



Centraal Bureau voor de Statistiek

Divisie Methodologie en Kwaliteit

Sector Methodologie

Postbus 4481

6401 CZ Heerlen

**A Comparison of Different Estimation Methods of
Voting Transitions with an Application in the
Dutch National Elections**

Carin van der Ploeg

Samenvatting: Er zijn vele uitdagingen in het schatten van het stemgedrag van kiezers. Eén hiervan is het schatten van verschuivingen in het kiesgedrag, zogenaamde kiezersstromen. Deze kiezersstromen kunnen worden gepresenteerd in een transitie matrix. Modelleren en survey methodologie zijn twee verschillende algemene methoden die deze transitie matrix kunnen schatten. Onderzoek laat zien dat beide methoden voordelen en nadelen hebben. In dit paper worden deze twee methoden met elkaar vergeleken op basis van de voorspelde tweede verkiezingsuitslag. Die wordt verkregen door de uitslag van het eerste moment te vermenigvuldigen met de transitie matrix. De focus van het onderzoek vindt plaats binnen de context van de Nederlandse tweede kamerverkiezingen van 2003 en 2006, waar de uitslagen kunnen worden geobserveerd op gemeenteniveau. De resultaten laten zien dat op gemeenteniveau er realistische schattingen kunnen worden gemaakt van het stemgedrag. Tevens blijkt dat surveytechnieken en modelleringstechnieken elkaar kunnen aanvullen, wat de validiteit van de resultaten verder versterkt. Dit biedt een goed uitgangspunt voor verdere studies.

Trefwoorden: *Stemgedrag, Lineaire regressie, Kwadratisch programmeren, Latente Markov ketens, Iteratief Proportioneel Fitten, Ontrouwe-trouwe kiezersmodel, Nederlands KiezersOnderzoek (NKO)*

Summary: There are many challenges in the estimation of voting behavior. One of these is the estimation of shifts in voting preferences, so called voting transitions. These voting transitions can be represented in a transition matrix. The estimation of changes in voting behavior can be pursued in two ways, by modeling and with a survey. Research shows that both have their advantages but also their downsides. We compare these two approaches using the election results on the first moment multiplied by this transition matrix to estimate the election results on the second moment. The focus of this research is on the Dutch National Elections of 2003 and 2006 where the results can be observed on municipality level. This paper shows that realistic estimations of the election results on municipality level can be made. It also appears that survey techniques and model based techniques are able to complement each other, improving the validity of the results providing basis for new research.

Keywords: Voting transitions, Linear Regression, Quadratic programming, Latent Markov, Iterative Proportional Fitting, Mover-Stayer model, Dutch National Survey (NKO)

Acknowledgements

I wish to thank my supervisors dr. René Bekker, dr. Jarl Kampen and dr. Frank Pol for their guidance, comments and advice during the internship period. Especially the cooperation with dr. Jarl Kampen and the discussions and his contributions have been interesting and useful. The input of my second supervisor Prof. dr. Bert Kersten has also been very valuable. I'd like to thank dr. Peter Kruiskamp for his love, feedback and patience and also my other fellow colleagues at the CBS for their assistance and comments. Furthermore I am very grateful to Joost Bosman for his various useful contributions.

Contents

Contents.....	4
1. Introduction.....	6
1.1 Introduction to the organization	8
1.2 Research question.....	9
2. Established methodology for describing mover-stayer behavior.....	11
2.1 Introduction.....	11
2.2 Benchmark methods	13
2.3 Survey methodology.....	13
2.4 Model based methodology.....	16
2.4.1 Linear regression.....	16
2.4.2 Quadratic programming.....	17
2.4.3 Latent Class Analysis	18
2.4.4 Iterative Proportional Fitting.....	21
2.4.5 Combinations	21
2.4.6 Short summary	23
3. Application in the Dutch territory.....	24
3.1 Introduction.....	24
3.2 Available data	27
3.3 Fit measures	28
4. Results.....	31
4.1 Goodness of fit statistics	31
4.2 Benchmark methods	33
4.3 NKO	36
4.4 Quadratic Programming.....	36
4.5 Latent Class Analysis	37
4.6 Iterative Proportional Fitting.....	37
4.7 Combinations	38
4.8 Discussion about the elections 2003-2006.....	39
4.9 Seating distributions 2003-2006.....	41
5. Discussion and further research.....	45
5.1 Ecological inference	45
5.1.1 EI estimator.....	46
5.2 Discussion.....	47
6. Bibliography.....	50
7. Appendix A	55
7.1 Analysis of parties	55

