Using Conjunctural Indices in Prediction Models for gas sales –A Case Study

Adrian Sigot

Lund University
Department of Automatic Control
September 2010
Abstract
This thesis evaluates whether it is possible to use conjunctural indices from Statistics Sweden (Statistiska Centralbyrån) in order to improve the sales predictions at a Swedish gas company. For this purpose sales data for three of the main gases are gathered from the year 1995 and forth. In the first step, this data is used to create ARX and State Space models. Sales predictions are made from these models and the quality of these predictions is measured. In the second step, the conjunctural indices from Statistics Sweden for the corresponding period is used as input values and new ARX and State Space models are created. Predictions are made also with these models and are then compared to the predictions without help from the conjunctural indices. The results are however not significantly better. Correlation analysis show strong correlation between the conjunctural indices and the sales for some intervals in the available sales period, but no thoroughgoing correlation and no strong correlation at the current time (year 2008-2010). The conclusion is thus that the available conjunctural indices cannot be used to improve the sales predictions. However, since the correlations have been strong during come periods, it could be worth monitoring the correlations in case they reappear.
1 Introduction

1.1 Background

Most large companies show interest in analyzing their business environment. Sales, production figures and other data is stored in vast data warehouses, available for business analysts to extract valuable information. For a long time much of the focus has been on historical data. However, the recent trend seems to be in the predictive analytics field, where historical data together with real-time information is used to predict future patterns. IBM’s recent $1.2 billion acquisition of SPSS, a company well-known for its predictive analytics technology, and the SAP module “Advanced Planning & Optimization” (APO) indicates the demand for not only insight, but also foresight.¹

Making predictive models based on solely past and new sales is however not an easy task, as sales are affected by countless numbers of factors. It thus important to identify explanatory variables, i.e known or measurable facts or phenomena which are in some way connected to the sales and could explain parts of it. The connection does not have to be direct, i.e the origin or the direct reason for the sales pattern, which clearly would be very difficult to identify or measure. More realistic might be to find some indirect connection where the observed phenomena could be describing something else.² At the Swedish industrial gas company AGA Gas, some preliminary research has been made concerning the connection between the sales of a number of their main product and various conjunctural indices provided by statistical institutions. It has been found that the sales of these gases has some grade of correlation with some of these conjunctural indices, and AGA Gas anticipates this connection being useful in creating sales predictions. The objective of this thesis is to analyze this connection in detail and evaluate whether it could be used to improve prediction models.

AGA Gas is northern Europe’s leading industrial gas company, with operations in Scandinavia and the Baltic region. The company is also market leader in liquefied petroleum gas, i.e. propane sold in cylinders and has ventures in biogas which it distributes. AGA Gas has over 400,000 customers in the manufacturing and process industry and in healthcare. Since 2000 AGA Gas is owned by the German concern The Linde Group.

Three of AGA Gas’ main products will be analyzed in this thesis. These are Acetylene, Gaseous Oxygen (GOX) and Shielding gases, which includes Argon, Argon mixes, Mison and Mison mixes.

Acetylene is a highly flammable gas and it is produced by a reaction between calcium carbide and water. When burned with oxygen, Acetylene produces the hottest flame of all the fuel gases, which reaches a temperature of 3200°C. This makes Acetylene a very useful fuel gas for welding, cutting, heat treatment and coating across a wide range of industries. Due to its highly localized heating it

¹ Chikcowski (2009)
² Angus (2008)
is also used in laboratories and research for optical spectrometry as an instrumentation gas.

Apart from being used together with Acetylene to create the oxy-acetylene torch explained above, another main area for Oxygen is in the steel manufacturing industry. Steel is produced by reducing the carbon content in iron from the blast furnace by injecting pure Oxygen into the molten iron under pressure, which reduces the carbon content from 4% to between 0.2% and 2%. Examples of other areas where Oxygen is used is in medicine to create breathing mixtures. Oxygen is also used to clean polluted rivers and lakes.

Shielding gases are used in welding to protect the molten weld pool and hot surrounding metal against ambient air, and to provide the electric arc with favorable conditions. Without such protection the metal would oxidize and the nitrogen in air would make the weld porous. Other factors that the shielding gas has an impact on include the speed of welding, corrosion resistance, mechanical properties and the working environment.\(^3\)

### 1.2 Tasks

The main tasks of this thesis is to evaluate whether it is advantageous to use conjunctural indices in order to improve models describing the sales of gases. In order to do this conjunctural indicies from Statistics Sweden are used and a casestudy is performed at AGA Gas, Sweden. Three gases are selected for the study. The following three subtasks are performed.

1. Analyze the sales of three main gases at AGA Gas, and map essential industrial segments.
2. Analyze the conjunctural indices for cross-correlation with the identified industrial segments from 1).
3. In the event of found cross-correlated conjunctural indices, compare and evaluate models created with and without help from these conjunctural indices.

### 1.3 Thesis disposition

The disposition for this thesis is as follows:

- **Chapter 1 - Introduction**: In this chapter the background to this Master Thesis and the issues and problems it addresses are introduced as well as the purpose of the study.

- **Chapter 2 - Methodology**: The methods used in the thesis are introduces as well as the available data.

- **Chapter 3 - Theory**: The theory used in the thesis is introduced and explained.

\(^3\)http://www.aga.se
Chapter 4 - Preparation of Data: In this chapter the available data is prepared for the later modeling.

Chapter 5 - Analysis and Discussion: The sought models are created and analyzed, followed by a discussion of the results.

Chapter 6 - Conclusions: In this chapter the obtained results are summarized.

1.4 Notes/remarks
Due to confidentiality agreements, some figures in this thesis have been omitted (sales figures 3-24). This should not affect the comprehensibility of this thesis in a negative manner as the analysis could be performed on an arbitrary sales sequence.

2 Methodology
2.1 Method
As the first subtask is to analyze the sales of the three main gases at AGA and to map the essential industrial segments within each of them, the first step is to attempt to segment the sales data. This is made possible through a detailed database of the customers, which specifies the monthly sales for each of AGA’s customers along with the customers’ industrial classification (ISIC). Through summation of the sales within each of the ISICs, the ISICs that are essential for each of the gases are determined. If the sales pattern for the essential ISICs differ enough from the total sales, the idea is to model the essential ISICs and the remaining ISICs separately. Furthermore, the essential ISICs are matched with various conjunctural indices from Statistics Sweden, in order to find indices which are cross-correlated with these ISICs. If such cross-correlations exists for some conjunctural index, it means that it in a way describes the sales of the essential ISICs, and can thus be used to help estimating models describing the sales. Suppose that, for instance, 75% of the sales are represented by 7 ISICs. Suppose further that the sales curves for these 7 essential ISICs differ from the total sales curve, and that individual models for these essential ISICs could be created and improved with help from various conjunctural indices. After having created a model for the remaining 25% of the sales as well, the underlying assumption is that a model for the total sales could be created by some kind of superposition of the individual models and that this model would be superior to a model created based only on total sales. The approach is illustrated in figure.

The next steps after the identification of the essential ISICs is thus analyzing whether the sales data differ from the total sales, followed by cross-correlation analysis between the identified ISICs and various conjunctural indices.

It was soon discovered that the essential ISICs did not differ significantly from the total sales. This meant that the original idea of creating individual
models for the essential ISICs had to be discarded, as the created models would resemble each other too much and thus not contribute to an assembled model in a positive manner. The new approach thus became to try and create models for the total sales with help from conjunctural indices cross-correlated with the total sales. As the essential ISICs represents the major part of the sales, they could still indicate what kind of conjunctural indices to look for.

A number of conjunctural indices cross-correlated with the total sales could be identified for each of the three gases. These indices was then used in the estimation of ARX and State Space models for the total sales. The created models could then be evaluated and compared to AR and State Space models created with the same methods, but without help from conjunctural indices.

![Figure 1: The used approaches](image)

### 2.2 Data

In this thesis there are three available datasets; two with volume sales figures for three dominating gases at AGA, and one containing conjunctural index series from Statistics Sweden. The sales data come in excel-files from AGA Gas’ SAP, and the index series come in excel-files from the WWW-page of Statistics Sweden. All data is imported to MATLAB with the command `xlsread`.

The first sales dataset ranges between 2004 and 2010, and shows sales per customer per month. Each customer is categorized in a specific industry code,
according to the characteristics of the customers business, which is called ISIC. ISIC codes could be, for instance, 'Motor Vehicles' or 'Construction'.

Example:

<table>
<thead>
<tr>
<th>ISIC-code</th>
<th>ISIC</th>
<th>Customer</th>
<th>Jan 2004</th>
<th>Feb 2004</th>
<th>...</th>
<th>Mars 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>4500</td>
<td>Construction</td>
<td>Builder AB</td>
<td>31.2</td>
<td>18.3</td>
<td></td>
<td>25.9</td>
</tr>
<tr>
<td>4500</td>
<td>Renting of automobiles</td>
<td>Car Rent AB</td>
<td>2.1</td>
<td>1.8</td>
<td></td>
<td>2.2</td>
</tr>
</tbody>
</table>

This dataset is vast as it contains sales figures for every single customer at AGA, which for a single gas can range up to 60 000 customers.

The second sales dataset ranges between 1995 and 2005, and do not have the customer segmentation as described for the first dataset. It thus shows only total sales per month for each gas.

The conjunctural index series from Statistics Sweden comprises the New Orders in Industry (NOI) and Industrial Production Index (IPI), please see section 3.5.2 for more details.
3 Theory

3.1 Time Series

A time series is a discrete sequence of data points over time, often at regular intervals. It is expressed as \( \{y_k, y_{k+1}, \ldots, y_n\} \) where \( y \) is the measured value at time \( k \) and \( n \) is the total number of samples. The characteristics of the time series is often affected by several factors, i.e seasonal variations or trends. The observations can be described as an \( n \)-dimensional random variable with the distribution \( p(y_1, y_2, \ldots, y_n) \).

3.1.1 Seasonal Variations

Time series sometimes show cyclical behavior and for time series observed on a monthly basis this cyclical pattern is often called seasonal variation. The seasonal cycle completes a whole cycle through a year, and repeats itself every subsequent year. An example of such seasonal variation in the Swedish industry is during the vacation during the July, where large parts of the industry is idle. It is often important to isolate and remove these seasonal effects from a time series in order to analyze it properly. This can be done in numerous ways and the method that will be presented here is called the ratio-to-moving-average method.

