# Forecast of mail volumes. Predicting daily mail volumes for sorting centers

Ma, J.

Mathematical Institute

University of Leiden

THESIS

# Forecast of Mail volumes

## Predicting daily mail volumes for sorting centers

by

Jinxia Ma

Submitted in partial fulfillment of the requirements for the degree of Master in Science (Mathematics).

Thesis advisors:

Prof. dr S.A. van de Geer
dr E.W. van Zwet

date: 30 September 2005

# Table of contents

# 1 Introduction

TNT B.V. is an international company with over 163,000 employees, whose network covers 200 countries. It has three divisions: Mail, Express, and Logistics. The Mail Division, which TPG POST is part of, is a very important operator in the world and accounts for more than 70% of the total revenue. (See figure 1)

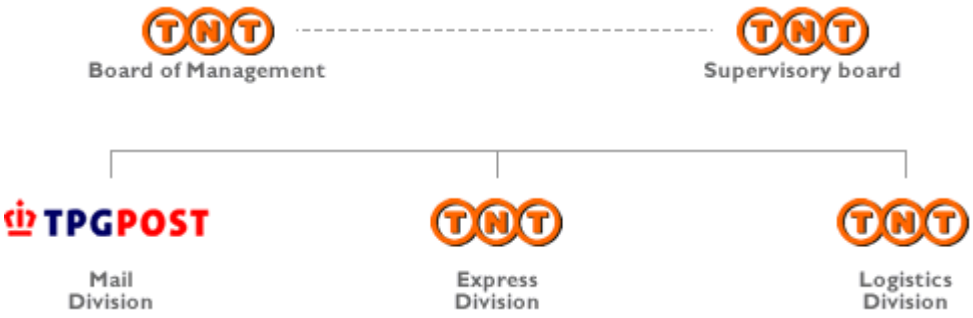



Figure 1: Structure of TNT B.V.

Since the Mail Division of TNT B.V. is very important among the three divisions of the company, making the Mail Division operate efficiently is significant to TNT. The mailing system contains three processes: collection, sorting, and distribution. This project is focused on the sorting part. A prediction of mail volume would help the planning of workers, and ensures that the sorting to operate more efficiently with relatively low cost.

The first chapter of this report contains general information about TPG POST, then in the second chapter, the mailing system and sorting process will be introduced. After we give an introduction to the project, which includes its goal, data collection, and data exploration. Then the methods will be introduced mathematically in the fourth chapter. In chapter five the estimations are made and results are shown. How to predict future will be introduced in chapter six and the conclusions are in chapter seven.

# 2 Mailing system and sorting process

## 2.1 Mailing system and sorting centres

The mailing system is a system of collection, sorting, and distribution. Every day, tens of millions of mail is collected at the mail boxes, post offices, post agencies, service points, and business counters. Then the mail is sent to the sorting centres to be sorted. After being sorted, the mail is distributed to delivery offices and from there to the addresses where they should go.

The sorting is done at sorting centres. There are eight sorting centres in The Netherlands totally, two of which deal with international mail and registered mail. Since this project concentrates on the national mail, it is enough to focus on the other six sorting centres: Amsterdam, Nieuwegein, Rotterdam, s-Hertogenbosch, s-Gravenhage, and Zwolle. (See Figure 2.1)


Figure 2.1: Sorting Centres

## 2.2 Sorting process

After the mail is collected, it is sent to the “nearest” sorting centre. (Here the word

“nearest” is between quotation marks because it doesn’t mean nearest in the sense of geography. In fact TPG POST divides The Netherlands into six areas, one sorting centre at each area. And all the mail in the area is sent to the sorting centre in the charge of this area. ) Then it gets the 1st sort, which means that it is sorted according to the first 4 digits of the postal code. The output is called “semi-sorted products”. Then all the sorting centres exchange the semi-sorted products between each other, so that the mail enters the areas where it should go. At the 2nd sorting centres, the mail gets the 2nd sort, this time according to all digits of the postal code, and the output is the end-products. Through the two sortings, the mail is sorted for individual postmen’s walks. (See Figure 2.2) Since all the mail that goes through the 2nd sorting also goes through the 1st, we only need to focus on the 1st sorting. Next paragraph will show how the mail is sorted for the 1st time at the sorting centres.

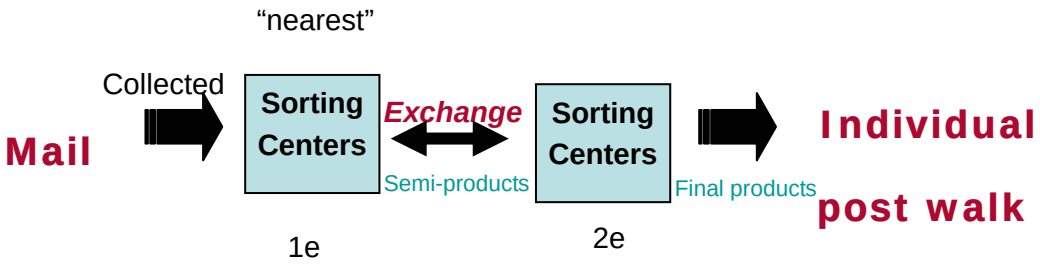



Figure 2.2: Sorting process

When the mail arrives at sorting centres, it is first put into SOSMA, which puts all the mail properly. It means that all the mail is put face-up with all stamps at the right-up corners. After the SOSMA, the mail will be distributed to different sorting machines according to size and weight. Generally, the SMK is used to sort small mail, SMG for big mail, SMB for trays of mail, and SMO for other mail. All the mail that can not be sorted by machines will be sorted by hand. The sorting machines can read the postal codes on the mail and change the mail into semi-sorted products, which are put into containers and exchange with other sorting centres. (See Figure 2.3)

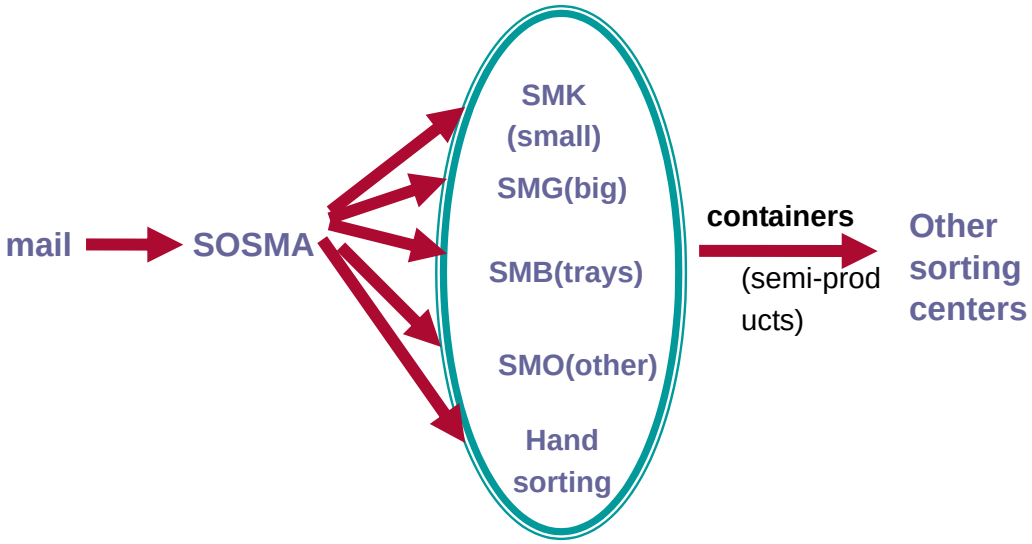



Figure 2.3: The 1st sorting process at sorting centres

TPG POST provides two services: 24-hour service and 48-hour service. Figure 2.4 shows that the 24-hour mail is sorted on the same day as it is collected, while the 48-hour mail is sorted one day later than the day it is collected. This should be taken into account with data-exploration.

