noise and drift and therefore this question deserves more attention.

3 Constant Model with Drift and Noise

A very simple example for systems with a random change over time is a one-di- mensional random walk, see, for instance, Spitzer (2001). A random walk \( (Y_t)_{t \in \mathbb{N}} \) is obtained by adding up values from independent and identically distributed (i.i.d.) random variables \( X_i \) with expected value zero \( E(X_i) = 0 \) and variance \( \sigma^2 = Var(X_i) = Var(X) \). We assume the random variables \( X_i \) to be normally distributed with mean \( \mu = 0 \) and unknown variance \( \sigma^2 \), i.e. \( X_i \sim N(0, \sigma^2) \).

\[
Y_t = \sum_{i=1}^{t} X_i
\]

(1)

The expected value for the random walk is then equal to zero:

\[
E(Y_t) = E\left( \sum_{i=1}^{t} X_i \right) = \sum_{i=1}^{t} E(X_i) = 0
\]

(2)

Furthermore the expected value is independent of \( t \), whereas the variance of the random walk increases linearly with \( t \).

\[
Var(Y_t) = Var\left( \sum_{i=1}^{t} X_i \right) = t \cdot Var(X) = t \cdot \sigma^2.
\]

(3)

According to equations (1), (2) and (3), the random variables \( Y_t \) follow a normal distribution, i.e. \( Y_t \sim N(0, t \cdot \sigma^2) \).

Furthermore the covariance of a random walk is given by

\[
Cov(Y_t, Y_s) = s \cdot \sigma^2, \quad (s < t)
\]

(4)

and also tends to infinity with increasing difference between the time points \( t \) and \( s \).

The theoretical analysis between an extremely simplified evolving system based on a windowing technique and a model tailored to the known assumptions that the data generating process is a random walk is discussed by Tschumitschew & Klawonn (2010b).

A slightly more complex model in comparison to the random walk is the following:
where $Y_t$ is a random walk $Y_t = \sum_{i=1}^{t} X_i$ and $X_i \sim N(0, \sigma^2)$.

The process (5) can be understood as a constant model with drift and noise, at least when the variance $\sigma^2$ of the underlying random walk $Y_t$ is small compared to the noise, generated by the normal distribution (5). $\sigma^2$ determines how fast or strong the drift is. Without loss of generality, we have chosen the variance $\sigma^2_{noise} = 1$ for the noise generating normal distribution, since for the analysis of the process only the proportion of the noise in comparison to the drift is of importance.

Figure 1 shows data generated by the process (5) with $\sigma = 0.2$. This data set exhibits drift as well as noise. In this case, the noise has a stronger effect on the data than the drift. The best strategy for the prediction of the next value for a simple random walk in terms of minimising the squared error of the prediction is the naïve approach, simply to use the last value as a prediction for the next one. However, in the process (5) the data have not only drift (random walk), but are also contaminated with noise. Hence we need to find out a better strategy for the prediction for the data generated by the process (5). In the following, we analyse how such a strategy depends on the proportion of the drift compared to the noise.

For the prediction we analyse a very simple evolving system based on a time window technique with a fixed window size of $T$. As a prediction for the next value, the mean of the last $T$ values is used. Therefore, the question that arises is now: how does the noise affect the optimal size $T$ of the window? For this purpose, we try to minimise the expected quadratic error. The expected quadratic error for the prediction of the next value can be computed as follows:

\[
Z_t \sim N(Y_t, 1)
\]
\[
E \left( \left( \frac{1}{T} \sum_{i=t_0}^{t} Z_i - Y_{t+1} \right)^2 \right) = \frac{(2t^2 - 4t \cdot t_0 + 7t + 2t_0^2 - 7t_0 + 6) \sigma^2 + 6}{6(t-t_0+1)} \tag{6}
\]

where \(t_0\) is the first value in the window, i.e. \(t_0 = t - T + 1\). The details for the derivation of this equation and also all following equations are provided in the appendix.

In order to determine the minimum of the function (6), we take the derivative with respect to the parameter \(t_0\), which yields

\[
\left( E \left( \left( \frac{1}{T} \sum_{i=t_0}^{t} Z_i - Y_{t+1} \right)^2 \right) \right)' = \frac{4t_0 - 4t - 7}{6(t-t_0+1)} \sigma^2 + \\
\frac{2t^2 - 4t \cdot t_0 + 7t + 2t_0^2 - 7t_0 + 6}{6(t-t_0+1)^2} \sigma^2 + 6. \tag{7}
\]

The right-hand side of the function (7) is zero, when \(t_0\) is chosen as

\[
t_0 = t + 1 \pm \sqrt{\frac{1}{2} + \frac{3}{\sigma^2}}. \tag{8}
\]

Hence the optimal window size is

\[
T = t - t_0 + 1 = \sqrt{\frac{1}{2} + \frac{3}{\sigma^2}}. \tag{9}
\]

From Equation (9) we can see – as would be expected – that with increasing drift (\(\sigma^2\)), the window size decreases. When \(\sigma^2 \geq 6\), i.e. the drift becomes too large in comparison to the noise, the window size shrinks to 1 as in the ordinary random walk.

We have assumed that \(Z_t\) has a constant variance of one. As mentioned before, the optimal window size depends only on the quotient of the noise and the drift. By assuming that the data generating process (5) has variance \(\sigma_1^2\), i.e.

\[
Z_t \sim N \left( y_t, \sigma_1^2 \right), \tag{10}
\]

and denoting the variance of \(X_t\) by \(\sigma_2^2\), we can recalculate the expected quadratic error. Hence, Equation (6) becomes

\[
E \left( \left( \frac{1}{T} \sum_{i=t_0}^{t} Z_i - Y_{t+1} \right)^2 \right) = \frac{1}{\sigma_2^2} \frac{(2t^2 - 4t \cdot t_0 + 7t + 2t_0^2 - 7t_0 + 6) \sigma_1^2 + 6}{6(t-t_0+1)} \tag{11}
\]

Accordingly, the optimal window size is then

\[
T = t - t_0 + 1 = \sqrt{\frac{1}{2} + \frac{3}{\sigma_1^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}}. \tag{12}
\]
Figure 2: Optimal windows size depending on the ratio of the drift and the noise.

Figure 2 shows Equation (12) as a function of the ratio of the drift and the noise $\frac{\sigma_1^2}{\sigma_2^2}$. We can see from the figure that with decreasing $\frac{\sigma_1^2}{\sigma_2^2}$, the windows size increases. That means if the noise in the data is stronger in comparison to the drift, for the prediction of the next value more previous values should be used (larger window size). And for stronger drift compared to the noise, the windows size tends to 1, which relates to the prediction for an ordinary random walk, since the noise is negligible in comparison to the drift.