The Ratio-to-Moving-Average Method assumes that the time series can be explained by a so-called multiplicative model, described by \( Z_t = T_t S_t C_t I_t \), where \( Z_t \) is the observed time series, \( T_t \) the trend component, \( S_t \) the seasonal component, \( C_t \) the cyclical component (other period than \( S_t \)) and \( I_t \) is the irregular component (the error). The main idea is to find \( S_t \) and to deseasonalize the model by dividing it by \( S_t \). A way to find \( S_t \) is to remove or reduce the other products of \( Z_t \), namely \( T_tC_tI_t \). This is done in two steps. First a moving average of the time series is computed according to

\[
Z_{MA,t} = \frac{z_t + z_{t-1} + \ldots + z_{t-n}}{n+1},
\]

where \( n \) denotes the number of implicated values in the moving average. The calculation is repeated for every value of the time series starting at \( t = n \), and generates a new time series \( Z_{MA,t} \) with \( t - n \) values. This resulting smoothed series will contain the trend and cycle component, but most of the seasonality and error has been smoothed out by the moving average. By dividing each observation in the original time series by the corresponding value of the new series, the \( T_t \) and \( C_t \) components in the multiplicative model will be reduced and the \( S_t \) and \( I_t \) isolated according to

\[
\frac{Z_t}{Z_{MA,t}} = \frac{T_tC_tI_t}{T_tC_t} = S_t I_t,
\]

\[\text{Box (1994)}\]
which is called the ratio to moving average.

The second step is to deal with the irregular component \( I_t \). By averaging each seasonal value by the average of \( Z_t/Z_{MA,t} \) for the very same season, most of the irregular component \( I_t \) will be cancelled out and the seasonal component \( S_t \) is successfully isolated. Normally the values \( S_t \) are multiplied by 100 in order to get the seasonal component expressed as percentage. It is now a matter of simple division to obtain a deseasonalized time series from the original series:

\[
Z_{\text{deseasonalized},t} = \frac{T_t S_t C_t I_t}{S} = T_t C_t I_t.
\]  

(3)

It should be noted, however, that the new series contains less measured values than the original series as the moving average requires a certain number of values before it can be computed (depending on the number of included data points).

### 3.1.2 The covariance and correlation functions

For a stationary stochastic process, \( \{X(t)\} \), the autocovariance function for the time difference \( \tau = t_2 - t_1 \) is defined by

\[
\gamma_{XX}(\tau) = \text{Cov}[X(t_1), X(t_2)] = E[(X(t_1) - \mu(t_1))(X(t_2 + \tau) - \mu(t_2 + \tau))],
\]

where \( \mu = E[X(t)] \) and \( \tau \) is called the lag. The variance for the same process becomes \( \sigma^2(t) = \gamma_{XX}(0) \). The autocovariance is the covariance of the process against a time-shifted copy of itself. By dividing the autocovariance function with the variance, the autocorrelation function is obtained as

\[
\rho_{XX}(\tau) = \frac{\gamma_{XX}(\tau)}{\sigma^2_X} = \frac{\gamma_{XX}(\tau)}{\gamma_{XX}(0)}.
\]

The autocorrelation function is a normalized version of the autocovariance function and puts the measure in the range \([-1, 1]\), where 1 and -1 denotes perfect correlation and perfect anti-correlation, respectively. To illustrate this, \( \rho_{XX}(0) = \rho_{XX}(t_1, t_1) = 1 \), as a value always is perfectly correlated with itself. As both the autocovariance function and the autocorrelation function are even functions, and thus symmetric around lag \( \tau = 0 \), these functions are often plotted only for \( \tau \geq 0 \). Specifically, the plot for the autocorrelation is called a correlogram.

In practical situations where \( \mu \) might not be defined for a time series \( Y_1, Y_2, ..., Y_N \), the following function is used as an estimator for the autocovariance function

\[
C_{YY}(k) = C(k) = \frac{1}{N} \sum_{t=1}^{N-|k|} (Y_t - \bar{Y})(Y_{t+|k|} - \bar{Y}),
\]

for \( |k| = 0, 1, ..., N - 1 \) and \( \bar{Y} = (\sum_{t=1}^{N} Y_t)/N \). Similar to the theoretical autocorrelation function, the estimated correlations function is given by dividing

\[\text{Aczel (2002)}\]
the estimated autocovariance with the estimated variance, i.e the estimated
autocovariance function at lag $k = 0$, which gives
\[
\hat{\rho}(k) = r_k = \frac{C(k)}{C(0)},
\]
for $k = 0, 1, \ldots, N - 1$.

Now consider two stationary stochastic processes, \{\(X(t)\)\} and \{\(Y(t)\)\}. The
covariance between these processes is described by the \textit{cross-covariance function}
\[
\gamma_{XY}(\tau) = \text{Cov}[X(t_1), Y(t + \tau)] = E[X(t) - \mu_X(t)](Y(t + \tau) - \mu_Y(t + \tau)).
\]
With similarities to the autocorrelation function, the \textit{cross-correlation function}
is defined by
\[
\rho_{XY}(\tau) = \frac{\gamma_{XY}(\tau)}{\sigma_X \sigma_Y},
\]
which describes correlation measures in the same way as the autocorrelation
function, but between two different processes. It should be noted, just like
for the autocorrelation function, that \(|\rho_{XX}(\tau)| \leq 1\). The cross-correlation
function is not symmetric around lag $\tau = 0$, and therefore often plotted for both
negative and positive $\tau$ in order to analyze the function for both positive and
negative lags. This gives insight about not only the strength of the correlation
for different lags, but also its direction, i.e. whether values of the first process
is correlated with future or past values from the second process. A plot of the
cross-correlation function is called a \textit{cross-correlogram}.

The cross-covariance is in practical situations calculated by an \textit{estimator for
the cross-covariance function}, which is given by
\[
C_{XY}(k) = \frac{1}{N} \sum_{t=1}^{N-k} (X_t - \bar{X})(Y_{t+k} - \bar{Y})
\]
\[
C_{XY}(-k) = \frac{1}{N} \sum_{t=1}^{N-k} (X_{t+k} - \bar{X})(Y_t - \bar{Y}),
\]
for $k = 0, 1, \ldots, N - 1$ and \(\bar{X} = (\sum_{t=1}^{N} X_t)/N\) and \(\bar{Y} = (\sum_{t=1}^{N} Y_t)/N\). Based on
this estimated cross-covariance function, the \textit{estimated cross-correlation func-
tion} is given by
\[
\hat{\rho}_{XY}(k) = \frac{C_{XY}(k)}{\sqrt{C_{XX}(0)C_{YY}(0)}},
\]
where $C_{XX}(0)$ and $C_{YY}(0)$ are the estimated variances of $X_t$ and $Y_t$, respectively.\(^6\)

\(^6\)Madsen (2008)
3.2 AR(X)-Models

The AutoRegressive model of order \( p \), \( AR(p) \), is defined by the difference equation

\[
Y_t + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} = e_t,
\]

where \( a_p \) are model parameters and \( \{e_t\} \) is white noise, which is defined as a series of \( n \) uncorrelated stochastic variables \( \{e_t\}_{t=1}^{n} \) with \( E[e_t] = 0 \) and \( \text{Var}[e_t] = \sigma^2 \).

A more compact way of writing the difference equation is

\[
A(q) Y_t = e_t,
\]

where \( A(q) \) is a polynomial of order \( p \), and \( q \) is the delay operator, with the property \( q^{-k} Y_t = Y_{t-k} \). More specifically,

\[
A(q) = 1 + a_1 q^{-1} + \ldots + a_p q^{-p}.
\]

It is possible to add an input \( U_t \) to the \( AR(p) \)-model. In such cases the model is called AutoRegressive with eXogenous input, \( ARX(p) \), and is defined by the difference equation

\[
Y_t + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} = b_1 U_{t-1} + \ldots + b_p U_{t-p} + e_t,
\]

with parameters \( a_p \) and \( b_p \), and white noise-sequence \( \{e_t\} \). Similar to the \( AR \)-model, the \( ARX \)-model may be written in a compact form, given by

\[
A(q) Y_t = B(q) U_t + e_t,
\]

where \( A(q) \) and \( B(q) \) are polynomials of order \( p \).

The main task for \( AR \) identification is to identify the polynomial \( A(q) \) (together with \( B(q) \) in case of an \( ARX \) model). How this can be done is explained in the next section.

3.2.1 Identification of \( AR(X) \) models

A common method to estimate an \( AR(X) \) model is the \textit{Least Squares method}, which has its origins in the linear regression analysis.

Consider the general linear regression model

\[
y_t = \theta_0 + \theta_1 z_t + e_t,
\]

\[
\mathbf{x}_t = (1, z_t)^T, \quad \theta = (\theta_0, \theta_1)^T,
\]

\[\text{Madsen (2008)}\]
\[\text{Madsen (2008)}\]
\[\text{Ljung (1999)}\]
where the parameters $\theta$ are unknown. This general linear regression model can be expressed including $N$ available measurements by the matrix notation

$$
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_N
\end{bmatrix} =
\begin{bmatrix}
X_1^T \\
\vdots \\
X_N^T
\end{bmatrix} \theta +
\begin{bmatrix}
e_1 \\
\vdots \\
e_N
\end{bmatrix}
$$

or

$$
Y = x\theta + e,
$$

which is still linear in its parameters.

An estimation of the parameters $\hat{\theta}$ may be obtained by using the $N$ observations to minimize the sum of the squared errors between the model output, $f(x_t; \theta)$, and the observations, such as

$$
\hat{\theta} = \min_{\theta} S(\theta),
$$

where

$$
S(\theta) = \sum_{t=1}^{N} [y_t - f(x_t; \theta)]^2.
$$

Now consider the ARX($p$) model

$$
Y_t = -a_1 Y_{t-1} - ... - a_p Y_{t-p} + b_1 U_{t-1} + ... b_p U_{t-p} + e_t.
$$

Assume that the observations $Y_1, ..., Y_N$ are known. Introduce

$$
X_t =
\begin{bmatrix}
Y_1 & U_1 \\
\vdots & \vdots \\
Y_{N-1} & Y_{N-1}
\end{bmatrix}
$$

and

$$
\theta =
\begin{bmatrix}
\hat{A} \\
\hat{B}
\end{bmatrix} =
\begin{bmatrix}
a_1 & \cdots & a_p \\
b_1 & \cdots & b_p
\end{bmatrix}.
$$

and reformulate (5) in the linear form

$$
Y_t = X_t^T \theta + e_t.
$$

By further introducing

$$
Y^T = (Y_2, ..., Y_N) \\
X^T = (X_1, ..., X_N) \\
e^T = (e_1, ..., e_N),
$$

13
all the observations can be included in the model in the compact linear form

\[ Y = x\theta + e, \]

and the estimation of the parameter vector \( \hat{\theta} \) is again found by minimizing the sum of the squared errors

\[ S(\theta) = (Y - X\theta)^T(Y - X\theta). \]

The solution is given by

\[ \hat{\theta} = (X^T X)^{-1} X^T Y, \]

which is solved explicitly.\(^{10,11}\)