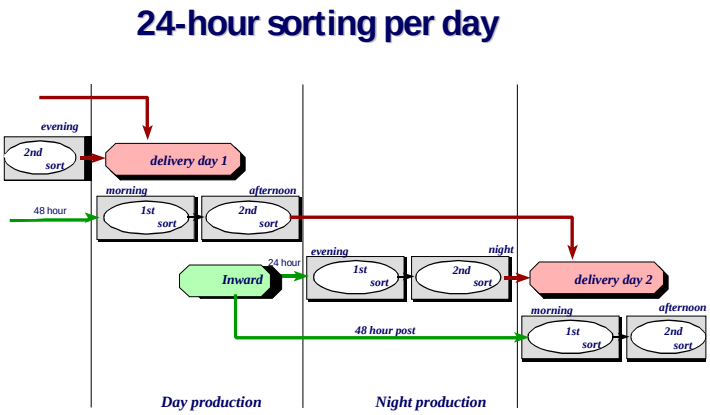



Figure 2.4: 24-hour service and 48-hour service

# 3 Introduction to project

The goal of this project is to make predictions of the mail volume per sorting centre per day.

The first question is why do we need to forecast? The reason is that different sorting machines require different numbers of workers (Generally speaking, the hand sorting needs the most and the SMK needs the least.) So the Department Sorting of TPG POST needs an indication of the different mail volumes per day, so that they can plan the workers.

## 3.1 Plan of project

The project will be focused on mail volumes per day. The data of SMK(sorting of small mail) in MIS are used for modelling and predictions.

It starts with data collection and data exploration. Since there are several kinds of data available, the most suitable data will be chosen after the data exploration. Also the data exploration will give some main ideas of estimations and predictions. Then further analysis will be made on a mathematical basis and estimation methods are provided. Lastly the estimation methods will be realized through programming and predictions will be made.

## 3.2 Data collection

The data mainly come from three sources: MIS, INDOOR, and L.V.R..

(1) MIS (Management Information System)
When the mail goes through the sorting machines, the machines count the mail volume. This is how the MIS data come out.
(2) INDOOR
This is the billing system for customers with groups of mail. For example, a bank sends the account balances to its customers every two weeks. Then the bank has a long-term contract with TPG POST to send mail regularly. INDOOR contains all

the bills sent to the customers for the groups of mail. (So figures of INDOOR are about all the groups of mail collected by TPG POST)

(3) L.V.R.

This estimates the total mail volume, both hand sorted and machine sorted mail. The estimates are partly based on sampling.

In this project, MIS and INDOOR are analysed and MIS SMK (sorting of small mail) is used for the modelling. The following (Figure 3.1) are the conditions for making queries when doing Data Collection. Notice that only SMK is considered because this data is the most reliable and it is one of the biggest mail streams.

| MIS (SMK ) | INDOOR (SMK ) |
|---|---|
| • 1st sort<br>• net-sorted | • Weight <= 50 g<br>• SV = 7<br>• Sort depth = 1 or 3 |

Figure 3.1: Query conditions when doing data collection

Here are some explanations of Figure 3.1. During the query of MIS, the "1st sort" and "net-sorted" are used. The reason that why use "1st sort" has been explained in the section "Sorting Process". The "net-sorted" differentiates itself from "gross-sorted". In fact, every time the mail goes through the sorting machines, some mail is rejected by the machines. Then the rejected mail is put into the machines again. Still some mail is rejected. The rejected mail now will be sorted by hand. This "rejection" course brings the difference between "net-sorted" and "gross-sorted". This is explained by an example. If 100 pieces of mail are put into a sorting machine and 10 are rejected, then the 10 pieces are put into the machine again. This time 4 are rejected. Altogether 110 pieces of mail are put into the machine, and 96 pieces go through successfully. The number "110" is called "gross-sorted", and "96" is "net-sorted".

The "net-sorted" data is used instead of "gross-sorted" because it is more similar to the real number of mail pieces.

As for the query conditions of INDOOR, the "sort depth" is set to "1" or "3". It is

because the companies can sort the mail by themselves before they send the mail to TPG POST. Through setting “sort depth”, we can get the mail that is sorted by TPG. And through setting “Weight <=50 g”, it ensures that the mail chosen all goes through the machines SMK. In the condition “SV=7”, the “SV” means the mail in groups. Then all the mail that is delivered at the post offices as a group (from about 100 to over 1,000,000) are booked under “SV=7”.

After the Data Collection, the next job is to do Data Exploration.

## 3.3 Data exploration

### 3.3.1 Analysis of INDOOR

Before starting the work of data exploration, it is necessary to know about the composition of mail. Where does the mail come from? What kind of mail accounts for the biggest proportion? In fact, the mail comes from two sources: business and individual. The business mail is from customers, especially long-term customers of TPG POST, and it always arrives regularly or is told to TPG POST in advance. So this part of mail is not the focus of our predictions, because the predictions only care the unknown of future, that is, the individual mail that come randomly.

Then how to get the individual mail from our data? The idea is to use MIS minus INDOOR. the reason is that “INDOOR” is the bills for customers, so it stands for the “business part”, while the “MIS”, as introduced above, is the total mail. So the difference between total mail and business mail is just the individual mail, i.e., “*MIS – INDOOR*”.

Exploration of data 2004

In order to see whether it is appropriate to use the idea of MIS-INDOOR, the data of MIS and INDOOR for 2004 are firstly explored.

Both the data of MIS and INDOOR reveal a decline in the period August and September, because during holidays there is relatively little mail to be sorted. December is a very special month, because there are always a lot of greeting cards

and other mail during the Christmas period. This phenomenon will continue to the beginning of next January, which is called in TPG POST as “KNJ”-period (Kerstmis en Nieuw Jaar, which means “Christmas and New Year”). So December is excluded from this analysis, and TPG POST has a special model for it.

If analyse the data of MIS and INDOOR together then something strange happens.

As is known, INDOOR is only mail from companies, so its values should be less than the MIS. But the data shows that values of INDOOR is much bigger than MIS at some days. While in the weekly data the INDOOR is normally less than the MIS. How can this be explained?

The reason is that the INDOOR volume can be for a whole month or week; in fact, the mail is sorted and delivered throughout the month or week, not just on the day when it is recorded into INDOOR. This is not the case for all records, only for large companies that send mail on a regular basis. So now the difficulty is to find out how the INDOOR mail is delivered, especially for the big customers. And since TPG POST has a lot of customers, it probably is impossible to look at every customer’s mail delivery separately. So, this idea about modelling MIS-INDOOR is difficult to be realized, and might not give a prediction of the actual daily mail volume.

### 3.3.2 Analysis of MIS

According to the data available, the focus is finally put on the “net-sorted” MIS data. There are three years of data (2002, 2003, and 2004). The daily mail volumes of the 3 years show an obvious decline in July and August, which accounts for the summer holidays. There are four holidays not on Saturday or Sunday, i.e., Eastern Monday, Queen’s Day, Ascension Day, and Pentecost Monday, which have much fewer mail volumes than normal days.

Another feature that can be noticed from the data is that it reveals a rule for the weekdays, e.g., a weekday is always a relatively busy day and another weekday always has a relatively small mail volume. The people of Sorting Department also certify this.

The analysis of MIS SMK reveals that it is the better data to be used for the predictions. The next part is the mathematical introduction of modelling methods.

# 4 Mathematical introduction of methods

In this part, some estimation methods for the data of net-sorted SMK of MIS will be summarized on a mathematical basis. The main idea is to separate the mail volumes into two parts, one is a stable generalized linear trend, and the other is the seasonal effects, which will describe the decline during summer holidays. A model with these two elements will firstly be introduced, and then estimation methods will be presented: Naïve method, Kernel method, Projection estimators, and Penalized least squares. After the mathematical introductions of these methods in this chapter, two of them will be implemented and modelled in the next chapter, i.e., Naïve method and Kernel estimation.