4 Linear model with drift and noise

The constant model with drift and noise described in the previous section can be considered as a regression problem with a constant function as the regression function. In this section, we discuss an example for linear regression with drift and noise. The model for linear regression is given by

$$y_i = B_{1,i} \cdot x_i + B_{0,i} + \epsilon_i$$

(13)

where the random variables $B_0$ and $B_1$ are random walks (see Section 3, Equation (1)) and follow normal distributions: $B_{0,i} \sim N(0, i\sigma_0^2)$, $B_{1,i} \sim N(0, i\sigma_1^2)$. The random variable $\epsilon_i$ represents noise and is normally distributed with expected value zero and variance $\sigma^2$, i.e. $\epsilon_i \sim N(0, \sigma^2)$. It is assumed that the $\epsilon_i$-s are independent. Therefore, the model (13) has drifts in the slope $B_1$ and intercept $B_0$ of the regression line. Furthermore, the random variable $\epsilon_i$ adds noise to the linear relationship between the dependent variable $y$ and the predictor $x$.

In order to simplify the analysis, the sampling values $x_i$ for the predictor are assumed to be fixed in the following way: we sample repeatedly $n$ points $x_1, \ldots, x_n$. 
The values $x_l$ are equi-distant values from the sampling interval $[-a,a]$, i.e. $x_l = -a + (l-1) \frac{2a}{n-1}$ for $l = 1, \ldots, n$.

The sum of the values $x_l$ is equal to zero (14) and according to that the expected value of $X$ and the mean are also equal to zero: $\bar{x} = 0$.

$$\sum_{l=1}^{n} x_l = 0 \quad (14)$$

Figure 3 shows data generated by the process (13) with the following settings: $\sigma_1 = 0.7$, $\sigma_0 = 0.5$ and $\sigma = 0.1$. The values for the predictor $x$ are selected from the interval $[-1,1]$ with $n = 100$, the start value for the slope is chosen as 1 and for the intercept as 0.

We assume that the drift occurs only after a whole cycle of sampled predictor values $x_1, \ldots, x_n$, i.e. the random walks $B_0$ and $B_1$ are only updated after one sequence of the values $x_1, \ldots, x_n$. The line starting at the highest point corresponds to the first sampling cycle: $y_{1,l} = B_{1,1} x_l + B_{0,1} + \varepsilon_{1,l}, l = 1, \ldots, n$. It should be noted that the horizontal axis shows the sampled $x$-values of the corresponding to the sample cycle, not necessarily in the order in which they were sampled. The line starting at the lowest point corresponds to the second cycle where the slope and the intercept have drifted the first time. The line in the middle corresponds to the third sampling cycle, where another drift of the slope and the intercept has taken place. It can be seen clearly that the first two lines have almost the same intercept, but different slope, the second and third line have almost the same slope, but differ in their intercept. This means that after the first cycle, $B_{1,1}$ did a larger jump, whereas $B_{0,1}$ changed only very little. In contrast, in the next step the larger drift occurred for the intercept.

In order to predict the dependent variable $y$, we need to estimate the parameters of the model $B_0$ and $B_1$ first. A large variety of methods have been developed for the
estimation of the parameters of a linear model. In this work we restrict our considera-
tions to the simplest and most common method for linear regression: the least squares
estimator. The least squares principle for simple linear regression can be explained as
follows: choose the estimates for $B_0$ and $B_1$ in such a manner that the sum of squared
residuals becomes minimal. A formal development of the least squares estimates can
be found, for instance, in Weisberg (2005); Groß (2008). Furthermore, the least squares
estimator has nice statistical properties. The estimator is unbiased, consistent and effi-
cient under certain assumptions (see Groß (2008)).

Let $\hat{B}_0$ and $\hat{B}_1$ be the unbiased estimators for $B_0$ and $B_1$ (when no drift occurs). The
least squares estimates are in this case given by

$$\hat{B}_1 = \frac{\sum_{i=t_0}^{t} \sum_{l=1}^{n} (x_l - \bar{x}) y_{i,l}}{\sum_{i=t_0}^{t} \sum_{l=1}^{n} (x_l - \bar{x})^2},$$

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}.$$ (15)

In order to distinguish the estimators of $B_1$ and $B_0$ from the true parameters $B_1$ and
$B_0$, which are also random variables in our case, we denote the estimators by $\hat{B}_0$ and
$\hat{B}_1$.

Similar to the previous section, we use a sliding window technique for the estima-
tion of the next value of the slope and intercept. As already in Section 3, the window
starts at the value with index $t_0$ and the window size is denoted by $T$. According to the
fact that the model has noise and drifts in the slope and intercept the question presently
is: which effects have drifts and noise on the optimal window size?

First we analyse the optimal window size for $B_1$, without taking $B_0$ into account.
The estimator for $B_1$ is given by

$$\hat{B}_1 = \frac{\sum_{i=t_0}^{t} \sum_{l=1}^{n} x_l y_{i,l}}{\sum_{i=t_0}^{t} \sum_{l=1}^{n} x_l^2},$$

(15)

To define the best possible window size we have to minimise the expected quadratic
error. The expected quadratic error for the prediction of the slope can be computed as
follows:

$$E \left( \left( \frac{\sum_{i=t_0}^{t} \sum_{l=1}^{n} x_l y_{i,l}}{\sum_{i=t_0}^{t} \sum_{l=1}^{n} x_l^2} - B_{1,t+1} \right)^2 \right) = \sigma_1^2 \left( \frac{3 (t + t_0) + 2 (t - t_0) (t + 2t_0 - 1)}{6 (t - t_0 + 1)} + (1 - t_0) \right)$$

$$+ \frac{\sigma^2}{(t - t_0 + 1) \sum_{l=1}^{n} x_l^2}.$$ (16)

The minimum of the function (16) is at

$$t_0 = t + 1 - \sqrt{\frac{3}{2} \frac{\sigma^2}{\sigma_1^2} + \frac{1}{n}}$$

(17)
where $\tilde{n} = \sum_{l=1}^{n} x_{l}^{2} = \frac{a^{2}n(n+1)}{3(n-1)}$. According to Equation (17), the optimal window size is given as a function of $\sigma_{1}^{2}$ and $\sigma^{2}$ when $a$ and $n$ are considered as constant values.

$$T = \sqrt{\frac{1}{2} + \frac{3 \cdot \sigma^{2}}{\tilde{n} \cdot \sigma_{1}^{2}}}.$$ (18)

It is obvious that the optimal window size in this case is independent of the drift of the intercept ($\sigma_{0}^{2}$). The reason for this is the following. The random walks $B_{1}$ and $B_{0}$ are independent that means that the drift of the slope occurs independently of the drift of the intercept. Consequently, $B_{0,t+1}$ is irrelevant for the estimation of $B_{1,t+1}$.