### 3.3 State Space Models

The relationship between the input, noise and output signals can be written as a system of first-order difference equations using an auxiliary state vector \( x(t) \) which is explaining the state of the system. This way of representing the system is called a state space representation. The state vector is a set of numbers \( x_1(t_0), x_2(t_0), \ldots, x_m(t_0) \), which, together with possible input-signals and noise disturbances at time \( t > t_0 \), uniquely describes the system at times \( t \geq t_0 \). A system with \( m \) states is called an \( m \)th order system. Representing the system in a state space form facilitates incorporation of physical mechanisms or other insights about the observed system into the model. The earlier described AR(\( X \)) model is an external description of the system since only the input to output relationship are modelled, whereas the state space representation defines an internal description as the state vector is not observable. When there is no available knowledge about the system, such as physical laws, the state vector must be determined in another way. This could be done either with a Kalman filter or with a subspace identification method, which both will be explained in later sections.\(^{12,13}\)

The (discrete time) linear stochastic state space model is given by

\[
\begin{align*}
X_t &= A_t X_{t-1} + B_t u_{t-1} + w_t \quad (7) \\
Y_t &= C_t X_t + D_t u_t + v_t \quad (8)
\end{align*}
\]

where \( X_t \) is the \( m \)-dimensional, not observable, state \( Y_t = C_t X_t + v_t \) vector. \( u_t \) is an optional input vector, \( Y_t \) the observed output, and \( A_t, B_t, C_t \) and \( D_t \) are the so called system matrices, which have to be estimated initially. \( w_t \) and \( v_t \) are process noise and measurement noise respectively, which are uncorrelated

\(^{10}\)Johansson (2009)  
\(^{11}\)Lang (1999)  
\(^{12}\)Ljung (1999)  
\(^{13}\)Madsen (2008)
and with the properties $E[w_t] = E[v_t] = 0$ and $Var[w_t] = \Sigma_w$, $Var[v_t] = \Sigma_v$. The system matrices $A_t, B_t, C_t$ and $D_t$ might depend on time, but are often constant, in which case the $t$-subscript is omitted. In the state space representation above, equation (7) is called the system equation which describes how the system states evolve and equation (8) is called the observation equation and describes what can be directly observed.

When talking about predictions, the State Space model (7)-(8) is often written as

$$\hat{X}_{t+1|t} = A\hat{X}_{t|t-1} + Bu_{t-1} + K\varepsilon_t$$  \hspace{1cm} (9)$$
$$Y_t = C\hat{X}_{t|t-1} + \varepsilon_t,$$  \hspace{1cm} (10)

where $\hat{X}_{t|t-1}$ is the prediction of $X_t$ at time $t-1$ and $\varepsilon_t$ equals the innovation, i.e the prediction error

$$\varepsilon_t = Y_t - \hat{Y}_{t|t-1}.$$

This form is thus called the innovation form. It should be noted that if $\hat{X}_{1|0}$ is known, then $\varepsilon_t$ can be calculated through (11) and thus all values $\hat{X}_{t|t-1}$ can be computed from past observations $Y_{t-1}, Y_{t-2}, \ldots$. The matrix $K$ is called the Kalman gain and will be discussed in a later section.  \footnote{Johansson (2009)}

### 3.3.1 Optimal Reconstruction

The Kalman filter yields the optimal reconstruction and prediction of the state vector $x(t)$ for a system expressed in a state space form according to (7)-(8), based on measurements of input and output signals $u$ and $y$. This being possible is guaranteed if the system is observable, i.e the observability matrix

$$O = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix}$$  \hspace{1cm} (12)

has full rank $n$. For systems without input $u_t$, (12) ensures that the $m$ equations

$$Y_t = CX_t$$
$$Y_{t+1} = CX_{t+1} = CA X_t$$
$$\vdots$$
$$Y_{t+m-1} = CX_{t+m-1} = CA^{m-1} X_t,$$

can be solved with respect to $X_t$. In other words, the $m$-dimensional state vector $X_t$ can be solved based on $m$ observations of the output.
The fundamental principles behind the Kalman filter are linear projections. Consider two random vectors \( \mathbf{X} = (X_1, \ldots, X_m)^T \) and \( \mathbf{Y} = (Y_1, \ldots, Y_n)^T \) and define the \((m + n)\)-dimensional vector \( \begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} \) with mean and covariance

\[
\begin{pmatrix}
\mu_Y \\
\mu_X \\
\end{pmatrix},
\begin{pmatrix}
\Sigma_{YY} & \Sigma_{YX} \\
\Sigma_{XY} & \Sigma_{XX} \\
\end{pmatrix}.
\]

It can then be shown that also \( \mathbf{X} | \mathbf{Y} \) is normally distributed with mean and variance

\[
E[\mathbf{X} | \mathbf{Y}] = \mu_X + \Sigma_{XY} \Sigma_{YY}^{-1} (Y - \mu_Y)
\]

\[
Var[\mathbf{X} | \mathbf{Y}] = \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX}.
\]

These equations also hold for the conditioned vector \( \begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} | Z \), and can then be rewritten as

\[
E[\mathbf{X} | \mathbf{Y}, Z] = E[\mathbf{X} | Z] + C[\mathbf{X}, \mathbf{Y} | Z] \Sigma_{YY}^{-1} (Y - E[\mathbf{Y} | Z]),
\]

\[
Var[\mathbf{X} | \mathbf{Y}, Z] = Var[\mathbf{X} | Z] - C[\mathbf{X}, \mathbf{Y} | Z] \Sigma_{YY}^{-1} C^T[\mathbf{X}, \mathbf{Y} | Z].
\]

Now define a vector \( \mathcal{Y}_t \) which contains all observations of \( y \) up to time \( t \):

\[
\mathcal{Y}_t = (Y_1^T, \ldots, Y_t^T).
\]

The reconstruction \( \mathbf{X}_{t|t} \) of \( \mathbf{X}_t \) can now be constructed, and the prediction \( \mathbf{X}_{t+k|t} \) of \( \mathbf{X}_t \) is expressed as

\[
\mathbf{X}_{t+k|t} = E[\mathbf{X}_{t+k} | \mathcal{Y}_t], \quad k \geq 1.
\]

It is known from the theory of regression that when minimizing the expected value of the squared prediction error the optimal reconstruction and prediction are obtained by the conditional mean. The prediction errors are introduced as

\[
\tilde{\mathbf{X}}_{t+k|t} = \mathbf{X}_{t+k} - \mathbf{X}_{t+k|t},
\]

\[
\tilde{\mathbf{Y}}_{t+k|t} = \mathbf{Y}_{t+k} - \mathbf{Y}_{t+k|t}.
\]

Using (13) we get the estimation error covariance

\[
\Sigma_{\tilde{X}_{t+k|t}} = Var[\mathbf{X}_{t+k} | \mathcal{Y}_t]
\]

\[
= E \left[ \left( \mathbf{X}_{t+k} - \mathbf{X}_{t+k|t} \right) \left( \mathbf{X}_{t+k} - \mathbf{X}_{t+k|t} \right)^T | \mathcal{Y}_t \right]
\]

\[
= E \left[ \tilde{\mathbf{X}}_{t+k|t} \tilde{\mathbf{X}}_{t+k|t}^T | \mathcal{Y}_t \right]
\]

\[
= Var \left[ \tilde{\mathbf{X}}_{t+k|t} | \mathcal{Y}_t \right].
\]

\[\text{Madsen (2008)}\]
The conditional means are calculated using linear projections, and it is thus known that
\[
C \{X_{t+k} - E[X_{t+k}|Y_t], Y_t\} = 0,
\]
which should be interpreted as the prediction error \( \tilde{X}_{t+k|t} \) and \( Y_t \) being uncorrelated. We can therefore rewrite (14) as
\[
\Sigma^{xx}_{t+k|t} = \text{Var} \{X_{t+k}|Y_t\} = \text{Var} \{\tilde{X}_{t+k|t}\}. \tag{15}
\]
Similarly, it can be shown that
\[
\Sigma^{yy}_{t+k|t} = \text{Var} \{Y_{t+k}|Y_t\} = \text{Var} \{\tilde{Y}_{t+k|t}\}, \tag{16}
\]
\[
\Sigma^{xy}_{t+k|t} = C \{X_{t+k}, Y_{t+k}|Y_t\} = C \{\tilde{X}_{t+k|t}, \tilde{Y}_{t+k|t}\}. \tag{17}
\]
To conclude, the optimal reconstruction of \( X_{t|t} \) is obtained by
\[
X_{t|t} = E\{X_t|Y_t, Y_{t-1}\} \\
= E\{X_t|Y_{t-1}\} \\
+ C \{X_t, Y_t|Y_{t-1}\} \text{Var}^{-1} \{Y_t|Y_{t-1}\} (Y_t - E\{Y_t|Y_{t-1}\}),
\]
which also can be expressed as
\[
\tilde{X}_{t|t} = \tilde{X}_{t|t-1} + \Sigma^{xy}_{t|t-1} \left( \Sigma^{yy}_{t|t-1} \right)^{-1} (Y_t - \tilde{Y}_{t|t-1}). \tag{18}
\]
The covariance for the construction error becomes
\[
\Sigma^{xx}_{t|t} = \Sigma^{xx}_{t|t-1} - \Sigma^{xy}_{t|t-1} \left( \Sigma^{yy}_{t|t-1} \right)^{-1} \Sigma^{xy}_{t|t-1}^T. \tag{19}
\]
### 3.3.2 Prediction and the Kalman Filter

By introducing the Kalman gain
\[
K_t = \Sigma^{xy}_{t|t-1} \left( \Sigma^{yy}_{t|t-1} \right)^{-1}, \tag{20}
\]
the optimal reconstruction and the related error covariance, equations (18) and (19), can be reformulated into
\[
\tilde{X}_{t|t} = \tilde{X}_{t|t-1} + K_t \left(Y_t - \tilde{Y}_{t|t-1}\right), \tag{21}
\]
\[
\Sigma^{xx}_{t|t} = \Sigma^{xx}_{t|t-1} - K_t \Sigma^{yy}_{t|t-1} K_t^T. \tag{22}
\]
It should be noted how the estimate of the state vector $X_t$ is improved by combining old values of the state vector together with a weighed measure of the error between new measurement $Y_t$ and the prediction $\hat{Y}_t$ based on past measurements. In order to obtain the new improved estimate of the state vector the one-step predictions of $\hat{X}_{t+1|t}$ and $\hat{Y}_{t+1|t}$ and the related variances and covariances need to calculated.\textsuperscript{17} By using equations (7) and (8) the one-step prediction is simply

$$
\hat{X}_{t+1|t} = A \hat{X}_t + Bu_t, \quad (23)
$$
$$
\hat{Y}_{t+1|t} = C \hat{X}_{t+1|t}. \quad (24)
$$

The prediction error then becomes

$$
\tilde{X}_{t+1|t} = X_{t+1} - \hat{X}_{t+1|t} = A (X_t - \hat{X}_t) + w_{t+1} = A X_t + w_{t+1},
$$
$$
\tilde{Y}_{t+1|t} = Y_{t+1} - \hat{Y}_{t+1|t} = C (X_{t+1} - \hat{X}_{t+1|t}) + v_{t+1} = C X_{t+1} + v_{t+1},
$$

and thus follows for the variances and covariances

$$
Var [\tilde{X}_{t+1|t}] = A Var [\tilde{X}_t] A^T + \Sigma_w \quad (25)
$$
$$
Var [\tilde{Y}_{t+1|t}] = C Var [\tilde{X}_{t+1|t}] C^T + \Sigma_v \quad (26)
$$
$$
C [\tilde{X}_{t+1|t}, \tilde{Y}_{t+1|t}] = Var [\tilde{X}_{t+1|t}] C^T. \quad (27)
$$

Using (15)-(17) we finally get

$$
\Sigma_{xx}^{t+1|t} = A \Sigma_{xx}^t A^T + \Sigma_w
$$
$$
\Sigma_{yy}^{t+1|t} = C \Sigma_{xx}^{t+1|t} C^T + \Sigma_v
$$
$$
\Sigma_{xy}^{t+1|t} = \Sigma_{xx}^{t+1|t} C^T.
$$

By using equations (20)-(22), (23)-(24) and (25)-(27) in turn, the state vector $X_t$ is estimated.\textsuperscript{18} This is called the Kalman filter, which is here summarized.