## 4.1 The data

Our data consist of the mail volumes at working days during three years. Let $\tilde{Y}_{i,j}$ be the mail volume at day $t_{i,j}, i = 1,...,n_j$, $j = 1, 3$. Here, $n_j$ is the number of days at which we observed the mail volume in year $j$, and $t_{i,j}$ is the date, i.e., the day number when we number the days in a year from one to 365. We also keep track of the days in a week using a seven dimensional dummy variable d:

d = (1, 0, 0, 0, 0, 0, 0) is Monday,
d = (0, 1, 0, 0, 0, 0, 0) is Tuesday,
d = (0, 0, 1, 0, 0, 0, 0) is Wednesday,
d = (0, 0, 0, 1, 0, 0, 0) is Thursday,
d = (0, 0, 0, 0, 1, 0, 0) is Friday.
d = (0, 0, 0, 0, 0, 1, 0) is Saturday.
d = (0, 0, 0, 0, 0, 0, 0) is Sunday.

Let $d_{i,j}$ denote the value of this dummy on day $t_{i,j}$.

## 4.2 The general introduction to models

The seasonal effects are estimated by taking the average mail volume over the years. Next, the average curve can then be further smoothed. The dependence on weekdays can then be estimated by looking at the mail volumes after subtracting the seasonal effects. The rationale behind this approach is that seasonal effects may show a periodic character, i.e., the seasonal curve has the same form every year. We model this idea as follows.

Let $Y_{i,j} = \log \tilde{Y}_{i,j}$ be the log mail volumes. (One reason to take a 'log' is that when a seasonal effect *m* is added to $Y_{i,j}$, it is in fact multiplying a *exp(m)* to $\tilde{Y}_{i,j}$. Since every weekday has difference levels of mail volumes, it is not reasonable to add the seasonal values to them on a same level. Through taking 'log' and 'exp', we transfer the 'add' to 'multiply by a scale', so that every weekday can get the seasonal effect by a scale.*)* We now assume that

(1) $$Y_{i,j} = \alpha(t_{i,j} + 365j) + \beta d_{i,j} + m(t_{i,j}) + \varepsilon_{i,j},$$

$$i = 1,\cdots,365,$$

$$j = 0,1,2,\cdots \text{ years since 2002,}$$

where $\alpha$ and $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^T$ are unknown parameters, $m(.)$ is an unknown function, and where $\varepsilon_{i,j}$, $i = 1,...,n_j$, $j = 1,2,3$, are independent errors with mean zero and finite variance. The $\alpha(t_{i,j} + 365j)$ models a linear trend. It is expected that this trend is decreasing due to competition and substitution. The function *m* represents the seasonal effects. Note that it is assumed to be the same function for the three years. Moreover, the parameter β representing the influence of the day of the week, is also assumed to be the same for the three years.

The model (1) is called *semiparametric*, because it contains a *parametric* part and a *nonparametric* or *infinite dimensional* part. This terminology refers to the "dimensionality" of the unknown parameters. The parameters $\alpha = (\alpha_{0,1}, \alpha_1)^T$

and β are clearly finite dimensional. They represent the parametric part. The function m is assumed to be smooth, but we do not assume a parametric form (a parameterization using finitely many parameters) for it. Therefore, *m* represents the nonparametric part. We remark here that also the distribution of the errors is not assumed to be known. This distribution may not be of primary interest to us, in which case we refer to it as an (infinite-dimensional) *nuisance* parameter.

α and β are put together in the finite dimensional parameter $\gamma^T = (\alpha^T, \beta^T)$ . Moreover, write $x = (1, t, d)$ with values $x_{i,2002} = (1, t_{i,2002}, d_{i,2002})$ for year 2002, $x_{i,2003} = (1, t_{i,2003} + 365, d_{i,2003})$ for year 2003, and $x_{i,2004} = (1, t_{i,2004} + 365 + 365, d_{i,2004})$ for year 2004. Our model can now be written in the form

$$Y_{i,j} = x_{i,j}\gamma + m(t_{i,j}) + \varepsilon_{i,j}, \quad i = 1,...,n_j, \quad j = 2002,2003,2004$$

The intuitive idea is to find the seasonal effects by smoothing the mean mail volume over the three years, and then finding weekday effects by subtracting the seasonal effects.

The first method we are trying to use is Poisson quasi-likelihood model, which is one of the generalized linear models.

## 4.3 Estimation

### 4.3.1 Naïve way of estimation

Let $Y_{2002} = (Y_{1,2002},...,Y_{365,2002})^T$ , $Y_{2003} = (Y_{1,2003},...,Y_{365,2003})^T$ , $Y_4 = (Y_{1,2004},...,Y_{366,2004})^T$ be the data of year 2002, 2003, and 2004. Firstly, get the means of $Y_{2002}, Y_{2003}, Y_{2004}$, which respectively are $E(Y_{2002}), E(Y_{2003})$, and $E(Y_{2004})$. Then get the *0-mean* data by subtracting the mean:

$$\Delta Y_{2002} = Y_{2002} - E(Y_{2002}),\ \ \Delta Y_{2003} = Y_{2003} - E(Y_{2003}),\ \ \Delta Y_{2004} = Y_{2004} - E(Y_{2004}).$$

Calculate the average of the *0-mean* data:

$$\Delta Y = \frac{\Delta Y_{2002} + \Delta Y_{2003} + \Delta Y_{2004}}{3}.$$

Smooth $\Delta Y$, then get the seasonal effects $S$. (Refer to Graph 5.2.3)

Subtract the seasonal effects from the $E(Y_{2002}), E(Y_{2003}), E(Y_{2004})$, and use a *Poisson* model with dummies for weeks to estimate it.

The utility of *Poisson* model on this project is based on some statistics theories, especially the results of Nelder(1974).

In statistics, there are special methods for dealing with discrete events rather than with continuously varying quantities. The enumeration of probabilities of configurations in cards and dice was a matter of keen interest in the eighteenth century, and from this grew methods for dealing with data in the form of counts of events, which is just the case of this project. The basic distribution here is Poisson, and it has been widely applied to the analysis of such data.

In this project, the mail volume every day can be regarded as following a multinomial distribution (The mail volume $y_i$ on day $i$ arrives with the probability $p_i$, and $\sum p_i = 1$.); such a distribution, as is well known, can be regarded as a set of independent Poisson distributions, subject to the constraint that the total count is fixed. A suitable initial (or minimal) log-linear model can be fitted to the data to fix the fitted total counts at their given values; additional terms to test the effect of other factors on the response factor will then be fitted conditional on those totals being kept fixed (Nelder, 1974).

Let the mail volume on day $i$ is $\tilde{Y}_i$, $i = 1,...,365 \times 3 = n$. Suppose that $\tilde{Y}_1,...,\tilde{Y}_n$ are independent *Poisson* random variables with means $\mu_1,...,\mu_n$ . Let $\mu_i = \exp[x_i^T \beta]$ with $\beta = (\beta_1,...,\beta_p)^T$ being unknown parameters and $x_i = (x_{i_1},...,x_{i_p})^T$ known constants.

Since $\tilde{Y}_i$ has the *Poisson* distribution, it has the probability function:

$$P(\tilde{Y}_i = y_i) = e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!}, \quad y_i = 0,1,2,...$$

So

$$\log P(\tilde{Y}_i = y_i) = -\mu_i + y_i \log \mu_i + (-\log(y_i!)).$$

We can neglect the last term because it has nothing to do with estimating the parameter $\mu$.

Furthermore, $\tilde{Y}_1,...,\tilde{Y}_n$ are independent, so

$$\begin{aligned}
&\log(P(\tilde{Y}_1 = y_1)...P(\tilde{Y}_n = y_n)) \\
&= \sum_{i=1}^{n} \log P(\tilde{Y}_i = y_i) = -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \log(y_i!) \\
&= -\sum_{i=1}^{n} \exp[x_i^T \beta] + \sum_{i=1}^{n} y_i x_i^T \beta - \sum_{i=1}^{n} \log(y_i!) \\
&= l(\beta) - \sum_{i=1}^{n} \log(y_i!)
\end{aligned}$$

Using Maximum Likelihood Estimation, we can find $\hat{\beta}$ which maximizes $l(\beta)$. So,

$$\hat{\mu}_m = \exp(x_m^T \hat{\beta}).$$

$$\hat{Y}_{i,j} = E(Y_{2002+j}) + S_i + \hat{\mu}_i .$$

Here S is the seasonal effects, which can be referred to the beginning of this section.