In the next step we compute the expected quadratic error for the intercept $B_{0}$. The estimator of $B_{0}$ is $\hat{B}_{0} = \bar{y} - \hat{B}_{1} \bar{x}$, where in our case $\bar{x} = 0$ holds. Therefore the expected quadratic error for $B_{0,t+1}$ is given by

$$E \left( (\tilde{y} - B_{0,t+1})^{2} \right) = \frac{6 + 7t + 2t^{2} - 7t_{0} - 4t \cdot t_{0} + 2t_{0}^{2}}{6(t-t_{0}+1)} \sigma_{0}^{2} + \frac{\sigma^{2}}{n \cdot T}.$$ (19)

The minimum of the function (19) is similar to the one for the slope (17).

$$t_{0} = t + 1 - \sqrt{\frac{1}{2} + \frac{3 \cdot \sigma^{2}}{n \cdot \sigma_{0}^{2}}}.$$ (20)

Figure 4 shows the optimal window size for $\hat{B}_{0}$ and $\hat{B}_{1}$ corresponding to the functions (20) and (17) with $n = 90$ and $X \in [-1;1]$.

So far, we have computed the optimal window size separately for the estimator for the intercept $\hat{B}_{0}$ and for the estimator for the slope $\hat{B}_{1}$. However, usually the values for $B_{0,t+1}$ and $B_{1,t+1}$ are unknown. Only the value for $y_{t+1}$ can be observed. Therefore it makes sense to determine the optimal window size with respect to the prediction of the next value for $y$. For this purpose, the minimum of the expected quadratic error $E \left( (y - y_{t+1})^{2} \right)$ needs to be computed.

$$E \left( (y - y_{t+1})^{2} \right) = \left( \frac{6 + 7t + 2t^{2} - 7t_{0} - 4t \cdot t_{0} + 2t_{0}^{2}}{6T} \right) \left( x^{2} \sigma_{1}^{2} + \sigma_{0}^{2} \right)$$

$$+ \left( \frac{n \cdot x^{2} + \tilde{n}}{n \cdot \tilde{n} \cdot T} + 1 \right) \sigma^{2}.$$ (21)

The function (21) has its minimum at $t_{0} = t + 1 - \sqrt{\frac{1}{2} + \frac{3 \cdot \sigma^{2}}{b \cdot (x^{2} \sigma_{1}^{2} + \sigma_{0}^{2})}}$, therefore the optimal window size for the prediction is in this case

$$T = \sqrt{\frac{1}{2} + \frac{3 \cdot \sigma^{2}}{b \cdot (x^{2} \sigma_{1}^{2} + \sigma_{0}^{2})}},$$ (22)

where $b = \frac{n \cdot \tilde{n}}{n \cdot x^{2} + \tilde{n}}$. 

10
As we can see from Equation (22), the optimal window size is a function of \( \sigma^2 \), \( \sigma^2_1 \) and \( \sigma^2_0 \). Hence the choice of the optimal window size depends on the drifts in slope and intercept, but also on the noise. For highly noisy data with relatively small drifts, the best option is to look “more backward”, i.e. \( \lim_{\sigma^2 \to \infty} T = \infty \), whereas for very little noise only the last value should be used for the prediction, \( \lim_{\sigma^2_1 \to 0} T = \frac{1}{\sqrt{2}} \), i.e. \( T = 1 \) has to be selected in this case.

Note that a window of size \( m \) in our case contains \( m \cdot n \) points for the regression, since one sampling step corresponds to obtaining the points \((-a + (l-1) \frac{2a}{(n-1)}, y_l)\) \((l \in \{1, \ldots, n\})\). This means even a window of size 1 contains \( n \) sampling points. Apart from that, the effect of \( \sigma^2_1 \) depends on the sampled \( X \)-values and can be decreased or increased with decreasing or increasing \( X \). Since it is impossible to compute a window size for each new value \( x \), the expected value of \( X^2 \) could be used for \( x^2 \) in Equation (22).

By using different values \( t_0 \) in Equation (21), one for \( B_0 \) and one for \( B_1 \) (respectively \( t_0^{(0)} \) and \( t_0^{(1)} \)), the result would be equal to Equations (20) and (17), respectively.

Figure 5 shows the function (22), where \( n = 30 \), \( X \in [-1, 1] \) and instead \( x^2 \) we used the expected value of \( X^2 \), \( E (X^2) = 1/3 \).
5 Estimation of the optimal window size

In this section we present an experimental evaluation of our theoretical analysis of the optimal window size. For this purpose we consider an artificial data set. The data were generated according to the models described in Sections 3 and 4.

For the linear model with drift and noise, different settings for $\sigma_1$, $\sigma_0$ and $\sigma$ were used. In this way we simulate cases with different optimal window sizes. For each case the mean squared errors (MSE) for different values of $T$ were computed. The computations were carried out for samples of size 10000 that were generated according to Equation (10). The mean squared error over 10 repeated runs is shown in Figures 6, 7 and 8. Figure 6 shows the case with optimal theoretical window size of $T = 3$. The model has the following parameters: $\sigma_1 = 0.3$, $\sigma_0 = 0.3$ and $\sigma = 1$. This means that the drift and the noise have almost the same effect on the data, with the noise being slightly stronger than the drift. It is obvious that the empirical MSE-function in Figure 6 has a minimum point at $T = 3$ which is equal to the theoretical minimum. It is easy to see that using a window size different from the optimal size, the prediction error increases drastically. In that case, a window of size 25 will double the error.

Almost the same situation is shown in Figure 7. The data was generated with the following parameters: $\sigma_1 = 0.06$, $\sigma_0 = 0.06$ and $\sigma = 1$. Here the optimal theoretical window size is 12, which is also the minimum point for the empirical error function. By using a window size different from 12, the mean squared error increases clearly. The situation is even more dramatic when only the last cycle of the sampled data should be used ($T = 1$).

The situation in Figure 8 is different. Here the model has the following settings $\sigma_1 = 0.005$, $\sigma_0 = 0.005$ and $\sigma = 1$. Therefore, the drift is negligible in comparison to the noise. This is also reflected in the optimal size of the data window $T = 144$. As we can see in Figure 8, the empirical error function has a plateau like minimum at the point $T = 144$, so that a slight change of the window size does not have much effect on
Figure 6: MSE of a linear model depending on the window size with optimal window size $T = 3$.