The Kalman Filter. The optimal linear reconstruction $X_{t|t}$ and prediction $X_{t+1|t}$ of a system in state space form according to (7)-(8) is obtained by alternating a correction and prediction scheme.\textsuperscript{19} The correction is given by

\textsuperscript{17}Lung, 1999
\textsuperscript{18}Johansson (2009)
\textsuperscript{19}Johansson (2009)
\[
\begin{align*}
\hat{X}_{t|t} &= \hat{X}_{t|t-1} + K_t (Y_t - C\hat{Y}_{t|t-1}), \\
\Sigma_{xx}^{t|t} &= \Sigma_{xx}^{t|t-1} - K_t \Sigma_{yy}^{t|t-1} K_t^T = \Sigma_{xx}^{t|t-1} - K_t C \Sigma_{xx}^{t|t-1},
\end{align*}
\]
with the Kalman gain
\[
K_t = \Sigma_{xx}^{t|t-1} C^T (\Sigma_{yy}^{t|t-1})^{-1},
\]
and the prediction is given by
\[
\begin{align*}
\hat{X}_{t+1|t} &= A\hat{X}_{t|t} + B u_t, \\
\Sigma_{xx}^{t+1|t} &= A \Sigma_{xx}^{t|t} A^T + \Sigma_w, \\
\Sigma_{yy}^{t+1|t} &= C \Sigma_{xx}^{t+1|t} C^T + \Sigma_v.
\end{align*}
\]

### 3.3.3 k-step ahead predictions

The Kalman filter, when used as a predictor, only generates a one-step ahead prediction. It is however straightforward to generate a \(k\)-step ahead prediction in a state space model by using the recursive formulas
\[
\begin{align*}
\hat{X}_{t+k|t} &= A\hat{X}_{t+k|t} + B u_{t+k}, \\
\Sigma_{xx}^{t+k|t} &= A \Sigma_{xx}^{t+k+1|t} A^T + \Sigma_w,
\end{align*}
\]
which follows directly from the state space model.

### 3.3.4 Identification of State Space Models

Identifying a state space model is equivalent to finding the system matrices \(A, B, C, D\) which fully describes the system. Classical methods consist of first identifying the unique transfer function for the system, and then with some realization technique obtain the state space model, i.e. the system matrices. Finding the transfer function can however be troublesome, especially for non-physical systems. Another approach are the subspace methods, who first construct a state estimate from available input and output data using tools of numerical linear algebra and then obtain the system matrices by solving a least-squares problem. These two fundamentally different approaches are illustrated in Figure 2.

The classical identification method will not be discussed here, instead a subspace identification method called \(N4SID\) (Numerical algorithms for Subspace state space System iDentification) is presented.
3.3.5 The N4SID Subspace Identification Method.

Define the matrix $Y_{0|k-1}$ as

$$Y_{0|k-1} = \begin{bmatrix} y(0) & y(1) & \cdots & y(N-1) \\ \vdots & \vdots & \ddots & \vdots \\ y(k-1) & y(k) & \cdots & y(k+N-2) \end{bmatrix},$$

where $y(k)$ is the output of the system at time $k$. Define in a similar manner $U_{0|k-1}$, $Y_{k|2k-1}$ and $U_{k|2k-1}$ where $u(k)$ is the input to the system. Also define $X_0$ and $X_k$ as the state vector before and after time $k$. For simpler notation, set $Y_p = Y_{0|k-1}$, $U_p = U_{0|k-1}$, $X_p = X_0$ and $Y_f = Y_{k|2k-1}$, $U_f = U_{k|2k-1}$, $X_f = X_k$, where the subscripts $p$ and $f$ denotes the past and the future, respectively.

It is straightforward to show that the input-output relation for the state space model can be written as

$$y_k(t) = O_k x(t) + \Psi_k u_k(t),$$

where $O_k$ is the extended observability matrix and $\Psi_k$ is a lower block Toeplitz matrix, which are defined by

$$O_k = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{k-1} \end{bmatrix}, \quad \Psi = \begin{bmatrix} D \\ CB & D \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \vdots \\ CB & D \\ CA^{k-2}B & \cdots & CB & D \end{bmatrix}$$

respectively.\(^\text{20}\) It follows directly that also the following equations hold:

$$Y_p = O_k X_p + \Psi_k U_p,$$
$$Y_f = O_k X_f + \Psi_k U_f.$$

\(^{20}\)Katayama (2005)
Now define
\[ W_p = \begin{bmatrix} U_p \\ Y_p \end{bmatrix}, \quad W_f = \begin{bmatrix} U_f \\ Y_f \end{bmatrix} \]
and consider \( W_p \) and \( W_f \) past and future data matrices. Under certain assumptions, it can be shown that the state vector \( X_f \) is a basis of the intersection of the past and the future subspaces. Recall from linear algebra that given row spaces \( A, B, C, \) we have for \( \alpha \in A \) the decomposition
\[
\hat{E} \{ \alpha | B \lor C \} = \hat{E} \{ \alpha | B \} + \hat{E} \{ \alpha | C \},
\]
The left-hand side is the orthogonal projection and the right-hand side is a direct sum decomposition where \( \hat{E} \{ \alpha | B \} \) is the oblique projection of \( \alpha \) onto \( B \) along \( C \) and \( \hat{E} \{ \alpha | C \} \) is the oblique projection of \( \alpha \) onto \( C \) along \( B \).

Let the oblique projection of \( Y_f \) onto \( W_p \) along \( U_f \) be given by
\[
\xi = \hat{E} \{ Y_f | W_p \}
\]
and the Singular Value Decomposition of \( \xi \) by
\[
\xi = [U_1 \ U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 \Sigma_1 V_1^T.
\]
It can now be shown that the following equations hold:
\[
\xi = \mathcal{O}_k X_f, \quad (28)
\]
\[
\mathcal{O}_k = U_1 \Sigma_1^{1/2} T, \quad (29)
\]
\[
X_f = T^{-1} \Sigma_1^{1/2} V_1^T, \quad (30)
\]
where \( T \) is an arbitrary nonsingular matrix. Note that equation (30) gives a state vector estimate for the states \( \hat{X}_k = [x(k) \ x(k+1) \ x(k+N-1)] \). By defining the matrices
\[
\hat{X}_{k+1} = [x(k+1) \ \cdots x(k+N-1)] \\
\hat{X}_k = [x(k) \ \cdots x(k+N-2)] \\
\hat{U}_k = [u(k) \ \cdots x(k+N-2)] \\
\hat{Y}_k = [u(k) \ \cdots x(k+N-2)],
\]
the observed system can be written as
\[
\begin{bmatrix} \hat{X}_{k+1} \\ \hat{Y}_k \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \hat{X}_k \\ \hat{U}_k \end{bmatrix},
\]
which is a linear equation system. 21 By using the Least Squares method the system matrices can finally be estimated as
\[
\begin{bmatrix} \hat{A} & \hat{B} \\ \hat{C} & \hat{D} \end{bmatrix} = \left( \begin{bmatrix} \hat{X}_k \\ \hat{U}_k \end{bmatrix}^T \begin{bmatrix} \hat{X}_k \\ \hat{U}_k \end{bmatrix} \right)^{-1} \begin{bmatrix} \hat{X}_k \\ \hat{U}_k \end{bmatrix}^T \begin{bmatrix} \hat{X}_{k+1} \\ \hat{Y}_k \end{bmatrix}.
\]
\[21\text{Katayama (2005)}\]
3.4 Model Validation

After having obtained a model structure and determined its parameters, the next step is to evaluate whether the model can adequately describe the observations. It is important that the model estimation and the model validation is made on different interval of the time series, i.e. if there are $N$ measurements then the estimation should be done based on the measurements $1, ..., k$, and the validation based on the measurements $k+1, ..., N$, where $k$ is sufficiently smaller than $N$. Often $k = N/2$ is a good value.

Different criteria are used to measure the quality of the model, some of which are presented below.

**FIT** is the percentage of the measured output that was explained by the model, and is given by

$$FIT = 100 \left(1 - \frac{\| \hat{y} - y \|}{\| y - \bar{y} \|} \right),$$

where $\hat{y}$ is the predicted output, $y$ is the measured output and $\bar{y}$ denotes the mean of the measured output. $\| \cdot \|$ represents the 2-norm. 22

**VAF** or Variance Accounted For, is given by

$$VAR = 100 \left(1 - \frac{\text{Var}[e]}{\text{Var}[y]} \right),$$

where $e = y - \hat{y}$ is the residual. 23

3.5 Index Numbers

3.5.1 What are Index Numbers

An index number is a number that measures the relative change in a set of measurements over time. A certain value in the measurement series is chosen as base value and usually defined as 100. The index $i_k$ for any other measurement $y_k$ is then defined as

$$i_k = 100 \frac{y_k}{y_{k, \text{base}}},$$

where $k = \text{base}$ is the measurement value at the base value. For example, if sales of a commodity cost 1.5 times as much 1980 as it did 1970, the index number would be 150 relative to 1970. The advantage of indices is that it gives a way to compare values in a standardized manner instead of i.e. real sales figures. It should be noted, however, that changes in the index from year to year should not be considered as percentages except when one of the years is the base year. 24

---

22MATLAB Documentation
24Aczel (2002)
3.5.2 Index series from Statistics Sweden

In this thesis, conjuncture index series from the Statistics Sweden ("Statistiska Centralbyråns") will be analyzed in order to try to find correlations between these index series and sales figures for different gases at AGA. Statistics Sweden is an administrative agency whose main task is to support customers with statistics for decision making, debate and research. It supports and coordinates the Swedish system for official statistics and takes part in the international statistical cooperation.