7.2	Transition matrices 1998-2002, 2003-2006	56
7.3	Comparison all models per election year	68
7.4	Extensions to the mover-stayer model	72
7.5	Used programs and code	73

1. Introduction

Politicians, political scientists and the media are very interested to map the transitions in party preference of voters between elections. This is visible in the large amount of literature on *movers* and *stayers* that can be found in the literature and goes back at least to Leo Goodman's article from 1961, *Statistical methods for the mover-stayer model*. Also, many Dutch opinion polls varying from Maurice D'Hondt, Interview NSS to TNS NIPO periodically publish predictions of political transitions and support for political parties. The goal has always been to estimate the voting behavior as precisely as possible. This is a big challenge because due to the confidentiality of the voting ballot there is only data available on an aggregated level (national, regional or municipal), while we want to know the behavior of individuals. Many researchers try to overcome the lack of data on individual voting behavior by organizing surveys. However, for a number of reasons explained below, surveys are not necessarily a guarantee for reliable results.

The background for this particular study is a paper of Keller and ten Cate (1977). Their article with the title "De verschuiving van kiezersvoorkeur" posed the interesting question 'where do voters go'. They constructed a single matrix to try to describe the voting transitions from 1972 until 1977 in the Netherlands. We took this work as a starting point for the development and comparison of several new and existing models. The original model used by Keller and ten Cate has also been included in this thesis. These mathematical models aim to describe the transitions of party preferences on the basis of aggregated information on municipality level.

To get a fast idea of the subject of this dissertation a short hypothetical example, given by Keller & ten Cate (1977) is presented. In this example, as in the analyses in this paper, the aggregate level corresponds to the electoral results in municipalities. Imagine that there are two parties (party A and party B) and two cities (municipality 1 and municipality 2). As can be seen in table 1 below, party A is the largest party in both municipalities and has obtained 1000 votes in 2003 in municipality 1 and 300 votes in municipality 2. The votes for party B can be derived in the same way. In municipality 2 party B obtained 200 votes in 2003 and won 70 votes in 2006 bringing their total number of votes at 270. In the 2006 elections party A loses votes in both municipalities whereas party B gains support. For the sake of this example it is assumed that the total number of votes are equal in both elections.

Table 1. Two parties, two cities and two elections				
	2003		2006	
	Party A	Party B	Party A	Party B
Municipality 1	1000	800	780	1020
Municipality 2	300	200	230	270

In order to calculate the transition rate from party A to party B we can present the number of votes casted in a simple equation. Let P_{AA} be the transition rate from party A to party A, the so called stayers, and P_{BA} be the transition from party B to party A, the so called movers. Then,

$$1000 \times P_{AA} + 800 \times P_{BA} = 780$$

$$300 \times P_{AA} + 200 \times P_{BA} = 230$$

We can make a simple calculation and find the transitions rates that belong to this equations. One can easily verify that the transition rates add up to 1. These values show that 70% of the support for party A remained, but that 30% of their followers transferred to party B in 2006. The supporters for party B where even more loyal since even 90% stayed with this party.

$$P_{AA} = 0.70 \quad P_{AB} = 0.30$$

$$P_{BA} = 0.10 \quad P_{BB} = 0.90$$

The last step is to present the transition rates in a matrix giving a very simple transition matrix. This matrix is extended to include all municipalities and more parties and is the center of the analysis.