Figure 7: MSE of a linear model depending on the window size with optimal window size $T = 12$.

Also for the constant model with drift and noise described in Section (3), data with different $\sigma_1$ and $\sigma_2$ were generated and the mean squared errors were computed. The achieved results for this model are similar to the results described above for the more complex linear model and are therefore not presented here.
Consequences for non-stationary meta-models

In the previous sections we have assumed that we have a non-stationary data generating process. But the meta-models were assumed to be stationary, i.e. the random drifts and the noise were assumed not to change over time. We would not have phases with higher noise or lower drift. Now we are interested in the following question: which effect would a non-stationary meta-model have on the quality of prediction? In this section we discuss some examples for non-stationary meta-models. As the underlying model we use the constant model described in Section 3.

There are two parameters in the meta-model: the step size \( \sigma_1^2 \) for the random walk and the intensity of the noise \( \sigma_2^2 \). We assume that only one of these two parameters changes randomly. For reasons of simplicity, we will assume that the random values for the standard deviation are generated by a normal distribution. This can in principle lead to a negative standard deviation, but since we only make use of the variance, i.e. the squared standard deviation in our models, this will not cause any problems.

For the first of our two non-stationary meta-models, we assume that the intensity of the noise \( \sigma_2^2 \) remains constant, but the step size in the random walk changes randomly. Therefore, in each step the (signed) standard deviation \( \sigma_1 \) of the normal distribution for the random walk is changed randomly, following a normal distribution with expected value \( \mu_{\sigma_1} \) and variance \( \sigma_{\sigma_1}^2 \).

\[
Z_t \sim N \left( \mu_{\sigma_1}, \sigma_{\sigma_1}^2 \right)
\]

\[
Y_t = \sum_{i=1}^{t} X_i, \quad X_i \sim N \left( 0, \sigma_1^2 \right)
\]

For the second type of a non-stationary meta-model assumes the step size of the random walk to be constant, but the intensity of the noise to be changing over time. So for
this model we assume
\[ \sigma_2 \sim N \left( \mu_{\sigma_2}, \sigma_{\sigma_2}^2 \right). \]  

Both kinds of non-stationary meta-models cannot be distinguished from a stationary meta-model with a comparatively much larger variance of the random walk or noise, respectively. Therefore, in fact, from such a non-stationary meta-model we can easily come to a stationary meta-model and all we need, is to estimate the optimal window size for the already known stationary meta-model with unknown variances.

![Figure 9: MSE curves for non-stationary (24) and stationary meta-models.](image)

The MSE curves for the non-stationary (upper line) and the stationary (lower line) meta-models are shown in Figure 9. Here the following settings are used: \( \sigma_2 = 0.5, \mu_{\sigma_1} = 0.1 \) and \( \sigma_{\sigma_1} = 0.5 \), therefore the change occurs in the meta-model for the random walk (24). The expected optimal window size for the stationary meta-model is \( T = 9 \), whereas the optimal window size of the non-stationary one is \( T = 2 \). The situation is similar for the changes of the variance of the noise (see Equation (25)). Here the optimal window size for the non-stationary meta-model is larger than the window size for the stationary meta-model. For sufficiently small variances \( \sigma_{\sigma_1} \) and \( \sigma_{\sigma_2} \), the optimal window size does not differ from the optimal size for the stationary meta-model. For instance, with \( \sigma_{\sigma_1} = 0.03 \) the optimal window size is exactly the same as for the stationary model. It is \( T = 9 \) in both cases.

As already mentioned before, such kind of non-stationarity is in fact a stationary meta-model. Therefore we consider the following non-stationary meta-model.

\[ Z_t \sim N \left( \gamma_t, \sigma_{Z_t}^2 \right), \]
\[ Y_t = \sum_{i=1}^t X_i, \quad Y_t \sim N \left( 0, \sigma_1^2 \right) \]
\[ \sigma_{Z_t} = \sum_{i=1}^t X_i^{(\sigma_1)}, \quad X_i^{(\sigma_2)} \sim N \left( 0, \sigma_{\sigma_2}^2 \right) \quad \text{and} \quad X_i^{(\sigma_2)} = \mu_{\sigma_2}. \]  

Hence the standard deviation of the random variables \( Z_t \) is the value of another random walk. Moreover, since the variance of this random walk tends to infinity with increasing \( t \), we define a threshold, which should not be exceeded by this random walk.
This kind of non-stationarity could also be considered for the variance \( \sigma_1 \) of the original random walk. However, first we restrict our considerations to the situation described in Equation (26).

Figure 10: Data generated from the non-stationary meta-model in Equation (26)

Figure 10 shows the data generated by the process (26) with \( \sigma_1 = 0.05 \), \( \mu \sigma_2 = 0.5 \) and \( \sigma_2 = 0.01 \). Therefore in each step we have a different optimal window size \( T \) and according to this a different MSE curve (see Figure 11). Here we have computed the MSE curves for the next 5000 data points at different time points. For instance, the uppermost line shows the MSE curve for the first 5000 values, the lowest one for the data points with the indices between 25000 – 30000. The second line from below represents the MSE curve for a stationary meta-model. As can be seen in Figure 11, each MSE curve has a different minimum. The minimum points are indicated in the figure by vertical dashed lines.

Hence, instead of computing one MSE curve for all data as before, we analyze the time behaviour of the data. For this purpose we use the windowing technique. The optimal window size for the stationary meta-model (\( \sigma_2 = 0.5 \)) is \( T = 17 \). For a time frame of size 5000, the MSE using \( T = 17 \) is computed. Afterward the window is moved and so forth. The measurements are shown in Figure 12. The highly fluctuating curve corresponds to the non-stationary meta-model and the almost constant one to the stationary. The MSE for the non-stationary meta-model is mostly larger than for the stationary, partly because of the larger variance of the noise, partly because the MSE of the non-stationary meta-model is computed with suboptimal \( T \).

As Figure 11 shows, the quality of prediction could be improved by using the optimal window size. For instance for the MSE curve for the time point \( t = 0 \) the optimal window size is 45. However, this improvement is not very noticeable, since the MSE curves have very flat minima. Hence, by using \( T = 17 \), we will gain slightly worse results.
Now we consider the situation where the use of the “wrong” window size could cause an even more drastic increase of the error.

\[ Z_t \sim N (y_t, \sigma_2^2), \]
\[ Y_t = \sum_{i=1}^{t} X_i, \quad X_i \sim N \left(0, \sigma_{1,t}^2\right) \]
\[ \sigma_{1,t} = \left| \sum_{i=1}^{t} X_i^{(\sigma_1)} \right|, \quad X_i^{(\sigma_1)} \sim N \left(0, \sigma_{\sigma_1}^2\right) \quad \text{and} \quad X_i^{(\sigma_1)} = \mu_{\sigma_1} \]  

(27)

For the model (27) we have the following settings: \( \sigma_1 = 0.003, \mu_{\sigma_2} = 0.5 \) and \( \sigma_{\sigma_1} = 0.0001 \). Therefore, the optimal window size for the corresponding stationary meta-model is \( T = 289 \). As Figure 13 shows, using a constant window size of 289 will lead by some of the MSE curves to a rather poor prediction, since each MSE curve has
a different minimum. The minimum point for stationary meta-model is indicated by vertical dashed lines.