### 4.3.2 Kernel estimation

In this part, "Kernel Estimation" will be introduced by two steps.

**Step 1: Find the parameter for the linear part**

The first step is to estimate the important parameter $\gamma$. When this $\gamma$ is found, then the linear part is decide. Since the linear part can be obtained, then the seasonal effects can be estimated through separating the linear part from the data and smoothing. The estimator of $\gamma$ can be found through the formula in *Lemma 1*. In more detail:

Firstly we describe kernel estimation in regression (see for example see for example Nadaraya (1964) and Watson (1964)). Let *k* be a kernel, i.e., a function with finite support, satisfying

$$\int k(z)dz = 1, \quad \int k(z)zdz = 0, \quad \int k(z)z^2 dz < \infty .$$

Let *h* be a bandwidth, also called *tuning parameter*, and define the weights

$$w(s,t) = \frac{k((s-t)/h)}{\sum_{i,j} k((s-t_{i,j})/h)}.$$

If our model would not contain the parametric part $\gamma$, the kernel estimation of *m* would be

$$\hat{m}_0(t) = \sum_{i,j} w(t,t_{i,j}) Y_{i,j} .$$

To handle the parametric part, Speckman (1988) proposes the following. Let

$$\hat{m}_\gamma(t) = \sum_{i,j} w(t, t_{i,j})(Y_{i,j} - x_{i,j}\gamma) .$$

We now use the least squares estimator $\hat{\gamma}$ of $\gamma$, substituting $\hat{m}_\gamma$ for the function *m*, i.e. $\hat{\gamma}$ minimizes the sum of squares

$$\sum_{i,j} | Y_{i,j} - x_{i,j}\gamma - \hat{m}_\gamma(t_{i,j}) |^2 .$$

**Lemma 1**: An explicit expression for $\hat{\gamma}$ is

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

where X is the matrix with rows $x_{i,j}$ and $\hat{X}$ is the matrix with rows

$$\hat{x}_{i,j} = \sum_{k,l} w(t_{i,j}, t_{k,l}) x_{k,l} .$$

Likewise, Y is the vector with entries $Y_{i,j}$ and $\hat{Y}$ is the vector with entries

$$\hat{Y}_{i,j} = \sum_{k,l} w(t_{i,j}, t_{k,l}) Y_{k,l} .$$

Proof.

The right side of (4) can be rewritten as

$$\sum_{i,j} | Y_{i,j} - x_{i,j}\gamma - \sum_{i,j} w(t,t_{i,j})(Y_{i,j} - x_{i,j}\gamma) |^2$$
$$= (Y - X\gamma - (\hat{Y} - \hat{X}\gamma))^T (Y - X\gamma - (\hat{Y} - \hat{X}\gamma))$$
$$= (Y - \hat{Y} - (X - \hat{X})\gamma)^T (Y - \hat{Y} - (X - \hat{X})\gamma)$$

Set the derivative with respect to $\gamma$ equal to $0$,

$$(Y - \hat{Y} - (X - \hat{X})\gamma)^T (X - \hat{X}) = 0,$$

Which we can rewrite as,

$$Y - \hat{Y} - (X - \hat{X})\gamma = 0,$$
$$\Leftrightarrow (X - \hat{X})\gamma = Y - \hat{Y},$$
$$\Leftrightarrow (X - \hat{X})^T (X - \hat{X})\gamma = (X - \hat{X})^T (Y - \hat{Y}).$$

So

$$\hat{\gamma} = [(X - \hat{X})^T (X - \hat{X})]^{-1} (X - \hat{X})^T (Y - \hat{Y}).$$

**Step 2: Choose bandwidth *h***

This step is about how to find the best bandwidth *h* for the *smoothing*. The method is to use "RMSE" (Root Mean Squared Error) to choose the bandwidth. Make estimations with a specific bandwidth *h*, then use the model to make predictions for the *n* days in the future, for example, the predictions are $\hat{Y}_p^{(h)} = (\hat{Y}_{1,p}^{(h)}, \cdots, \hat{Y}_{n,p}^{(h)})$. If the real values on these *n* days are $Y_p = (Y_{1,p}, \cdots, Y_{n,p})$, then the RMSE is calculated as:

$$RMSE_h = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_{i,p}^{(h)} - \hat{Y}_{i,p}^{(h)})^2}.$$

For every bandwidth h, calculate this RMSE value, then draw a graph whose y-axis is

RMSE and x-axis is bandwidth. The best bandwidth is the one with smallest RMSE value, i.e. the lowest point in the graph.

### 4.3.3 Projection estimators

An alternative method to estimate the unknown parameters $\gamma$ and the unknown seasonal effect $m$ is the following. It is supposed that the function $m$ is smooth. We may write this mathematically as assuming that $m$ can be well approximated by a linear combination of given functions $\psi_1, \cdots, \psi_k$, say

$$m(\cdot) \approx \sum_{k=1}^{K} \theta_k \psi_k(\cdot) .$$

An example is choosing subintervals $I_u = [(u-1)h, uh]$, $u = 1, \cdots, \lceil 365/h \rceil = U$, where $h$ is the subinterval length, and

$$\psi_{u,v}(t) = t^{v-1} 1\{t \in I_u\}, \quad v = 1, \cdots, V .$$

The length(bandwidth) $h$ and the degree $V$ are again tuning parameters, to be chosen. For example, one may decide taking $V = 2$ and choose $h$ by leave-one-out cross-validation.

Recall our discussion of identifiability, and the restrictions given in (2). One can easily incorporate these restrictions by taking the part of the functions $\psi_k$ orthogonal to $(1, t)$.

Estimators $\hat{\gamma}$ and $\hat{\theta}$ can now be obtained using the least squares criterion, i.e., $\hat{\gamma}$ and $\hat{\theta}$ minimize the sum of squares

$$\sum_{i,j} | Y_{i,j} - x_{i,j}\gamma - \sum_k \theta_k \psi_k(t_{i,j}) |^2 .$$

The estimator of $m$ becomes

$$\hat{m} = \sum_k \hat{\theta}_k \psi_k \ .$$

### 4.3.4 Penalized least squares

In penalized least squares (see e.g. Wahba (1984)), estimators $\hat{\gamma}$ and $\hat{m}$ of $\gamma$ and $m$ are obtained by minimizing the quantity

(**)
$$\sum_{i,j} | Y_{i,j} - x_{i,j}\gamma - m(t_{i,j}) |^2 + n\lambda^2 J^2(m) .$$

Here $J(m)$ measures the "roughness" of the function $m$:

$$J^2(m) = \int_0^{365} | m^V(t) |^2 \ dt .$$

The degree $V$ is to be chosen. The case $V = 2$ gives the *cubic spline*.

Concept of *cubic spline*

A piecewise polynomial function $f(X)$ is obtained by dividing the domain of $X$ into continuous intervals, and representing $f$ by a separate polynomial in each interval. If the function is continuous, and has continuous first and second derivatives at the dividing knots, it is known as a *cubics spline*.

Furthermore $\lambda$ is a tuning parameter. The problem can be reformulated by writing

$$m = \sum_k \theta_k \psi_k$$

for suitable function $\psi_k$, known as B-splines. The restriction (2) can again be incorporated by taking the part orthogonal to $(1, t)$. We now have

$$J^2(m) = \theta^T Q \theta$$

is some quadratic form in the parameters $\theta$.

**Lemma 2**

The penalized least squares estimators are

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\theta} \end{pmatrix} = \left[ \begin{pmatrix} X^T X & X^T \Psi \\ \Psi^T X & \Psi^T \Psi + n\lambda^2 Q \end{pmatrix} \right]^{-1} \begin{pmatrix} X^T Y \\ \Psi^T Y \end{pmatrix},$$

where $X$ is the matrix with rows $x_{i,j}$, where $\Psi$ is the matrix with rows $\psi_{i,j} = \psi(t_{i,j})$ and where $Y$ is the vector with entries $Y_{i,j}$.