	2006	
2003	A	B
A	0.7	0.3
B	0.1	0.9

These values can then be filled back into the equation obtaining the initial voting results:

$$1000 \times 0.7 + 800 \times 0.1 = 780$$

$$1000 \times 0.3 + 800 \times 0.9 = 1020$$

$$300 \times 0.7 + 200 \times 0.1 = 230$$

$$300 \times 0.3 + 200 \times 0.9 = 270$$

This simple and well chosen example gives a unique transition matrix. However, adding more parties and more municipalities increases complexity of the problem rapidly, and requires the application of more elaborate modeling approaches. Analysis shows that it is possible to identify a unique transition matrix in a 1x1, 2x2 and 3x3 situation. When increasing the problem to 3 parties and 3 cities it is still possible to estimate a unique transition matrix. It is not possible to find a unique solution anymore when the complexity increases. Under strict assumptions is Ordinary Least Squares indeed unbiased. But when dependencies arise it is easy to show that there is an infinite set of solutions and no unique transition matrix.

As an alternative to model based approaches based on aggregated data analysis, survey research is often employed to gather detailed information from a representative part of the population. Using the responses of a sample, researchers try to make inferences about the rest of the population. This method provides very useful data and is used in many applications. However, this method also has its drawbacks. A lot of research is done to overcome these problems, but some still remain. Therefore, it has been interesting to investigate possible combinations using both aggregated and survey information. in the present analyses of voting transitions, the model based approaches are compared to the survey based approach in order to gain more insight in voting transitions.

1.1 Introduction to the organization

Statistics Netherlands has the task to collect, edit and publish statistics that is relevant in practice, and for policy and research purposes. Next to the responsibility for the national (official) statistics, Statistics Netherlands is also responsible for the production of European (community) statistics. The legal ground for Statistics Netherlands is the "Law on the Central Bureau of Statistics" of November 2003 (Staatsblad, 2003, p516)¹. There are two establishments: one in Voorburg near the Dutch governmental centre and one in Heerlen in the south of the Netherlands.

¹ "Het Centraal Bureau voor de Statistiek (CBS) heeft tot taak het verzamelen, bewerken en publiceren van statistieken ten behoeve van praktijk, beleid en wetenschap. Naast de verantwoordelijkheid voor de nationale (officiële) statistieken is het CBS ook belast met de productie van Europese (communautaire) statistieken. De wettelijke grondslag voor het CBS is de "Wet op het Centraal Bureau voor de Statistiek" van 20 november 2003" (Staatsblad, 2003, p516)

Statistics Netherlands actively performs research in order to improve their methods in general, making research and development an important task of Statistics Netherlands. The sector Methodology and Quality (DMK) is situated in both locations. This sector contributes to the knowledge development of Statistics Netherlands and aims to continuously improve their statistical process.

1.2 Research question

The goal of this study is to improve and compare methodology regarding the estimation of voting transitions. An extensive comparison is made between the model based methodology and survey based methodology. The election results from two consecutive elections are used to estimate voting transitions between two subsequent elections. The first instant is the election year of 2003 and the second instant is the election year of 2006. The quality of the estimated transition matrices is evaluated by estimating the voting results at the second instant by the information contained in the first election and the transition matrix evaluated. These estimates of the second instant, can be compared to the real election results, and a measure of fit of the estimates (e.g., likelihoodratio) can be computed.

This methodology of estimating election results by means of a transition matrix is applicable in a broader field of cohort studies with categorical variables. Two advantages of the present study specific to electoral studies are emphasized here. First, there is an added value for the specific field of electoral studies in providing an answer to whether the used methodology is suitable for estimating voting transitions on municipality level. Second, there is possible improvement of the accuracy in estimating voting behavior. In comparison with the NKO, improvement can be possible using model based methodologies.