From Figures 11 and 13 it is obviously that for such kind of non-stationary meta-model the optimal window size changes in each step. Therefore, $T$, estimated for the stationary meta-model might differ very much from the actual optimal window size. However, under the assumption that the parameters of the model are known and the values for $\sigma_d^2$, $\sigma_d$, and the threshold value are small enough, it might still be a good alternative.

7 Conclusions

We have proposed a theoretical analysis of the optimal sliding window size for the prediction of the next value of a regression problem in the context of data streams with drift and noise. For that purpose two simple theoretical models for the data generating process have been set up: a constant model with drift and noise and a linear model with drift and noise. For both models the minimum of the expected quadratic error was computed and the optimal window size has been determined as a function of the data generating process parameters. In such a way, we can better understand and quantify the effects of noise and drift on the optimal window size. For a stronger drift a very small window (in the extreme case a window of size 1) should be used, whereas for noisy data more of the previous values should be used for prediction.

In machine learning, data streams are often analysed based on heuristic methods. For non-stationary data streams, it can be crucial to find a compromise between using as many data as possible from the past to reduce the effect of noise and using only recent data to rule out the influence of drift. In the first case, predictions might be based on completely outdated data. In the latter case, one might tend to track the noise instead of the drift.
Our models are too simple for most real world applications. Nevertheless, they provide important insights on the combined effects of drift and noise. Furthermore, our models could be used as simple benchmark tests that can give an idea of how much machine learning algorithms designed for more complex data streams might be misled by drift and noise.

Our future work will focus on more complicated models. Moreover, the theoretical analysis for non-stationary meta-models will be carried out.

References

Angelov, P., & Filev, D. (2004). An approach to on-line identification of evolving takagi-sugeno models. *IEEE Trans. on Systems, Man and Cybernetics, part B, 34*(1), 484–498.

Angelov, P., & Lughofer, E. (2008). A comparative study of two approaches for data-driven design of evolving fuzzy systems: ets and flexfis. *Intern. Journal of General Systems, 37*(1), 45–67.

Beringer, J., & Hüllemeyer, E. (2007). Efficient instance-based learning on data streams. *Intell. Data Anal., 11*(6), 627–650.

Chu, F., Wang, Y., & Zaniolo, C. (2004). An adaptive learning approach for noisy data streams. In *ICDM*, (pp. 351–354).

Davies, P. L., Fried, R., & Gather, U. (2004). Robust signal extraction for on-line monitoring data. *J. Statist. Plann. Inference, 122*, 65–78.

Gather, U., Schettlinger, K., & Fried, R. (2006). Online signal extraction by robust linear regression. *Computational Statistics, 21*(1), 33–51.

Groß, J. (2008). *Linear Regression: Vol 175 (Lecture Notes in Statistics)*. Berlin: Springer.

Ikonomovska, E., Gama, J., & Džeroski, S. (2011). Learning model trees from evolving data streams. *Data Mining and Knowledge Discovery, 23*(1), 128–168.

Ikonomovska, E., Gama, J., Sebastião, R., & Gjorgjevik, D. (2009). Regression trees from data streams with drift detection. In *11th int conf on discovery science, LNAI, vol 5808*, (pp. 121–135). Berlin: Springer.

Klawonn, F., & Angelov, P. (2006). Evolving extended naive bayes classifiers. In S. Tsumoto, C. Clifton, N. Zhong, X. Wu, J. Liu, B. Wah, & Y. Cheung (Eds.) *Sixth IEEE International Conference on Data Mining: Workshops. IEEE, Los Alamitos*, (pp. 643–647).

Klinkenberg, R. (2004). Learning drifting concepts: Example selection vs. example weighting. *Intell. Data Anal., 8*(3), 281–300.
Appendix

Derivation of Equation (6)

\[
E \left( \left( \frac{1}{T} \sum_{t=0}^{t} Z_t - Y_{t+1} \right)^2 \right) = E \left( \frac{1}{T^2} \left( \sum_{t=0}^{t} Z_t \right)^2 - \frac{2}{T} \sum_{t=0}^{t} Z_t \cdot Y_{t+1} + Y_{t+1}^2 \right) \\
= \frac{1}{T^2} E \left( \left( \sum_{t=0}^{t} Z_t \right)^2 \right) - \frac{2}{T} \sum_{t=0}^{t} E (Z_t \cdot Y_{t+1}) + E (Y_{t+1}^2). \tag{28}
\]

where \( t_0 \) is the first value in the window, i.e. \( t_0 = t - T + 1 \).
The expression \(E(Z_t \cdot Y_{t+1})\) can be further simplified as follows:

\[
E(Z_t \cdot Y_{t+1}) = E \left( Z_t \cdot \left( Y_t + \sum_{j=t+1}^{t+1} X_j \right) \right)
\]
\[
= E(Z_t \cdot Y_t) + \sum_{j=t+1}^{t+1} E \left( Z_t \cdot X_j \right)
\]
\[
= E(Z_t \cdot Y_t) + \sum_{j=t+1}^{t+1} E(Z_t) \cdot E(X_j)
\]
\[
= E(Z_t \cdot Y_t).
\]  

(29)

Taking equations (29) and \(E(Y_{t+1}^2) = \text{Var}(Y_{t+1}) = (t+1) \cdot \sigma^2\) into account, Equation (28) is transformed to

\[
E \left( \frac{1}{T} \sum_{i=0}^{T} \left( Z_i - Y_{t+1} \right)^2 \right) = \frac{1}{T^2} \sum_{i=0}^{T} E(Z_i^2) + \frac{2}{T^2} \sum_{i=0}^{T} \sum_{j=0}^{i-1} E(Z_i \cdot Z_j) - \frac{2}{T} \sum_{i=0}^{T} E(Z_i \cdot Y_t) + (t+1) \cdot \sigma^2.
\]  

(30)

In order to find the optimal window size, the minimum of Equation (30) as a function in \(t_0\) should be computed. For this purpose and to facilitate the readability of the text, the terms from formula (30) will be treated separately.