Proof

The (**) can be rewritten as

$$(Y - X\gamma - \Psi\theta)^T (Y - X\gamma - \Psi\theta) + n\lambda^2 \theta^T Q \theta .$$

Let its differentiate equals to 0,

$$\begin{pmatrix} X^T (Y - X\gamma - \Psi\theta) \\ \Psi^T Y - \Psi^T X\gamma - (\Psi^T \Psi + n\lambda^2 Q)\theta \end{pmatrix} = 0 ,$$

e.g.,

$$\begin{pmatrix} X^T Y \\ \Psi^T Y \end{pmatrix} = \begin{pmatrix} X^T X & X^T \Psi \\ \Psi^T X & \Psi^T \Psi + n\lambda^2 Q \end{pmatrix} \begin{pmatrix} \gamma \\ \theta \end{pmatrix}.$$

So

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\theta} \end{pmatrix} = \begin{pmatrix} X^T X & X^T \Psi \\ \Psi^T X & \Psi^T \Psi + n\lambda^2 Q \end{pmatrix}^{-1} \begin{pmatrix} X^T Y \\ \Psi^T Y \end{pmatrix}. \#$$

Using the representation (3), the estimator of $m$ becomes $\sum \hat{\theta}_k \psi_k$ .

One may also use finite differences instead of derivatives in the roughness penalty. For example, corresponding to the case of order $V = 2$ , one can choose

$$J^2(m) = \sum_{u=1}^{U-1} | m(t_{u+1}) - 2m(t_u) + m(t_{u-1}) |^2 .$$

Here, $\{t_u : u = 0, \cdots, U\}$ is assumed to be an equidistant set of time points containing the measured time points $\{t_{i,j}\}$ .

### 4.3.5 Asymptotic normality

All three methods above lead, under an appropriate set of conditions, to an estimator $\hat{\gamma}$ of $\gamma$ which is asymptotically normally distributed. For the kernel estimator, this is shown in Speckman(1988). Mammen and van de Geer(1997) establish asymptotic normality of the penalized least squares estimator. Asymptotic normality means in this case that $\hat{\gamma}$ is approximately normally distributed with mean $\gamma$ and covariance matrix $\sigma^2 \Sigma^{-1}$. The parameter $\sigma^2$ is the variance of the noise $\varepsilon_{i,j}$ (assuming equal variance here). The definition of a matrix $\Sigma$ for which the approximation hold true is rather involved. Roughly speaking, if we have an (infinite) expansion $m = \sum_k \theta_k \psi_k$ for $m$, and if $\tilde{x}$ is the projection of $x$ on the space spanned by $\{\psi_k\}$, and $\xi = x - \tilde{x}$ is the part of $x$ orthogonal to this space, then $\Sigma = \xi^T \xi$ is a choice for which the approximation holds true.

For the asymptotic normality of $\hat{\gamma}$, it is not necessary to assume that $x$ and $\psi_k$ are independent. It requires some involved theory to indeed show that such an independence assumption is not necessary. However, in our case, independence may actually hold true. There seems to be little reason to assume that whether or not a day is a Monday (say) depends on the time of the year (except perhaps that second Easter

day is on a Monday).

### 4.3.6 Bootstrap

To estimate $\sigma^2$ we propose

$$\hat{\sigma}^2 = \sum_{i,j} | Y_{i,j} - x_{i,j}\hat{\gamma} - \hat{m}(t_{i,j}) |^2 \Big/ (n_1 + n_2 - df).$$

Here, $df$ is some estimate of the *degrees of freedom*. For the asymptotic theory, the choice $df = 0$ is fine. In practice, $df = 0$ may be too optimistic. Other methods to estimate $\sigma^2$ are also possible.

To estimate the asymptotic covariance matrix, we propose a wild bootstrap. This works as follows.

- Generate $\{\varepsilon^*_{i,j}\}$ *i.i.d.* $N(0, \hat{\sigma}^2)$.
- Calculate

$$Y^*_{i,j} = x_{i,j}\hat{\gamma} + \hat{m}(t_{i,j}) + \varepsilon^*_{i,j}.$$

- Calculate the new estimates $\hat{\gamma}^*$ and $\hat{m}^*$ using the bootstrap data $\{Y^*_{i,j}\}$.
- Do this $N$ times, with $N$ large (for instance $N = 10000$).
- The distribution of $\hat{\gamma} - \gamma$ can now be approximated by the empirical distribution of $\hat{\gamma}^* - \hat{\gamma}$.
-

The wild bootstrap gives one confidence intervals for $\gamma$.

For the estimator of $m$ it is not so easy to give confidence intervals. If the estimator is not under-smoothed, there will be an unknown bias which is not asymptotically

negligible when compared to the variance.

### 4.3.7 Quasi-likelihood estimation

The above three methods have their obvious extensions to quasi-likelihood estimation. For example, one may also use a Poisson model for $\tilde{Y}_{i,j}$ with intensity

$$\mu_{i,j} = \exp[x_{i,j}\gamma + m(t_{i,j})] .$$

The computations then become more difficult, because even for fixed tuning parameters, the estimators will not be linear in (some given function of) $\tilde{Y}_{i,j}$. Since the mail volumes on each day are rather large, we believe that a normal approximation s sufficiently accurate and have therefore only treated the least squares method in detail.

# 5 Programming and results

This chapter mainly contains the programming and results of the two estimation methods introduced in Chapter 4, that is, the Naïve method and the *Kernel estimation*. After this, the analyses of results are provided.

## 5.1 Further data analysis

If the data is looked into more carefully, the following features can be noticed:

(1) December. This month is a very unusual month. The rules in normal weekdays don't apply to this month.

A Sunday in December can have a big volume, and almost all the days have a relatively larger mail volume than usual. The graph of December shows that there still exist some regular characteristics of this month and they can be modelled. Since TPG POST has a special model for it, this project will remove it from the data.

(2) Holidays. By referring to the original data, we can find the holidays have several kinds.

The data of holidays reveal that the *New Year* is different with other holidays. All the other holidays have a stable mail volume except for 9-5-2002, which we don't know why. And the *New Year* seems very unpredictable.

From the analysis, it seems that the holidays should be differentiated into two kinds: *New Years* and others. This can make the estimations for holidays more reliable.

(3) The days after some holidays. Referring to data, the day after *New Year*, the day after *Eastern Monday*, and the day after *Pentecost Monday* has abnormal mail volumes. Specially, the day after New Year always has less mail volume then other corresponding weekdays. For example, 2-1-2003 is Wednesday, yet it has about 4,000,000 less mail than other Wednesdays. So it will be better if set a

dummy for January 2nd.

As for the "*Eastern Tuesdays*" and "*Pentecost Tuesdays*", the people of Sorting Department say that they are always treated as Mondays. By looking up the data, it can be shown that they conform to the Monday mail volumes very well. So they can be dealt with as Mondays in the estimation.

After the further analysis of data, the following steps will be taken accordingly:

(1) Remove the Decembers.
(2) Set three dummies for holidays: $h1$ for normal holidays, $h2$ for *New Year's Day*, and $h3$ for *the day after New Year's Day*.
(3) Treat the "*Eastern Tuesdays*" and "*Pentecost Tuesdays*" as Mondays. This can be done by setting the values of these days in the dummies for Mondays as 1.

## 5.2 Modeling with Naïve method and Kernel estimation

### 5.2.1 Detailed modeling process of Kernel method

This section gives a detailed description of the Kernel method. The following steps are used:

(1) Read data of *X* and *Y*.