The index series that will be used in this thesis are the New Orders in Industry (NOI) and the Industrial Production Index (IPI), which both are classified as conjunctural statistic series by Statistics Sweden. This means that the series holds information about the development in the total Swedish industry as well as for individual industrial segments concerning new orders (NOI) and industrial production (IPI). Index numbers are added monthly and are calculated through different kinds of surveys among the companies in the Swedish industry. 25

3.6 International Standard Industrial Classification (ISIC)

The International Standard Industrial Classification (ISIC) is a standard created by the UN for classification of economic activities based on a set of internationally agreed definitions and classification rules.26 The classification framework is vast and segmented in divisions, each division describing a certain industry. Division 58: "Publishing activities" is displayed in Table 1 as an illustrative example of the classification standard. In this thesis, the Class number in the ISIC framework for a specific customer will be referred to as simply its ISIC.

<table>
<thead>
<tr>
<th>Division</th>
<th>Group</th>
<th>Class</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Division 58</td>
<td></td>
<td>581</td>
<td>Publishing activities</td>
</tr>
<tr>
<td>5811</td>
<td></td>
<td></td>
<td>Publishing of books, periodicals and other publishing activities</td>
</tr>
<tr>
<td>5812</td>
<td></td>
<td></td>
<td>Book publishing</td>
</tr>
<tr>
<td>5813</td>
<td></td>
<td></td>
<td>Publishing of directories and mailing lists</td>
</tr>
<tr>
<td>5819</td>
<td></td>
<td></td>
<td>Publishing of newspapers, journals and periodicals</td>
</tr>
<tr>
<td>582</td>
<td></td>
<td>5820</td>
<td>Other publishing activities</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Software publishing</td>
</tr>
</tbody>
</table>

Table 1: Sample from the International Standard Industrial Classification

25 www.scb.se
26 ISIC, Rev 4
4 Preparation of data

In order to get a better understanding of the characteristics of the sales at AGA, some preparation of the two available datasets is needed. This preparation aims towards a mapping of what customers and what industry segments are the most prevailing among the analyzed gases. Especially important is the industry segment mapping as it will enable the next phase of the data preparation, which is the identification of conjunctural indices from Statistics Sweden corresponding to these segments. Recall that only dataset 2 is segmented into ISIC-codes and will therefore be the one used for the mapping. However, this dataset is too short for a reliable model estimation and for this purpose it will be merged with dataset 1. The outline of this chapter is thus, for each gas:

1. Group the customers in dataset 2 according to ISIC and analyze the sales for each ISIC separately
2. Identify dominating ISIC
3. Identify corresponding conjunctural indices from Statistics Sweden
4. Merge dataset 1 and dataset 2 and prepare the resulting series for modeling and correlation analysis

Some additional results concerning customer size and intermutual relations will also be declared. These results are not needed for the main task of this thesis, but could still be of interest for AGA Gas. Due to confidentiality agreements, the figures have been either omitted or scrambled (i.e. the numbers do not represent the real sales) in the official version of the thesis. It does however not affect the analysis as it naturally was performed on the real sales figures.

4.1 Segmentation

The main target in this section is to identify the most important ISIC-codes (i.e. have the biggest sales figures) for each of the three gases. The second task is to analyze the sales curves for the identified ISIC-codes in order to evaluate whether they can be separately modeled. To avoid repeating, every method or calculation in the following chapters are performed on all three gases if not stated otherwise.

4.1.1 Totals

To get an overall feeling about how the sales are varying over time, the sales per customer from dataset 2 are summarized per month. Figure 3 to Figure 4 shows the monthly sales during the period 2004-2010 for Acetylene, GOX and Shielding gases, respectively. One could notice the seasonal dips around July every year due to the industrial holiday, and the decrease of sales at the end of 2008 which is related to the global financial crisis.
After having obtained this understanding about the total sales, it might also be interesting to examine the ratio of some portion of the most important customers in relation to the total sales. In order to obtain this, the total volume per customer, i.e., the monthly volumes summarized, is first ordered according to intermutual size. Second, the total volume of a number of the most important customers is divided by the total volume for all customers. Set the total volume for customer $k$ to $vol_k$, where the volumes are ordered according to size so that $k = 1$ is either the biggest customer (descending order) or the smallest customer (ascending order). The volume for the biggest customer $k = 1$ to the total volume for all customers ratio is then calculated as

$$\frac{vol_1}{\sum_{k=1}^{n} vol_k},$$

where $n$ is the total number of customers. The sales of a portion of the 10 biggest customers $vol_1, vol_2, ..., vol_{10}$ to total sales ratio is then calculated as

$$\frac{\sum_{k=1}^{10} vol_k}{\sum_{k=1}^{n} vol_k}. \quad (31)$$

By sequentially adding one customer and performing the calculation in equation (31) and plotting the result, it is easy to see what portion of the customers is representing a significant part of the sales. This aims towards an answer to the question “how many of the biggest clients represent $x$ % of the sales”. This is calculated for both ascending and descending order (which of course are mirrors of each other) and the results are presented in figures 6 to 7.
From these sales to total sales ratio figures, the conclusion is made that during the period January 2004 to March 2010, quite a small portion of the total customers represent a large portion of the total sold volume.

4.1.2 ISICs

In order to find out which industry segments that are the most important for the sales, the monthly sales from figures 3 to 4 need to be segmented into ISIC-codes. Every customer in dataset 2 is pre-categorized to a specific ISIC, and the total monthly sales per ISIC can thus be calculated. The results are presented in figure 8 to 9.

Omitted

Figure 8: ISIC per month, Acetylene

Omitted

Figure 9: ISIC per month, GOX

Omitted

Figure 10: ISIC per month, Shield gases

At least two conclusions can be drawn from the Total sales per ISIC figures. First, as for the individual customers, there seem to be a smaller portion of the ISICs which are essential to the sales. However, the number of essential ISICs cannot be determined from these figures. Second, the answer to the question of whether or not the pattern for these ISICs differ significantly from the total sales is inconclusive. Judging from the figures there are reasons to suspect such a behavior, but further analysis will be needed.

Starting with the first conclusion, that a portion of the ISICs are essential to the sales, the next step is to find out how big this portion is. In the same way as the ratio to total sales calculations for the customers in figures 6 to 7, the ratios to total sales for the essential ISICs are calculated. The results are presented in figures 11 to 12.

Omitted

Figure 11: Sales to total sales ratio, ISIC

Omitted

Figure 12: Sales to total sales ratio, ISIC
According to these figures, the ratio to total sales for the ISICs are is similar to the one for the individual customers. A smaller portion of the ISICs represent a significant part of the sold volume. As the next step will be to study these essential ISICs individually, a suitable number of ISICs must be identified and picked out to this group of essential ISICs. From now on the group of essential ISICs will be referred to as the essential group and the remaining ISICs will be referred to as the remaining group. In order to decide the number of ISICs in the essential group, the ratios are studied more in detail, see Table (2). The main task of this section is to group essential ISICs in order to model them separately, which means that the essential group requires (i) a significant but not dominating representation of the total volume since the remaining group also should be significant and (ii) a sale pattern which differs sufficiently from the remaining group to be of interest for separate modeling. With (i) in mind and while keeping the number of ISICs reasonable low, 75% was chosen as an appropriate part of the total sales. Thus the suitable number of ISICs to be included in the essential groups are 22, 21 and 25 for Acetylene, GOX and Shielding gases, respectively.

<table>
<thead>
<tr>
<th>Table 2:</th>
<th>Ratio to Sales for grouped ISICs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetylene</td>
<td></td>
</tr>
<tr>
<td># of ISICs</td>
<td>% of Sales</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>22</td>
<td>75%</td>
</tr>
<tr>
<td>28</td>
<td>80%</td>
</tr>
<tr>
<td>40</td>
<td>90%</td>
</tr>
<tr>
<td>72</td>
<td>95%</td>
</tr>
<tr>
<td>333</td>
<td>100%</td>
</tr>
</tbody>
</table>

With the essential group chosen so that it represent a significant, but not dominating, part of the total sales, it is now possible to compare it with the remaining group by summarizing the monthly sales. The results are presented in figures 13 to 14. From these figures the conclusion is drawn that the sale pattern of the essential group dose not differ significantly from the sale pattern of the remaining group. This means that requirement (ii) presented above is not fulfilled, and therefore there is no gain in trying to model the essential group separately, as it will result in a model very similar to the one for the remaining group. Several different proportions in the numbers of ISICs in the essential and the remaining group has been tried out, all leading to the same conclusion.
However, the segmentation and identification of essential ISICs has not been done in vain, as it still enables the identification of suitable conjunctural indices from Statistics Sweden, which could improve the quality of a model for the total sales (i.e. no separate modeling for essential ISICs). The essential ISICs for each of the gases are presented in Table 9 in Appendix.

Omitted

Figure 13: ISIC grouped, Acetylene

Omitted

Figure 14: ISIC grouped, GOX

Omitted

Figure 15: ISIC grouped, Shield gases

4.2 Identification of conjunctural indices

Table 9 Appendix displays the essential ISICs for each of the three gases. Based on this information it is now possible to identify appropriate conjunctural indices from Statistics Sweden for the upcoming modeling and correlation analysis.

The available industrial classifications for the NOI and IPI conjunctural indices are given in Table 10 in Appendix. After having compared the essential ISICs for each of the gases with these conjunctural indices, a few interesting indices could be matched and selected. The indices in question are given in Table 3, but due to lack of space only the corresponding codes are displayed.
Although there are reasons to believe that these chosen indices are more interesting to analyze for correlation with the sales than the remaining indices, all available indices will be analyzed. As the analysis is done algorithmically on a computer, the task can be automated for an arbitrary number of indices and thus not representing unnecessary extra workload.

Before this thesis, some basic correlation analysis with conjunctural indices was performed at AGA Gas. This work suggested the use of a weighted index series for the correlation analysis, where the different industry segments in the conjunctural index were weighted into a main series according to the gas level of exposure toward each respective industry segment. This suggested series will be included in the modeling and correlation analysis, where it will be referred to as the weighted series.

### 4.3 Further preparation of the data

The time series used to identify the essential ISICs for each of the three gases, dataset 2, is relatively short. In fact it has not enough length to properly estimate the parameters for the sought models. For this purpose dataset 1 (ranging from 1995 to 2005) and dataset 2 (ranging from 2004 to 2010) will be merged. As they overlap with one year, it is possible to ensure that they are consistent, i.e the merge will be continuous in at the changing point. Sales during 2004 are examined for both series and they coincide, thus a simple appending of dataset 2 onto dataset 1 is possible.