First of all, we calculate the expected value of \(Z_t\).

\[
E(Z_t) = \int_{-\infty}^{+\infty} \mu \cdot f_Y(\mu) d\mu = \frac{1}{\sqrt{2\pi a}} \int_{-\infty}^{+\infty} \mu \cdot e^{-\frac{\mu^2}{2a}} d\mu = 0
\]

where \(a\) is the variance of \(Y_t\) with \(a = i \cdot \sigma^2\).

The variance of \(Z_t\) is

\[
\text{Var}(Z_t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (z - \mu)^2 \cdot f_{Z_t|Y_t}(z) \cdot f_Y(\mu) \, dz \, d\mu
\]
\[
= \frac{1}{2\pi \sqrt{a}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} z^2 \cdot e^{-\frac{(z-\mu)^2}{2a}} \, dz \, d\mu
\]
\[
= \frac{1}{2\pi \sqrt{a}} \left( 2\pi a^{3/2} + 2\pi \sqrt{a} \right)
\]
\[
= a + 1.
\]  

(31)

For the expected value of the product \(Z_t \cdot Y_t\), we obtain from \(E(Y_t) = 0\) and \(E(Z_t) = 0\)
of Equation (30) the following expression.

\[
E(Z_i \cdot Y_i) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (z - \mu_z) (y - \mu_y) \cdot f_{Z_i Y_i}(z, y) \, dz \, dy
\]

\[
= \frac{1}{2\pi\sqrt{\sigma}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} z \cdot y \cdot e^{-(\frac{z^2 + y^2}{2\sigma})^2} \, dz \, dy
\]

\[
= \frac{1}{2\pi\sqrt{\sigma}} \cdot 2\pi a^{3/2}
\]

\[
= a. \quad (32)
\]

The last term to be computed is \(E(Z_i \cdot Z_j)\). As mentioned above, \(a\) is the variance of \(Y_i\) and \(a = i\sigma^2\). Moreover, we assume that \(j = i + k\) and \(k \neq 0\), corresponding to \(Y_j = Y_i + \sum_{l=i+1}^{i+k} X_l\). According to the infinite divisibility of the normal distribution, the sum of the random variables \(X_l\) follows also a normal distribution: \(\sum_{l=i+1}^{i+k} X_l \sim N(0, k\sigma^2)\). Furthermore, let us denote \(E(Y_i) = \mu\) and \(E(Y_j) = \mu + \eta\). Therefore, \(E(Z_i \cdot Z_j)\) can be computed as follows:

\[
E(Z_i \cdot Z_j) = \frac{1}{4\pi^2 \sqrt{a \cdot b} \cdot 4\pi^2 a^{3/2} \sqrt{b}} = \frac{1}{3} \cdot 4\pi^2 a^{3/2} \sqrt{b} = a. \quad (33)
\]

Based on equations (31), (32) and (33), we obtain the following expression for Equation (30)

\[
E\left(\left(\frac{1}{T} \sum_{i=0}^{T} Z_i - Y_{i+1}\right)^2\right) = \frac{1}{T^2} \sum_{i=0}^{T} (i\sigma^2 + 1) + \frac{2}{T^2} \sum_{i=0}^{T} \sum_{j=0}^{i-1} j\sigma^2 - \frac{2}{T} \sum_{i=0}^{T} i\sigma^2 + (t + 1) \cdot \sigma^2
\]

\[
= \frac{3(t + t_0) + 2(t - t_0)(t + 2t_0 - 1) + 6(1 - t_0)(t - t_0 + 1) \cdot \sigma^2 + 1}{6T}
\]

\[
= \frac{(2t^2 - 4t \cdot t_0 + 7t + 2t_0^2 - 7t_0 + 6) \cdot \sigma^2 + 6}{6(t - t_0 + 1)}. \quad (34)
\]

**Derivation of Equation (11)**

In fact only Equation (31) needs to be changed to

\[
Var(Z_i) = i \cdot \sigma_i^2 + \sigma^2. \quad (35)
\]
Hence, Equation (34) becomes

\[
E \left( \left( \frac{1}{T} \sum_{t=0}^{T} Z_i - Y_{i+1} \right)^2 \right) = \frac{(2t^2 - 4t \cdot t_0 + 7t - 2t_0^2 - 7t_0 + 6) \sigma_c^2 + 6 \sigma_t^2}{6 (t - t_0 + 1)}
\]

\[
= \frac{1}{\sigma_c^2} \frac{(2t^2 - 4t \cdot t_0 + 7t - 2t_0^2 - 7t_0 + 6) \sigma_t^2 + 6}{6 (t - t_0 + 1)}.
\]

(36)

**Derivation of Equation (16)**

\[
E \left( \left( \frac{\sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} y_{i+1,j} - B_{1,t} + 1}{\sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij}^2} \right)^2 \right) = E \left( \left( \frac{\sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} y_{i+1,j} - B_{1,t} + 1}{\sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij}^2} \right)^2 \right)
\]

\[
-2 \cdot B_{1,t+1} \left( \frac{\sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} y_{i+1,j} - B_{1,t} + 1}{\sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij}^2} \right) + B_{1,t+1}^2 \right)
\]

\[
= \frac{1}{(C)^2} E \left( \left( \sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} y_{i+1,j} \right)^2 \right) + E (B_{1,t+1})
\]

\[
- \frac{2}{C} E \left( B_{1,t+1} \sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} y_{i+1,j} \right)
\]

(37)

where \( C = \sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij}^2 = (t - t_0 + 1) \sum_{t=1}^{n} x_{ij}^2 \). It follows from Equation (14) that

\[
\sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} y_{i+1,j} = \sum_{t=0}^{t} B_{1,t} \cdot \sum_{j=1}^{n} x_{ij}^2 + \sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} \varepsilon_{i,j}
\]

(38)

holds. Now we compute each term separately.

\[
E \left( \left( \sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} y_{i+1,j} \right)^2 \right) = E \left( \left( \sum_{t=0}^{t} B_{1,t} \sum_{j=1}^{n} x_{ij} \right)^2 \right)
\]

\[
+ 2 \sum_{t=0}^{t} E (B_{1,t}) \sum_{j=1}^{n} x_{ij}^2 \sum_{t=0}^{t} E (\varepsilon_{i,j}) + E \left( \sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} \varepsilon_{i,j} \right)^2
\]

\[
= \sum_{t=0}^{t} E (B_{1,t}^2) \left( \sum_{j=1}^{n} x_{ij}^2 \right)^2 + 2 \sum_{t=0}^{t} \sum_{j=0}^{t} E (B_{1,t} B_{1,j}) \left( \sum_{j=1}^{n} x_{ij}^2 \right)^2
\]

\[
+ E \left( \sum_{t=0}^{t} \sum_{j=1}^{n} x_{ij} \varepsilon_{i,j} \right)^2.
\]