Firstly read data: $Y_{2002}, Y_{2003}, Y_{2004}, t_{2002}, t_{2003}, t_{2004}$, *the dummies of weekdays* $d1, d2, d3, d4, d5, d6$ *for the 3 years*, and holiday dummies $h1, h2, h3$. Here the $Y_{2002}, Y_{2003}, Y_{2004}$ are log of the mail volumes $\tilde{Y}_{2002}, \tilde{Y}_{2003}, \tilde{Y}_{2004}$, i.e., $Y_{2002} = \log \tilde{Y}_{2002}, Y_{2003} = \log \tilde{Y}_{2003}, Y_{2004} = \log \tilde{Y}_{2004}$. And $t_{2002} = (1,\ldots,334)^T$ $t_{2003} = t_{2002} + 365$, $t_{2004} = t_{2003} + 365$.

Then $Y = (Y_{2002}, Y_{2003}, Y_{2004})^T$, $X = (t, d1, d2, d3, d4, d5, d6, h1, h2, h3)$, in which

$$t = (t_{2002}, t_{2003}, t_{2004})^T,$$

$$d1 = (d1_{2002}, d1_{2003}, d1_{2004})^T,$$

$$d2 = (d2_{2002}, d2_{2003}, d2_{2004})^T,$$

$$\vdots$$

$$d6 = (d6_{2002}, d6_{2003}, d6_{2004})^T,$$

$$h1 = (h1_{2002}, h1_{2003}, h1_{2004})^T$$

$$h2 = (h2_{2002}, h2_{2003}, h2_{2004})^T$$

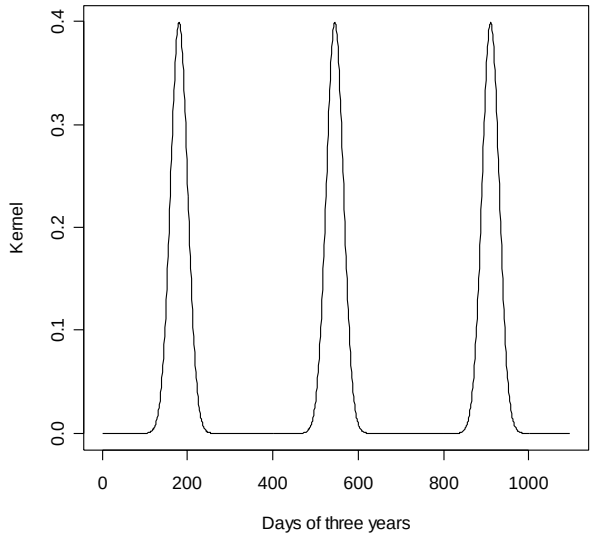
$$h3 = (h3_{2002}, h3_{2003}, h3_{2004})^T.$$

So $X$ is a matrix with 8 columns, and each column includes 3 years of data.

Now X and Y are both ready, next step is to calculate the parameters.

(2) Calculate $\hat{\gamma}$ according to the formula in Lemma 1 of Section 4.4.2.

In order to use this formula, we need $X, Y, \hat{X}, \hat{Y}$. Here $X$ and $Y$ have been obtained in Step (1), so we only need to calculate $\hat{X}$ and $\hat{Y}$.




Graph 5.1: Gaussian kernel graph at the 180th day every year

Graph 5.1 is the kernel graph at the 180th day of year 2002, 2003, and 2004. In the program we do the smooth for the average of three years, i.e.,

$$\hat{X} = smoothed \frac{1}{3}\sum_{j=1}^{3} X_{i,j} ,$$

$$\hat{Y} = smoothed \frac{1}{3}\sum_{j=1}^{3} Y_{i,j}$$

Then

$$\hat{\gamma} = [(\bar{X} - \hat{X})^{T}_{8\times 334}(\bar{X} - \hat{X})_{334\times 8}]^{-1}_{8\times 8}(\bar{X} - \hat{X})^{T}_{8\times 334}(\bar{Y} - \hat{Y})_{334\times 1}.$$

Since we get $\hat{\gamma}$, the estimator for the parametric part of the data is:

(3) Calculate the estimators for linear part and seadonal part.

Since the parameter $\hat{\gamma}$ has been obtained in step (3), the estimator for the linear part is $X\hat{\gamma}$.

And the seasonal part $\hat{m}$ is:

$$\hat{m}_{\hat{\gamma}}(t) = \sum_{i,j} w(t, t_{i,j})(Y_{i,j} - X_{i,j}\hat{\gamma}).$$

Again we take average across 3 years, and then apply the Gaussian kernel.

(4) Get the estimation for mail volumes.

Now we have estimated both the linear part and the seasonal part (i.e., seasonal effects), then the estimator for $Y$ is obtained by adding the seasonal part to the parametric part:

$$\hat{Y} = X\hat{\gamma} + \hat{m}.$$

The estimator for original data is: $\tilde{\hat{Y}} = \exp(\hat{Y}) = \exp(X\hat{\gamma} + \hat{m})$.

(5) Predict the future.

When predicting the future, the only thing to do is just set the matrix $X_{predict}$ properly, and the prediction is:

*Prediction* = exp ( linear part + seasonal part)

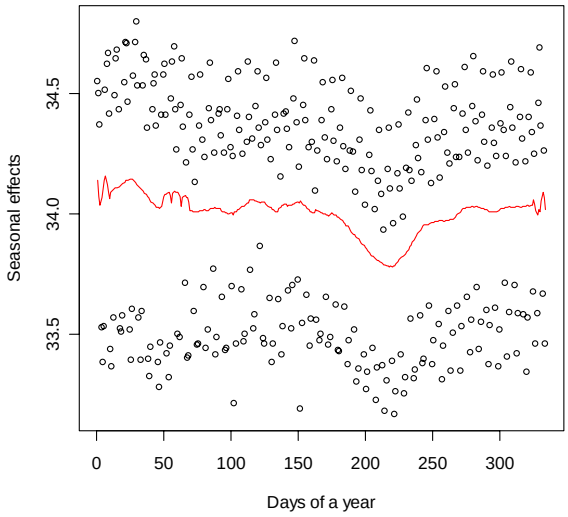
= exp ( $X_{predict}\hat{\gamma}$ + seasonal part) .

Here $X_{predict} = (t_{predict}, d1_{predict}, d2_{predict}, d3_{predict}, d4_{predict}, d5_{predict}, d6_{predict}, h_{predict})$ .

Again the $d1_{predict}, d2_{predict}, d3_{predict}, d4_{predict}, d5_{predict}, d6_{predict}, h_{predict}$ are the dummies of weekdays and holidays. While $t_{predict}$ is the continue of $t_{02}, t_{03}, t_{04}$. For example, in the prediction of the first three months of 2005, $t_{predict} = (1,2,\ldots,90)^T + 365 \times 3 + 1$. (Notice that 2004 has 366 days).

### 5.2.2 Seasonal effects of Naïve method

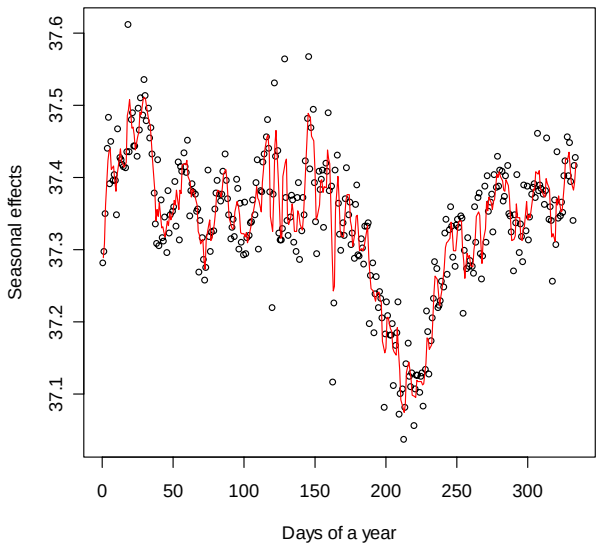
As stated in 4.4.1, a very important procedure is to smooth the seasonal effects. The graphs of smoothed seasonal effects with different bandwidths are given, then choose the one which is both smooth enough and at the same time able to describe the seasonal effects clearly enough. The black points in the following graphs are the data need to be smoothed, and the red lines are the smoothed results. (Note: The scales of the graphs have been changed in order to keep the business confidentiality.)