As been seen earlier the provided datasets are raw sales, i.e they include seasonal behavior and occasional fluctuations. It is desirable to remove these characteristics in order to simplify the model estimation. First, the series is unseasoned ed by using seasonal indices provided by AGA, which has been

<table>
<thead>
<tr>
<th>Acetylene</th>
<th>GOX</th>
<th>Shielding gases</th>
</tr>
</thead>
<tbody>
<tr>
<td>21+22</td>
<td>21+22</td>
<td>13+14</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>21.1</td>
<td>21.1</td>
<td>D</td>
</tr>
<tr>
<td>21.11</td>
<td>21.11</td>
<td>27+28</td>
</tr>
<tr>
<td>21.12</td>
<td>21.12</td>
<td>27</td>
</tr>
<tr>
<td>21.2</td>
<td>23+24</td>
<td>27.1-27.3</td>
</tr>
<tr>
<td>27+28</td>
<td>23</td>
<td>27.4+27.5</td>
</tr>
<tr>
<td>27</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td>27.1-27.3</td>
<td>27+28</td>
<td>29</td>
</tr>
<tr>
<td>27.4+27.5</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>27.1-27.3</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>27.4+27.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>
calculated according to the Ratio-to-Moving-Average method. The resulting series are shown, together with the original unseasoned series, in figures 16 to 17.

Omitted

Figure 16: Unseasoned sales, Acetylene

Omitted

Figure 17: Unseasoned sales, GOX

Omitted

Figure 18: Unseasoned sales, Shielding gases

The unseasoned series will however still contain occasional fluctuations. To eliminate these, a 5-value moving average is computed. Further, the series show a highly irregular behavior after the end of 2008, which is related to the recent financial crisis. As a model cannot consider such irregularities, the months directly affected by the crisis will not be part of further analysis. This concerns all months after September 2008. The resulting series is shown together with the unseasoned series in figures 19 to 20.

Omitted

Figure 19: Moving average, Acetylene

Omitted

Figure 20: Moving average, GOX

Omitted

Figure 21: Moving average, Shielding gases

Last, the annual rate based on the moving average series is calculated for each of the three series. The annual rate $\Delta_{12,k}$ is the relative difference between the sales $s_k$ for a month $k$ and the sales for the same month one year back, $s_{k-12}$, i.e.

$$\Delta_{12,k} = \frac{s_k - s_{k-12}}{s_k}.$$  

The resulting series are the final ones which the sought models will try to describe. The series are plotted in figures 22 to 23.
As for the sales data, the index series need to be prepared for the model estimation. They do not need to be unseasoned, as it has already been done by Statistics Sweden, but a 3-point moving average and the annual rate is calculated for the same period as for the sales data. As this is done on quite a few index series and the procedure is the same as above only the final result, i.e. the annual rate, and only for one of the indices is presented here in figure. The presented index is the weighted index which is the only index with different characteristics for the different gases (as the weights differ).
Figure 27: Annual rate, Weighted index for Shielding gases
5 Analysis and Discussion

In this section the AR and State Space models will be estimated and validated. First, AR and State Space models are estimated without input, i.e. without “help” from the conjunctural indices. Only the annual rates of the respective gases are used for the estimation. After this the cross-correlations between the sales’ annual rates and the different conjunctural indices’ annual rates are analyzed. This is done in order to find suitable conjunctural indices which could be used as input signals and hopefully improve the quality of the models. Finally, these new models with correlating conjunctural indices used as input are being estimated and validated.

5.1 Models without conjunctural indices

Using the annual rates of the respective gas, the first versions of the AR and State Space models were estimated using the techniques described in Chapter 3. The series contain 150 values, one value per month, of which the first half were used for the model estimation. The remaining half was then used for validation.

5.1.1 The Models

The AR model

\[ Y_t + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} = e_t, \]

was estimated to order \( p = 4 \) for both Acetylene and Shielding gases and order \( p = 6 \) for GOX. The polynomial \( A(q) = 1 + a_1 q^{-1} + \ldots + a_p q^{-p} \), which fully describes an AR(p)-model, was estimated to

\[
\begin{align*}
A(q) &= 1 - 1.168q^{-1} - 0.05757q^{-2} + 0.5446q^{-3} - 0.2833q^{-4}, \\
A(q) &= 1 - 0.9031q^{-1} - 0.3766q^{-2} + 0.1474q^{-3} + 0.0165q^{-4}
\end{align*}
\]

and

\[
\begin{align*}
A(q) &= 1 - 1.191q^{-1} - 0.1202q^{-2} + 0.6585q^{-3} - 0.2791q^{-4} + \\
&\quad -0.1501q^{-5} + 0.16q^{-6},
\end{align*}
\]

for Acetylene, GOX and Shielding gases, respectively.

The state space model

\[
\begin{align*}
X_t &= AX_{t-1} + Ke_t \\
Y_t &= CX_t
\end{align*}
\]

was estimated to order \( m = 3 \) for all three gases. The estimates of the matrices \( A, C \) and \( K \) and the initial state vector \( X_0 \) of course differ between the three gases, and the result is presented below.
Acetylene:
\[
A = \begin{bmatrix}
0.94683 & 0.30744 & 0.17111 \\
-0.16382 & 0.011616 & 0.7307 \\
0.05526 & -0.38027 & 0.34559
\end{bmatrix}
\]
\[
C = \begin{bmatrix}
63.876 & 1.6141 & 5.9152
\end{bmatrix}
\]
\[
K = \begin{bmatrix}
0.017012 \\
0.023994 \\
-0.0030828
\end{bmatrix}
\]
\[
X_0 = \begin{bmatrix}
-0.82279 \\
0.10372 \\
0.17707
\end{bmatrix}
\]

Shielding gases:
\[
A = \begin{bmatrix}
0.84636 & -0.10631 & 0.44058 \\
0.18951 & -0.58626 & -0.66136 \\
-0.20613 & -0.16547 & -0.44004
\end{bmatrix}
\]
\[
C = \begin{bmatrix}
31.085 & -0.13512 & 0.59956
\end{bmatrix}
\]
\[
K = \begin{bmatrix}
0.022436 \\
0.0014154 \\
0.019577
\end{bmatrix}
\]
\[
X_0 = \begin{bmatrix}
-1.5098 \\
-0.18655
\end{bmatrix}
\]

and GOX:
\[
A = \begin{bmatrix}
0.94855 & -0.29102 & 0.11418 \\
0.18739 & -0.03084 & 0.2171 \\
-0.044073 & 0.15929 & 0.64704
\end{bmatrix}
\]
\[
C = \begin{bmatrix}
64.519 & -2.1523 & -0.80783
\end{bmatrix}
\]
\[
K = \begin{bmatrix}
0.01777 \\
-0.024614 \\
0.0052209
\end{bmatrix}
\]
\[
X_0 = \begin{bmatrix}
-0.78358 \\
-1.984
\end{bmatrix}
\]

5.1.2 Model validation
In order to validate the estimated models, one-step ahead prediction series were calculated using each of the models. Based on these estimated annual rates, measures of how good the model are could be obtained in terms of FIT and VAF. These measures are presented in Table (4) and the estimated annual rates
together with the measured (real) annual rates are plotted in figure 28 and 29 for both the AR(p) and the State Space model. As the results are quite similar for all three gases, only the plots for Acetylene are displayed.

\begin{table}[ht]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
 & \textit{State Space Model} & & \textit{AR Model} & \\
 & FIT & VAF & FIT & VAF \\
\hline
Acetylene & 62.6 & 85.9 & 64.2 & 86.0 \\
GOX & 57.1 & 81.3 & 55.9 & 80.3 \\
Shielding gases & 54.2 & 79.0 & 54.8 & 79.2 \\
\hline
\end{tabular}
\caption{Table 4:}
\end{table}

Omitted

Figure 28: One step prediction, AR(4) model

Omitted

Figure 29: One step prediction, State Space model

The one-step ahead predictions are very similar for the state space model and the AR(p) models, but not very accurate.

5.2 Correlation Analysis

In order to improve the quality of the models presented in the previous section, cross-correlating conjunctural index-series could be used as input signal to the models. If the cross-correlation is strong, this should affect the quality of the estimated models in a positive manner. Thus, the next step is to analyze the cross-correlations between the available conjunctural indices and the sales series for the three gases. The two available conjunctural indices, NOI and IPI, are both segmented into industrial classification and presented in Table (10) in Appendix. Also, the \textit{weighted index} will be analyzed for cross-correlation.

In section 4.2 the essential ISICs for each of the gases was matched with the industrial classifications of the NOI and IPI indices. A number of segments were identified as especially interesting for cross-correlation analysis. However, as stated before, all available indices will be tested for cross-correlation. As the number of tests exceeds 300, only a few cross-correlograms will be presented here (figures 30 and 31). The final result, i.e the identified conjunctural indices with cross-correlation coefficients over 0.4 for some time lag \( \tau \), is presented in Table 5. Table 6 presents whether the \textit{weighted series} is cross-correlated with the sales. It actually does cross-correlate, but some of the conjunctural indices show stronger correlation which can be seen in some of the figures below.

35
Concluding, there are a number of conjunctural indices which seem to be cross-correlated with the sales series. These will be used as input signals in the second model estimation, hopefully improving the estimated models.