(39)
The first two terms in Equation (39) are easy to compute. For the computation of the third term further steps are needed.

$$E \left( \sum_{i=0}^{n} \sum_{l=1}^{n} x_l \varepsilon_{l,i} \right)^2 = E \left( \sum_{i=0}^{t} \sum_{l=1}^{n} x_l \varepsilon_{l,i} \right)^2$$

$$+ 2 \sum_{i=0+1}^{t} \sum_{j=0}^{i-1} \left( \sum_{l=1}^{n} x_l E(\varepsilon_{l,i}) \sum_{l=1}^{n} x_l E(\varepsilon_{l,j}) \right)$$

$$= \sum_{i=0}^{t} \sum_{j=0}^{n} x_l^2 E(\varepsilon_{l,i}^2) + 2 \sum_{i=0+1}^{t} \sum_{j=0}^{i-1} \sum_{m=1}^{n} x_l x_m E(\varepsilon_{l,i}) E(\varepsilon_{m,j}) .$$

Taking Equation (40) into account, (39) becomes

$$E \left( \sum_{i=0}^{t} \sum_{l=1}^{n} x_l y_{l,i} \right)^2 = \sum_{i=0}^{t} E(B_{1,i}^2) \left( \sum_{l=1}^{n} x_l^2 \right)^2 + 2 \sum_{i=0+1}^{t} \sum_{j=0}^{i-1} E(B_{1,i}B_{1,j}) \left( \sum_{l=1}^{n} x_l^2 \right)^2$$

$$+ E \sum_{i=0}^{t} \sum_{j=0}^{n} x_l^2 E(\varepsilon_{l,i}^2) .$$

Hence the first term of the expected quadratic error (37) is computed. Now we can compute the third term in Equation (37).

$$E \left( B_{1,t+1} \sum_{i=0}^{t} \sum_{l=1}^{n} x_l y_{l,i} \right) = \sum_{i=0}^{t} E(B_{1,t+1}B_{1,i}) \sum_{l=1}^{n} x_l^2 .$$

Therefore, we obtain the following equation for the expected quadratic error.

$$E \left( \left( \frac{\sum_{i=0}^{t} \sum_{l=1}^{n} x_l y_{l,i}}{\sum_{i=0}^{t} \sum_{l=1}^{n} x_l^2} - B_{1,t+1} \right)^2 \right) = \frac{1}{C^2} \left( \sum_{i=0}^{t} i \sigma_1^2 \left( \sum_{l=1}^{n} x_l^2 \right)^2 + 2 \sum_{i=0+1}^{t} \sum_{j=0}^{i-1} j \sigma_1^2 \left( \sum_{l=1}^{n} x_l^2 \right)^2 \right)$$

$$+ \sum_{i=0}^{t} \sum_{l=1}^{n} x_l^2 \sigma_1^2 \sum_{i=0}^{t} i \sigma_1^2 \sum_{l=1}^{n} x_l^2 + (t + 1) \sigma_1^2$$

$$= \sigma_1^2 \left( \frac{3(t + t_0) + 2(t - t_0)(t + 2t_0 - 1)}{6(t - t_0 + 1)} + (1 - t_0) \right)$$

$$+ \frac{\sigma_1^2}{(t - t_0 + 1) \sum_{l=1}^{n} x_l^2 .}$$

**Derivation of Equation (19)**

The estimator of $B_0$ is $\hat{B}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, where in our case $\bar{x} = 0$ holds. Therefore the expected quadratic error is $E \left( \left( \bar{y} - \hat{B}_0, t+1 \right)^2 \right)$. First we need to compute the expected
value of \( Y \).

\[
\bar{y} = \frac{1}{n \cdot T} \sum_{t=t_0}^T \sum_{i=1}^n (B_{t,i} x_t + B_{0,i} + \varepsilon_{t,i})
\]

\[
= \frac{1}{n \cdot T} \left( \sum_{t=t_0}^T B_{t,i} \sum_{j=1}^n x_j + \sum_{t=t_0}^T B_{0,i} \sum_{i=1}^n 1 + \sum_{t=t_0}^T \sum_{i=1}^n \varepsilon_{t,i} \right)
\]

\[
= \frac{1}{n \cdot T} \left( \sum_{i=0}^T n \cdot B_{0,i} + \sum_{t=t_0}^T \sum_{i=1}^n \varepsilon_{t,i} \right). \tag{44}
\]

According to Equation (44), we have

\[
E \left( (\bar{y} - B_{0,t+1})^2 \right) = \frac{1}{T^2} E \left( \left( \sum_{i=0}^T B_{0,i} \right)^2 \right) + \frac{1}{n^2 \cdot T^2} E \left( \left( \sum_{i=0}^T \sum_{i=1}^n \varepsilon_{t,i} \right)^2 \right)
\]

\[
+ E \left( (B_{0,t+1})^2 \right) + \frac{1}{n \cdot T} E \left( \sum_{i=0}^T B_{0,i} \sum_{i=1}^n \varepsilon_{t,i} \right)
\]

\[
- \frac{2}{T} E \left( \sum_{i=0}^T B_{0,i} : B_{0,t+1} \right) - \frac{2}{n \cdot T} E \left( B_{0,t+1} \sum_{i=0}^T \sum_{i=1}^n \varepsilon_{t,i} \right)
\]

\[
= \frac{1}{T^2} \sum_{i=0}^T i \sigma_0^2 + \frac{2}{T^2} \sum_{i=0}^T \sum_{j=0}^{i-1} j \sigma_0^2 - \frac{2}{T} \sum_{i=0}^T i \sigma_0^2 + (t+1) \sigma_0^2
\]

\[
+ \frac{1}{n^2 \cdot T^2} E \left( \left( \sum_{i=0}^T \sum_{i=1}^n \varepsilon_{t,i} \right)^2 \right). \tag{45}
\]

The only remaining term to be computed is \( \frac{1}{n^2 \cdot T^2} E \left( (\sum_{i=0}^T \sum_{i=1}^n \varepsilon_{t,i})^2 \right) \).