Graph 5.2: Smoothed seasonal effects with bandwidth=20
(Naïve method)

### 5.2.3 Seasonal effects of Kernel method

The **Step 2** of Section 4.4.2 gives a method to choose bandwidth. Draw a graph of "Error *versus bandwidth*" , and the smallest error is at bandwidth=2. The Graph 5.2.5 is the graph of smoothed seasonal effects when bandwidth is 2. Notice that Graph 5.2.5 shows some monthly bumps, which probably has something to do with companies sending out bills at the end of the month, or salary statements or other reasons. When discussing mail volumes at TPG no indication was given that there might be a monthly effect. Consequently, we did not include such an effect in our model. The fact that the optimal bandwidth is chosen very small reflects that the model is not able to properly accommodate the effect.




Graph 5.3: Smoothed seasonal effects with bandwidth=2
(Kernel method)

### 5.2.4 Coefficients of Naïve method and Kernel estimation

Except for the smoothing seasonal effects, another important part is the linear part, and this part is decided by the coefficients. The coefficients for Naïve method are obtained through fitting Poisson model, and the coefficients for Kernel method are obtained through a mathematical formula. Both of them describe the generalized linear trend of the mail volumes. Due to the competition and substitution, the coefficients of time are minus values, which stand for a slowing declining trend.

## 5.3 Results and analysis

### 5.3.1 Graphs of estimations

The estimation graphs for year 2002, 2003, and 2004 can be drawn now. From the graphs, the following features can be seen:

(1) The estimation can follow the data trend. The seasonal decline in summer holidays is described clearly by the model.
(2) The different levels of trends between weekdays are revealed. The low mail volumes on Sundays and the relatively high volumes on Thursdays, etc., are shown clearly. It is due to the use of dummies for weekdays.

There are still some big differences between estimated data and real data. This will be analyzed in the next Section.

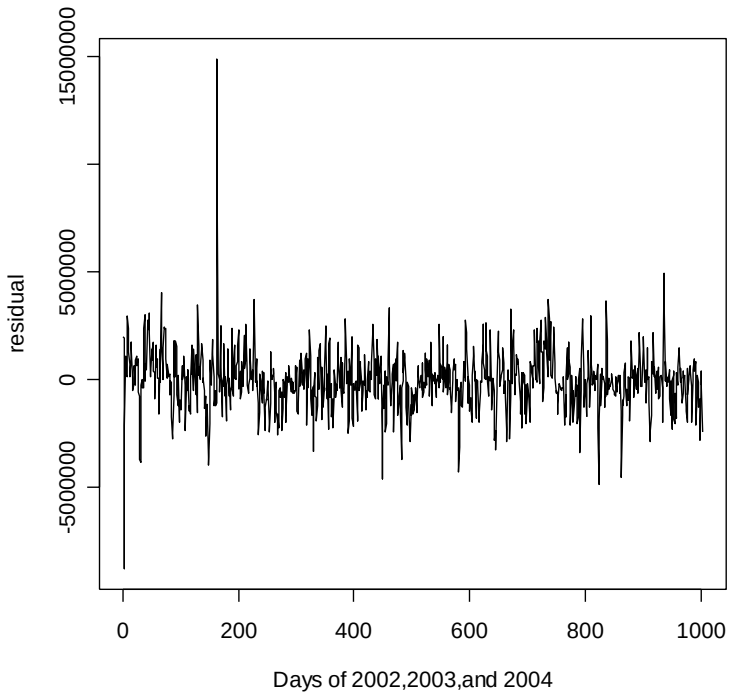
### 5.3.2 Analysis of differences between estimation and real data

This section concentrates on analyzing the residuals and difference percentages between estimation and real data. They are calculated as:

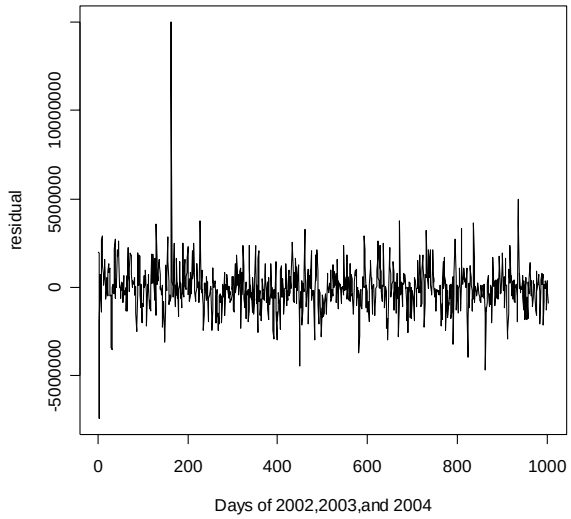
$$\varepsilon = residual = \hat{Y} - Y$$

$$difference\ percentage = \frac{\hat{Y} - Y}{Y}.$$

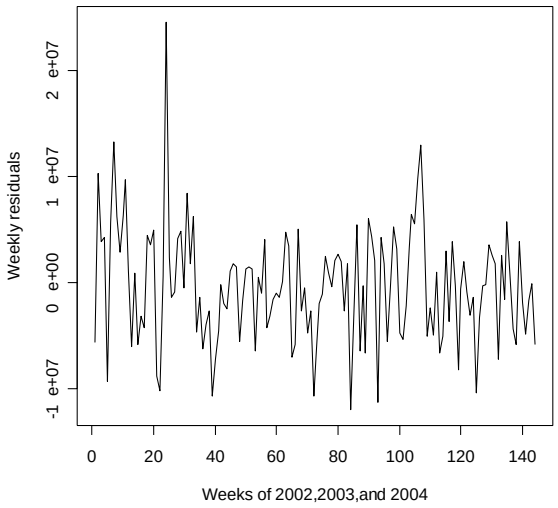
The following are the graphs of residuals and difference percentage. Firstly the graphs of the residuals and difference percentages are shown, and then the lists of data whose difference percentages are beyond [-0.3,0.3] are analysed.



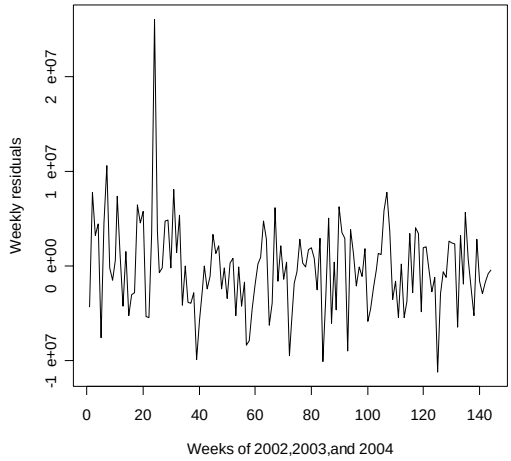

Graph 5.4: Residual of Naïve method




Graph 5.5: Residual of Kernel method



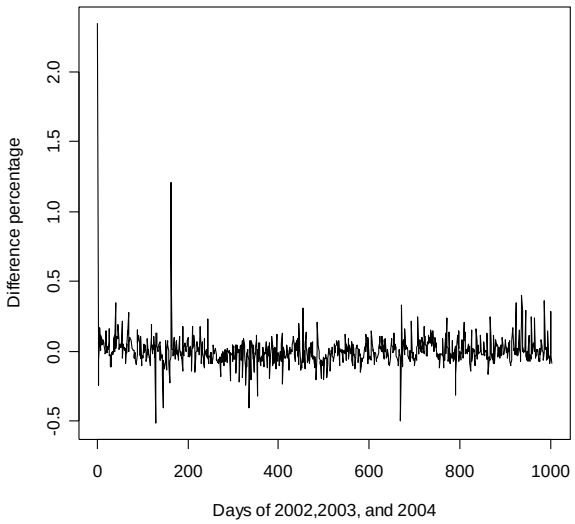

Graph 5.6: Residual of Naïve method (weekly)