Table 5: Found cross-correlations for NOI and IPI indices for the period 1996-2008

<table>
<thead>
<tr>
<th>Acetylene</th>
<th>GOX</th>
<th>Shielding gases</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>D+16</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>21.12</td>
<td>15+16</td>
</tr>
<tr>
<td>25</td>
<td>27</td>
<td>15</td>
</tr>
<tr>
<td>28</td>
<td>27.1-27.3</td>
<td>29</td>
</tr>
<tr>
<td>29</td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Acetylene</th>
<th>GOX</th>
<th>Shielding gases</th>
</tr>
</thead>
<tbody>
<tr>
<td>INV</td>
<td>INV</td>
<td>INV</td>
</tr>
<tr>
<td>26</td>
<td>C</td>
<td>13</td>
</tr>
<tr>
<td>28</td>
<td>13+14</td>
<td>13.2</td>
</tr>
<tr>
<td>29</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>13.1</td>
<td>29</td>
</tr>
<tr>
<td>28</td>
<td>13.2</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>24exkl24.4</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 6: Cross-correlation weighted index

<table>
<thead>
<tr>
<th>The Weighted Index</th>
<th>Acetylene</th>
<th>Shielding gases</th>
<th>GOX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

The cross-correlograms for the NOI index against Acetylene and the weighted index also against Acetylene is plotted in Figure 6. It is the positive x-axis (positive lag $\tau$) which is of main interest. Examining the left cross-correlogram in figure 6, one should conclude that the cross-correlation between the NOI (28) series and Acetylene is almost 0.6 at time lags $\tau = 2$ and $\tau = 3$, after which the correlation is decreasing. The blue horizontal lines indicate upper and lower bounds (confidence interval), within which the series are with 95% probability uncorrelated. As an example, for the left plot in Figure 30, it is safe to say that the series are uncorrelated for lags $\tau = 11$ to $\tau = 20$ (when still only considering the positive x-axis).
5.3 Models with conjunctural indices

The identified conjunctural indices and the weighted index will in this section be used as input to the estimated models in order to improve their accuracy in describing the annual rates of the sales. As there now is an input to the model, the AR model is replaced by the ARX model, which allows for an exogenous input.
5.3.1 Models

The conjunctural indices are used as input $u_t$ in the ARX model

$$y_t = -a_1 y_{t-1} - \ldots - a_{t-p} + b_1 u_{t-1} + \ldots b_p u_{t-p} + \epsilon_t,$$

and $u_t$ in the State Space model

$$X_t = AX_{t-1} + Bu_{t-1} + K\epsilon_t$$

$$Y_t = CX_{t-1} + \epsilon_t,$$

As a large amount of models were estimated, one for each conjunctural index with cross-correlation to the sales, the estimated parameters are not being presented here. The same estimation techniques was used as for the first models, i.e the least squares for the ARX model and the subspace identification for the State Space model. Orders vary in the range 3 – 8 for the ARX models and 2 – 4 for the State Space models.

5.3.2 Model Validation

Each of the created models was tested for quality with the same criteria as for the models without the conjunctural indices used as input. The results are presented in Tables 12 and 13 in Appendix, with one model per row for each of the gases. The indices for which the quality of the model was improved is marked as bold. From these results, one could easily tell that the improvements of the models unfortunately are not what could have been expected with such strong cross-correlations between the input signals and the measured values. Something in the model estimation has not been accounted for. Assuming that the model estimation algorithms are correct, one could suspect something not being right in the cross-correlations.

5.4 Discussion

The models with conjunctural indices as input were estimated using the first half of the sales and index series respectively, i.e the period 1996-2001. As each model then were validated against the second half, i.e the period 2002-2008, it is crucial for the model to be accurate that the character of the cross-correlation with the conjunctural index is similar for both periods. If not, one cannot expect the model to describe the validation data in a satisfying manner, as it would appear as the model was created for another data series. In this section the cross-correlations for the sales data with the different conjunctural indices will be analyzed during the first half and the second half of the data separately. The analysis is analogous to the one in section 5.2 and the result is presented in tables 14 and 13 in Appendix, where the displayed indices have a cross-correlation coefficient $> 0.5$ for at least one lag $\tau$. In addition, the indices that show correlation for both the first half and the second half of the series are marked in bold.
The results are interesting and explains the poor improvements of the estimated models when the indices are used as inputs. Judging from the tables, very few of the conjunctural indices that have significant cross-correlation with the sales during the first period also shows this behavior during the second half of the series. This means that while the cross-correlations might be strong in the first half and an estimated model would be excellent in describing a second half with the same cross-correlation, the model might be even worse off using the conjunctural index if the cross-correlation is weak or non-existent in the second half.

However, there seem to be a few indices which are cross-correlated both during the first half and the second half. They are listed in Table 7. As the models using these indices do not show significantly more improvements than the others, the cross-correlations need to be investigated even more in detail. It should be noted that not all displayed indices in Table 7 were included in the first cross-correlation results in Table 5, which means that the “extra” indices displayed here did not show any cross-correlation when analyzed over the whole period.

Also the weighted index is re-analyzed for cross-correlation for the two periods separately. The results are presented in Table 8, and the conclusion drawn is that the weighted index only shows thoroughgoing cross-correlation with Shielding gases.

In order to investigate these unclarities, the corresponding cross-correlograms were analyzed for each of the periods separately. Not all plots are shown here, but the cross-correlograms for Shielding gases with IPI (26) and Shielding gases with the weighted index are presented in Figure 32, where the two periods are plotted on top of each other to facilitate the comparison. The plots confirm the cross-correlation for the sales and the indices for both periods, but reveals that the characteristics of the cross-correlations differ between the first and the second period. This explains why not even these indices could improve the model, despite of the seemingly thoroughgoing cross-correlation. The plots for the other conjunctural indices differ in a similar manner and are therefore not shown here.

### Table 7: Thoroughgoing cross-correlations

<table>
<thead>
<tr>
<th>Gas</th>
<th>NOI</th>
<th>Gas</th>
<th>IPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shielding gases</td>
<td>27+28</td>
<td>Acetylene</td>
<td>28</td>
</tr>
<tr>
<td>GOX</td>
<td>INSexkIERV</td>
<td>Shielding gases</td>
<td>26</td>
</tr>
<tr>
<td>GOX</td>
<td>27+28</td>
<td>Shielding gases</td>
<td>28</td>
</tr>
<tr>
<td>GOX</td>
<td>27</td>
<td>GOX</td>
<td>INSexkIERV</td>
</tr>
<tr>
<td>GOX</td>
<td>27.1-27.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8: Cross-correlations, Weighted index

<table>
<thead>
<tr>
<th>Gas</th>
<th>First half</th>
<th>Second half</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetylene</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Shielding gases</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>GOX</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Figure 32: Cross-correlations for different periods
6 Conclusions

The largest part of the sales for the three dominating gases Acetylene, GOX and Shielding gases at AGA Gas are represented by a relatively small number of industrial segments. For all three gases, 75% of the sales are represented by in between 21-25 industrial segments from a total of almost 350 segments. By analyzing the sales of these segments separately it can be concluded that the sales pattern does not differ significantly from the remaining industrial segment. Thus, there is no advantage in trying to model these segments separately.

Cross-correlation analysis over the whole available period, 1996-2008, shows strong correlation between several of the conjunctural indices and the sales, and are therefore used as input signals to the models in order to improve their quality. The ARX and State Space models, estimated with data from the period 1996-2001, did however not show significant improvements from the correspondent models without indices as input when validated on the period 2002-2008. The one-step ahead predictions of the sales were even inferior for some of the models. More detailed cross-correlation analysis revealed that the original analysis had not taken into account that the correlations between the chosen indices and the sales varied over time. By partitioning the datasets in two halves, i.e the estimation period 1996-2001 and the validation period 2002-2008, cross-correlation analysis could be performed on the two halves separately. The conclusion from this analysis is that most of the conjunctural indices which had shown correlation with the sales over the whole period 1996-2008, had weak or non-existing correlation with the sales during either the first or the second half of the period. A few indices did correlate over both periods, but with significantly different characteristics in the correlation. As it is crucial for any model to work correctly that the data used for estimation is similar to the data where the model is supposed to operate, it should be considered normal that the models in this thesis did not show substantial improvements.

Concerning the weighted index, it could not be proved that it would be a better choice than straight-forward using the conjunctural indices as some indices had better correlation to the sales than the weighted index. Although, the idea to weight the indices is probably a good idea but the weights need to be revised.

As no significant thoroughgoing cross-correlation could be established between the sales and the conjunctural indices over the period 1996-2008, it is thus, at this time, not advantageous to use conjunctural indices from Statistics Sweden to improve models describing the sales at AGA Gas. However, cross-correlation have been proven with various conjunctural indices during sub-periods. This means that if the correlations are continuously monitored there is a good possibility of finding indices that currently (i.e over the past 4-5 years) cross-correlate with the sales which could be used to help estimate models for the sales. As the financial climate have been quite turbulent during the last couple of years, this could probably be difficult to do at this time. But it should definitely be worth trying in the longer perspective. It then is important to bear in mind that the cross-correlations seem to vary between different conjunctural indices over time and that it therefore is important to have models which allows for
dynamic selection of which indices to use in the estimators.

Regarding what kind of model to use, there seem to be small quality difference between the ARX and the State Space models. Due to the complexity of the State Space model, the recommendation would be to use the ARX model as it implicates a relatively simple least squares estimation.
7 References

References


## Appendix

### 8.1 Tables

<table>
<thead>
<tr>
<th>Acetylene</th>
<th>GOX</th>
<th>Shielding gases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building installations</td>
<td>Metal waste &amp; scrap</td>
<td>Treat/coating metals</td>
</tr>
<tr>
<td>Other business act.</td>
<td>Treat/coating metals</td>
<td>Manufacture of other fabr. metal prod.</td>
</tr>
<tr>
<td>Build, rep. of ships</td>
<td>Other business activities</td>
<td>Other business activities</td>
</tr>
<tr>
<td>Basic iron &amp; steel</td>
<td>Manufacture of other fabr. metal prod.</td>
<td>Structural metal products</td>
</tr>
<tr>
<td>Treat/coating metals</td>
<td>Basic iron &amp; steel</td>
<td>Lifting &amp; handling equipment</td>
</tr>
<tr>
<td>Manufacture of other fabr. metal prod.</td>
<td>Structural metal products</td>
<td>Other general purpose machines</td>
</tr>
<tr>
<td>Motor vehicles</td>
<td>Building installation</td>
<td>Other special purpose machines</td>
</tr>
<tr>
<td>Structural metal products</td>
<td>Site preparation</td>
<td>Parts/access motor vehicles</td>
</tr>
<tr>
<td>Other special purpose machines</td>
<td>Construction/Parts thereof</td>
<td>Building installations</td>
</tr>
<tr>
<td>Construction/Parts thereof</td>
<td>Railway, Tram-locos</td>
<td>Basic iron &amp; steel</td>
</tr>
<tr>
<td>Other general purpose machines</td>
<td>Construction</td>
<td>Motor vehicles</td>
</tr>
<tr>
<td>Precious, non-ferr metals</td>
<td>Other general purpose machines</td>
<td>Construction/parts thereof</td>
</tr>
<tr>
<td>Site preparation</td>
<td>Recycling</td>
<td>Adult &amp; Other Education</td>
</tr>
<tr>
<td>Construction</td>
<td>Lifting &amp; handling equipment</td>
<td>Machinery &amp; Equipment</td>
</tr>
<tr>
<td>Adult &amp; Other Education</td>
<td>Other special purpose machines</td>
<td>Agriculture, forest machines</td>
</tr>
<tr>
<td>Pulp, paper, paperboard</td>
<td>Pulp, paper, paperboard</td>
<td>Machines mining</td>
</tr>
<tr>
<td>Maintenance/repairing motor vehicles</td>
<td>Manufacture building stone</td>
<td>Construction</td>
</tr>
<tr>
<td>Railway, Tram-locos</td>
<td>Parts/access motor vehicles</td>
<td>Maintenance/Repairation motor vehicles</td>
</tr>
<tr>
<td>Mining of iron ores</td>
<td>Machines mining</td>
<td>Production of electricity</td>
</tr>
<tr>
<td>Freight transport by road</td>
<td>Glass &amp; glass production</td>
<td>Manufacture of furniture</td>
</tr>
<tr>
<td>General Public Service</td>
<td>Adult &amp; Other Education</td>
<td>Build, rep. of ships</td>
</tr>
</tbody>
</table>
Table 10: Available industrial segments for the conjunctural indices *New Orders in Industry* and *Industry Production Index*