\[
\frac{1}{n^2 \cdot T^2} E \left( \left( \sum_{i=0}^T \sum_{i=1}^n \varepsilon_{t,i} \right)^2 \right)
\]

\[
= \frac{1}{n^2 \cdot T^2} E \left( \left( \sum_{i=0}^{T+t_0-1} \varepsilon_{i,i} \right)^2 \right)
\]

\[
= \frac{1}{n^2 \cdot T^2} \sum_{i=0}^{nT+t_0-1} E(\varepsilon_{i,i}^2) + 2 \sum_{i=0}^{nT+t_0-1} \sum_{j=i+1}^{nT+t_0-1} E(\varepsilon_i \varepsilon_j)
\]

\[
= \sigma^2 \frac{nT+t_0-1}{n^2 \cdot T^2} \sum_{i=0}^{nT+t_0-1} 1
\]

\[
= \frac{\sigma^2}{n^2 \cdot T^2} (nT+t_0-1-t_0+1) = \frac{\sigma^2}{n \cdot T}. \tag{46}
\]
The expected quadratic error for $B_{0,t+1}$ is therefore

$$E \left( \left( \bar{y} - B_{0,t+1} \right)^2 \right) = \frac{6 + 7t + 2t^2 - 7t_0 - 4t \cdot t_0 + 2t_0^2 \sigma_0^2 - \sigma^2}{6(t - t_0 + 1)} n \cdot T. \tag{47}$$

**Derivation of Equation (21)**

$$E \left( (y - y_{t+1})^2 \right) = E \left( \left( \hat{B}_1 x + \hat{B}_0 - B_{1,t+1} x - B_{0,t+1} - \varepsilon_{t+1} \right)^2 \right)$$

$$= E \left( \left( \left( \hat{B}_1 - B_{1,t+1} \right) x + \left( \hat{B}_0 - B_{0,t+1} \right) - \varepsilon_{t+1} \right)^2 \right)$$

$$= E \left( \left( \hat{B}_1 - B_{1,t+1} \right)^2 \right)x^2 + E \left( \left( \hat{B}_0 - B_{0,t+1} \right)^2 \right)$$

$$+ E \left( \varepsilon_{t+1}^2 \right) + 2E \left( \left( \hat{B}_1 - B_{1,t+1} \right) \left( \hat{B}_0 - B_{0,t+1} \right) \right)$$

$$+ 2E \left( \left( \hat{B}_1 - B_{1,t+1} \right) \varepsilon_{t+1} \right) + 2E \left( \left( \hat{B}_0 - B_{0,t+1} \right) \varepsilon_{t+1} \right). \tag{48}$$

The first two terms are the same as in the analysis before, so that we can use Equations (16) and (19). The third term is the variance of $\varepsilon$ and is therefore equal to $\sigma^2$. So we just need to evaluate the three remaining terms. In order to simplify the computation, we rewrite $\hat{B}_1$ and $\hat{B}_0$ in the following forms.

$$\hat{B}_1 = \frac{1}{\sum_{i=t_0}^t \sum_{j=1}^n x_{i}^2} \left( \sum_{i=t_0}^t \sum_{j=1}^n x_i x_{ij} + \sum_{i=t_0}^t \sum_{j=1}^n x_i \varepsilon_{ij} \right). \tag{49}$$

$$\hat{B}_0 = \frac{1}{n \cdot T} \left( \sum_{i=t_0}^t \sum_{j=1}^n B_{0,i} + \sum_{i=t_0}^t \sum_{j=1}^n \varepsilon_{ij} \right). \tag{50}$$

From Equation (49) we obtain that $\hat{B}_1$ and $B_{0,t+1}$ are independent, so that $E \left( \hat{B}_1 B_{0,t+1} \right) = E \left( \hat{B}_1 \right) \cdot E \left( B_{0,t+1} \right) = 0$ holds. The same applies to $E \left( \hat{B}_0 B_{1,t+1} \right)$ and $E \left( B_{1,t+1} B_{0,t+1} \right)$ (see Equation (50)). Taking this fact into account, the remaining terms can be computed as follows.

$$E \left( \left( \hat{B}_1 - B_{1,t+1} \right) \left( \hat{B}_0 - B_{0,t+1} \right) \right) = E \left( \hat{B}_1 \hat{B}_0 \right) - E \left( \hat{B}_1 B_{0,t+1} \right) - E \left( \hat{B}_0 B_{1,t+1} \right)$$

$$+ E \left( B_{1,t+1} B_{0,t+1} \right). \tag{51}$$
Furthermore, we have

\[ E ( \hat{\beta}_1 \hat{\beta}_0 ) = \frac{1}{nT \sum_{i=0}^{n} \sum_{l=1}^{n} x_l^2} \left( E \left( \sum_{i=0}^{T} B_{1,i} \sum_{l=1}^{n} x_l^2 + \sum_{i=0}^{T} \sum_{l=1}^{n} x_l \epsilon_{l,i} \right) \right) \]

\[ \cdot E \left( \sum_{i=0}^{T} nB_{0,i} + \sum_{i=0}^{T} \sum_{l=1}^{n} \epsilon_{l,i} \right) \]

\[ = \frac{1}{nT \sum_{i=0}^{n} \sum_{l=1}^{n} x_l^2} \left( \sum_{i=0}^{T} \sum_{l=1}^{n} x_l E ( \epsilon_{l,i}^2 ) + \sum_{i=0}^{T} \sum_{l=1}^{n} \sum_{j=0}^{T} \sum_{m \in M} x_l E ( \epsilon_{l,i} ) E ( \epsilon_{j,m} ) \right) \]

\[ = \frac{\sigma^2}{nT \sum_{i=0}^{n} \sum_{l=1}^{n} x_l^2} \sum_{i=0}^{T} \sum_{l=1}^{n} x_l = 0 \quad (52) \]

where \( M = \{ h \in \{ 1, \ldots, n \} \mid j = i \Rightarrow h \neq l \} \).

According to Equation (52) we have

\[ E \left( ( \hat{\beta}_1 - B_{1,t+1} ) ( \hat{\beta}_0 - B_{0,t+1} ) \right) = 0. \quad (53) \]

Analogously to Equation (51), we obtain

\[ E \left( ( \hat{\beta}_1 - B_{1,t+1} ) \epsilon_{t+1} \right) = 0. \quad (54) \]

\[ E \left( ( \hat{\beta}_0 - B_{0,t+1} ) \epsilon_{t+1} \right) = 0 \quad (55) \]

Since the tree last terms in Equation (48) are equal to zero, we obtain the following expression for the expected quadratic error.

\[ E \left( (y - y_{t+1})^2 \right) = \left( \frac{6 + 7t + 2r^2 - 7h_0 - 4t \cdot t_0 + 2r^2}{6T} \right) \left( x^2 \sigma_1^2 + \sigma_0^2 \right) + \left( \frac{n \cdot x^2 + \bar{h}}{n \cdot \bar{h} \cdot T + 1} \right) \sigma^2. \quad (56) \]