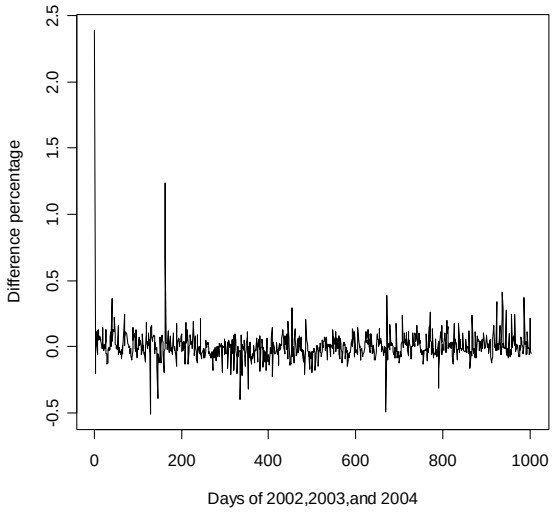

Graph 5.7: Residual of Kernel method (weekly)

The graphs above show the residuals and difference percentages for the two methods, daily and weekly. If analyse the results a little further, calculate and compare the means of residuals and difference percentages, it show that the Kernel method is better than the Naïve method because it has smaller residual values and difference percentages, both daily and weekly. On the other hand, if seen from the perspective of daily analysis and weekly analysis, the weekly estimation is better because it has smaller residual values and difference percentages. Yet the difference between daily and weekly are not very much, which is only improved by about 0.8%.

The below Graph 5.8 and 5.9 give the graphs of difference percentages, which show that there exist some big values.




Graph 5.8: Difference percentage of Naïve method




Graph 5.9: Difference percentage of Kernel estimation

If analyze the difference percentages carefully, it can be shown that most of the big difference percentages come from Sundays and holidays. Since the mail volumes on Sundays and holidays are very small, it is difficult to get a low difference percentage for them.

Other big differences happen when the mail volumes themselves on these days reveal big differences compared to the mean values, the big differences between these estimators and real data can be understood.

This is the analysis for the difference percentages. With this analysis now it is clear why the big differences happen. The next section will focus on the predictions for 2005.

## 5.4 Prediction for half year of 2005

The prediction for the first half year of 2005 shows that it can describe the data's trend. If see the difference percentage that are beyond [-0.3, 0.3], they are almost for Sundays or Holidays or the first beginning days of January, which are the same case as the analysis for estimations above.

Here shows the RMSE values for the two estimations. "RMSE" means "Root Mean Squared Error", which is one of the most commonly used measures to judge how close a prediction is to its target. In more detail: use the models to make predictions, for example for the first half year of 2005. And then find out what is the "difference" between the predicted values and the real values. This "difference" can be calculated with the method "RMSE". So it is a very important value, because by comparing the RMSE values, it shows which method is relatively better. The smaller RMSE, the better.

In this project, the RMSE for the first half year of 2005 is calculated as below:

$$RMSE = \sqrt{\frac{1}{202}\sum_{i=1}^{202}(Y_{i,2005} - \hat{Y}_{i,2005})^2} .$$

Here there are 202 values for the half year of 2005.

Though calculating, the *RMSE* value of Kernel method is less than Naive method, so it is better

# 6 Conclusions and recommendations

In this project we considered the problem of forecasting mail volumes. At the suggestion of TPG post we focused on a year effect (with mail volumes dropping in the summer), and a daily effect (with mail volumes depending on the day of the week). Also we included a linear trend to account for the loss of mail volume due to competition and substitution. Finally, we accounted for unusual volumes on holidays (New Year, Easter, Queen's day, Ascension and Pentecost Monday) and on the days that immediately follow them. We did not include December, because the mail volumes during that month behave very differently from the other months.

We started with a naive look at the data taking various simple averages. We refer to the results of this explorative work as the "naive method".

Next, in section 4.2, we formulated a semi parametric model. We estimated the parameters of this model by a combination of least squares and a kernel smoother. We refer to this method as the "kernel method". We determined the optimal bandwidth of the kernel by cross validation using data from 2005 which were not used for fitting. In terms of the root mean squared error, the principled kernel method outperforms the ad hoc naive method.

Firstly, the general information about sorting process and sorting centres are provided, including the six sorting centres, the different kinds of sorting machines, the 24-hour service and 48-hour service, etc.

Then is the exploration for data. The INDOOR and MIS are both focused on. The graphs of data give two general impressions. One is the obvious yearly decline during summer holidays; the other is the wiggly pattern which reveals the characteristics for different weekdays. According to these two features, the main idea comes out, i.e., separate the seasonal effects from the data and use dummies for different weekdays.

Since the main idea is decided, the mathematics ground is studied. There are more than one method to separate the seasonal effects from the data, and Chapter 4

introduces two of them: the Naïve method and the Kernel method.

Next step is to realize the mathematical models. Before doing the programming, look further into the data, and find that the following details should be noticed: one is the December is a very special period so its graph is not so regular as others; the other is that the holidays have many kinds so should not be treated the same. After this analysis, the December is excluded from our model, and two more dummies for different holidays are added.

Doing the estimations with these two methods and all of the analysis above, then get the estimation data. By analyzing the difference between estimation and real data, we can see that the models can describe the data set. If use the model to predict the first half year of 2005, the conclusion is that it can estimate the data. It is impossible that a model can describe every point precisely, since it can show the trend of data and is close to every weekday, we can say that it is doing its work satisfactorily.

Notice that the beginning few days of January are still influenced by the "KNJ" period, so the estimations around are probably not good enough. Also how long this model can be used is also in question. This model is composed of the linear trend and the seasonal effects. As we see, the linear part is a slowly declining trend due to the substitution and competition. We are not sure that after 5 year or 8 years, the trend is still stable. So how long this model can be used should be tested in practice.

If compare the results of the two methods, the Kernel method is better than the Naïve method. In practice, the Kernel method can give more precise estimations, so it is recommended to be used. The Naïve method comes from our intuitive ideas and lacks a solid mathematical background, so it is not recommended to be used for estimations.

The reason that why this project only focuses on these two methods is the following:

Our model is: $Y_{i,j}$ = parametric part + non-parametric part + noise.

It is only focused on estimating the parametric and non-parametric part. Other methods, e.g., Box-Jenkins, or the ARIMA time series model, has to do with the noise term. If the noise is correlated, then the current value of $Y_{i,j}$ can be used to predict a

future value. Our prediction, however, is essentially based on the assumption that the noise is uncorrelated. This seems
reasonable, especially since we typically want to predict more than a few days into the future.
Also there are three years of data available. For a time-series model, three years of data is not enough in order to make a good prediction. Combining all the above, it gives the reasons that why our project concentrates on these two methods.

If someone wants to improve this present model, the attention can be put on the monthly effects. As the graph of seasonal effects shows, there must exist some monthly effects. If the monthly effects can be modelled, then the results will be improved. Some ideas about dealing with the monthly effects are shown below:

Draw graphs for every month, and then see what happens. When doing so, the monthly effects are just like the seasonal effects that are discussed in this project, so similar method can be used on it. In more detail, firstly remove the day-of-week effect, and then draw a graph of the average month (think about how to deal with the different number of days 28,30, and 31). Then use dummies 1, … ,30 to encode the day of the month and include it in the present model. This is one of the possible solutions. Yet it needs more careful study and probe.

## References


Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist*. **16** 136-146

Mammen, E. and van de Geer, S.A. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist*. **25** 1014-1035

Nadarya, E.A. (1964). On estimating regression. *Theory Probab. Appl*. **9** 141-142

Speckman, P. (1988). Kernel smoothing in partial linear models. *J.R. Statist. Soc.* **B 50** 413-436

Wahba, G.(1984). Partial spline models for the semiparametric estimation of functions of several variables. In *Analyses for Time Series*, Japan-US Joint Sem. 319-329. Tokyo: Institute of Mathematical Statistics

Watson, G.S. (1964). Smooth regression analysis. *Sankhya* **A 26** 359-372

McCullagh & Nelder (1989). *Generalized Linear Models*, Chapman & Hall