<table>
<thead>
<tr>
<th>Code with industrial classification</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-E</td>
<td>Mining, quarrying, manufacturing and energy</td>
</tr>
<tr>
<td>ER VE</td>
<td>Energy-related goods, including E</td>
</tr>
<tr>
<td>C+D</td>
<td>Mining, quarrying and manufacturing</td>
</tr>
<tr>
<td>INS exkl ER V</td>
<td>Intermediate goods industries</td>
</tr>
<tr>
<td>ER V</td>
<td>Energy-related goods excluding E</td>
</tr>
<tr>
<td>INV</td>
<td>Capital goods industry</td>
</tr>
<tr>
<td>IVKON</td>
<td>Non-durable consumer goods industry</td>
</tr>
<tr>
<td>VKON</td>
<td>Durable consumer goods industry</td>
</tr>
<tr>
<td>C</td>
<td>Mines and quarries</td>
</tr>
<tr>
<td>10-12</td>
<td>Mining and quarrying of energy producing materials</td>
</tr>
<tr>
<td>10</td>
<td>Coal mines; peat industry</td>
</tr>
<tr>
<td>13+14</td>
<td>Metal ore mines and other mines and quarries</td>
</tr>
<tr>
<td>13</td>
<td>Metal ore mines</td>
</tr>
<tr>
<td>13.1</td>
<td>Iron ore mines</td>
</tr>
<tr>
<td>13.2</td>
<td>Other metal ore mines</td>
</tr>
<tr>
<td>14</td>
<td>Other mines and quarries</td>
</tr>
<tr>
<td>D</td>
<td>Manufacturing industry</td>
</tr>
<tr>
<td>15+16</td>
<td>Food product, beverage and tobacco industry</td>
</tr>
<tr>
<td>15</td>
<td>Food product and beverage industry</td>
</tr>
<tr>
<td>17-19</td>
<td>Textile industry, industry for wearing and Tanners</td>
</tr>
<tr>
<td>20</td>
<td>Industry for wood and products of wood, cork, cane etc., except furniture</td>
</tr>
<tr>
<td>20.1</td>
<td>Saw-mills and planing-mills; wood impregnation plants</td>
</tr>
<tr>
<td>20.2-20.5</td>
<td>Industry for wood, products of wood, except furniture</td>
</tr>
<tr>
<td>21+22</td>
<td>Industry for Pulp and paper and publishers and printers</td>
</tr>
<tr>
<td>21</td>
<td>Industry for pulp, paper and paper products</td>
</tr>
<tr>
<td>21.1</td>
<td>Industry for pulp, paper and paperboard</td>
</tr>
<tr>
<td>21.11</td>
<td>Industry for pulp</td>
</tr>
<tr>
<td>21.12</td>
<td>Industry for paper and paperboard</td>
</tr>
<tr>
<td>21.2</td>
<td>Industry for articles of paper and paperboard</td>
</tr>
<tr>
<td>22</td>
<td>Publishers and printers; other industry for recorded media</td>
</tr>
<tr>
<td>23-24</td>
<td>Industry for coke, refined petroleum products and nuclear fuel and Industry for chemicals</td>
</tr>
<tr>
<td>23-24</td>
<td>Industry for coke, refined petroleum products and nuclear fuel and Industry for chemicals</td>
</tr>
<tr>
<td>24</td>
<td>Industry for chemicals and chemical products</td>
</tr>
<tr>
<td>24.4</td>
<td>Industry for pharmaceuticals, medicinal chemicals and botanical products</td>
</tr>
<tr>
<td>24 exkl 24.4</td>
<td>Manufacture of chemicals and chemical products except pharmaceuticals, medicinal chemicals</td>
</tr>
<tr>
<td>24 exkl 24.4</td>
<td>Manufacture of chemicals and chemical products except pharmaceuticals, medicinal chemicals</td>
</tr>
<tr>
<td>25</td>
<td>Industry for rubber and plastic products</td>
</tr>
<tr>
<td>26</td>
<td>Industry for other non-metallic mineral products</td>
</tr>
<tr>
<td>27</td>
<td>Industry for basic metals</td>
</tr>
<tr>
<td>27.1-27.5</td>
<td>Manufacture of basic precious and non-ferrous metals and casting of metals</td>
</tr>
<tr>
<td>27+28</td>
<td>Industry for basic metals and Industry for fabricated metal products, except machinery and equipment</td>
</tr>
<tr>
<td>28</td>
<td>Industry for fabricated metal products, except machinery and equipment</td>
</tr>
<tr>
<td>29</td>
<td>Industry for machinery and equipment n.e.c.</td>
</tr>
<tr>
<td>30-33</td>
<td>Industry for electrical machinery and apparatus n.e.c. and optical instruments</td>
</tr>
<tr>
<td>31</td>
<td>Industry for electrical machinery and apparatus n.e.c.</td>
</tr>
<tr>
<td>32</td>
<td>Industry for radio, television and communication equipment and apparatus</td>
</tr>
<tr>
<td>33</td>
<td>Industry for medical, precision and optical instruments, watches and clocks</td>
</tr>
<tr>
<td>34-35</td>
<td>Industry for transport equipment</td>
</tr>
</tbody>
</table>
Table 12: FIT and VAF for the NOI index models, 1996-2008

**New Orders in Industry (NOI)**

<table>
<thead>
<tr>
<th></th>
<th>Acetylene</th>
<th>GOX</th>
<th>Shielding gases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The State Space Models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Input</td>
<td>FIT</td>
<td>VAF</td>
</tr>
<tr>
<td>Acetylene C+D</td>
<td>61.5</td>
<td>85.3</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>61.6</td>
<td>85.3</td>
<td></td>
</tr>
<tr>
<td>15+16</td>
<td>59.6</td>
<td>83.7</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>59.7</td>
<td>83.8</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>60.8</td>
<td>84.7</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>59.4</td>
<td>83.5</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>56.4</td>
<td>82.7</td>
<td></td>
</tr>
<tr>
<td><strong>The ARX Models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Input</td>
<td>FIT</td>
<td>VAF</td>
</tr>
<tr>
<td>Acetylene C+D</td>
<td>63.3</td>
<td>86.8</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>63.2</td>
<td>86.5</td>
<td></td>
</tr>
<tr>
<td>15+16</td>
<td>62.0</td>
<td>85.6</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>63.6</td>
<td>85.6</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>62.5</td>
<td>86.8</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>59.0</td>
<td>86.9</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>59.8</td>
<td>84.1</td>
<td></td>
</tr>
</tbody>
</table>
Table 13: FIT and VAF for the IPI index models, 1996-2008

<table>
<thead>
<tr>
<th></th>
<th>Acetylene</th>
<th></th>
<th>GOX</th>
<th></th>
<th>Shielding gases</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Input</td>
<td>FIT</td>
<td>VAF</td>
<td>Input</td>
<td>FIT</td>
<td>VAF</td>
</tr>
<tr>
<td>IPI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acetylene</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acetylene</td>
<td>25</td>
<td>62.0</td>
<td>85.7</td>
<td>Inv.</td>
<td>55.7</td>
<td>80.4</td>
</tr>
<tr>
<td>Acetylene</td>
<td>28</td>
<td>58.4</td>
<td>82.8</td>
<td>Inv.</td>
<td>56.2</td>
<td>81.0</td>
</tr>
<tr>
<td>Acetylene</td>
<td>29</td>
<td>50.9</td>
<td>82.9</td>
<td>13+14Inv.</td>
<td>53.6</td>
<td>79.9</td>
</tr>
<tr>
<td>Acetylene</td>
<td>24 exkl24.4</td>
<td></td>
<td>42.1</td>
<td>72.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOX</td>
<td>54.1</td>
<td>79.3</td>
<td>55.7</td>
<td>80.0</td>
<td>13</td>
<td>54.1</td>
</tr>
<tr>
<td>Shielding gases</td>
<td>13</td>
<td>54.1</td>
<td>79.3</td>
<td>56.2</td>
<td>81.0</td>
<td>Inc.</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>56.8</td>
<td>81.5</td>
<td>Inc.</td>
<td>56.8</td>
<td>81.5</td>
</tr>
<tr>
<td></td>
<td>24 exkl24.4</td>
<td></td>
<td>32.7</td>
<td>72.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 14: Cross-correlations for the NOI index models

<table>
<thead>
<tr>
<th>Acetylene</th>
<th>GOX</th>
<th>Shielding gases</th>
</tr>
</thead>
<tbody>
<tr>
<td>First half</td>
<td>Second half</td>
<td>First half</td>
</tr>
<tr>
<td>INSxklERV</td>
<td>15+16</td>
<td>C+D</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>21.1</td>
<td>28</td>
<td>21+22</td>
</tr>
<tr>
<td>21.12</td>
<td>30-33</td>
<td>21</td>
</tr>
<tr>
<td>27+28</td>
<td>30</td>
<td>21.1</td>
</tr>
<tr>
<td>27</td>
<td>31</td>
<td>27+28</td>
</tr>
<tr>
<td>27.1-27.3</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>27.1-27.3</td>
<td>27.1-27.3</td>
</tr>
</tbody>
</table>

Table 15: Cross-correlations for the IPI index models

<table>
<thead>
<tr>
<th>Acetylene</th>
<th>GOX</th>
<th>Shielding gases</th>
</tr>
</thead>
<tbody>
<tr>
<td>First half</td>
<td>Second half</td>
<td>First half</td>
</tr>
<tr>
<td>C-E</td>
<td>15+16</td>
<td>INSxklERV</td>
</tr>
<tr>
<td>INSxklERV</td>
<td>14</td>
<td>ERV</td>
</tr>
<tr>
<td>ERV</td>
<td>24exkl24.4</td>
<td>10-12</td>
</tr>
<tr>
<td>13</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>13.1</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>D</td>
<td>28</td>
<td>13.1</td>
</tr>
<tr>
<td>20.1</td>
<td>30-33</td>
<td>20</td>
</tr>
<tr>
<td>21.1</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>21.11</td>
<td>21.1</td>
<td></td>
</tr>
<tr>
<td>21.12</td>
<td>21.11</td>
<td>28</td>
</tr>
<tr>
<td>23</td>
<td>21.12</td>
<td>29</td>
</tr>
<tr>
<td>27</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>27.4+27.5</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>