The above considerations lead to the main research question:

To what extend can voting transitions in the Netherlands on municipality level be accurately described where the necessary transition matrices are lacking?

In order to answer this research question it is broken down into the following sub-questions:

- Which model based approach is best suited to estimate voting transitions on municipality level?
- How can survey methodology and model based methodology complement and improve each other?

The following general hypotheses will be drawn:

- Estimation of voting behavior in terms of estimated transition matrices is a reliable alternative for the methodology used to derive transitions in survey analysis.
- There are different transition matrices for each municipality which can in part be explained by demographic variability

The remainder of this paper is structured as follows: in chapter 2 the state of the art of the established methodology and explanation of all the used models is presented, followed by an introduction into the contextual setting of this research. In chapter 3 the data properties and constraints are presented. In the chapter 4 the results and analysis of the results are presented, and finally chapter 5 contains the conclusion and topics for further research.

2. Established methodology for describing mover-stayer behavior

2.1 Introduction

Electoral studies form a separate discipline in social science. Much of the research within electoral studies has focused on two-party systems and less on multi-party systems. In the last few years more research has been done on the latter subject, but this is still small compared to research on two-party systems (Quinn *et al.*, 1999). A short overview of important contributions both on two-party as well as on multi-party systems is provided in this chapter. The emphasis of the overview is on the methodology used within the context of this research, but other related research is also briefly mentioned so to provide a complete picture of the research that has been done in the area of electoral studies.

In order to describe voting behavior, it is necessary to study categorical data. A categorical variable is simply a variable for which the measurement scale consists of a set of categories. They can be on a nominal (not ordered) or ordinal (ordered) scale (Agresti, 1996, pp. 2-3; Kampen & Swyngedouw, 2000: p. 87). However, especially in social science, variables often are not directly observable, and latent variables are used to explain reality. If for instance, only electoral results are known for two subsequent elections, the transition matrix can be modeled in terms of latent variable. To see this more clearly, some notation needs to be introduced first. The formulas are expressed in conditional probabilities. In a Markov Chain transition probabilities are expressed as

$$P(X_n = j \mid X_{n-1} = i) = p_{ij},$$

Where the parties from the first election are represented with index $i=(1, \dots, m_i)$ and the parties from the second election with index $j=(1, \dots, m_j)$, In this paper, we apply the following basic notation for probabilities used the research:

- $p_i = \Pr(i)$, the probability of voting party i at the first election at national level,
- $p_j = \Pr(j)$, the observed probability of voting party j at the second election at the national level,
- $p_{ij} = \Pr(i, j)$, the unobserved joint probabilities of the observed marginals,

- $p_{ji} = \Pr(j|i) = \frac{\Pr(i,j)}{\Pr(i)}$, the *unobserved* proportion of voters transferring from party i to party j ,
- $p_{ig} = \Pr(i|g) = \frac{\Pr(i,g)}{\Pr(g)}$, the observed proportion of voters who voted for party i in municipality g (voting preference first instant, for example 2003),
- $p_{jg} = \Pr(j|g) = \frac{\Pr(j,g)}{\Pr(g)}$, the observed proportion of voters who voted for party j in municipality g (voting preference second instant, for example 2006).

After introducing this basic notation, the more generalized situation can be expressed in the following basic formula:

$$\Pr(j|g) = \sum_i^{m_i} \Pr(i|g) \times \Pr(j|i), \quad (2.1)$$

The values of $\Pr(j|g)$ and $\Pr(i|g)$ are known, and the value of $\Pr(j|i)$ (the elements of the transition matrix) must be estimated. As can be seen this formula can also be used in the example above. Analogous to Keller & ten Cate (1977) a single national transition matrix is adopted to calculate all transitions. This means that the elements of $\Pr(j|i)$ are independent of municipality g . The conditional probabilities or transition weights $\Pr(j|i)$ give the voting transitions between two elections. $\Pr(i|g)$ and $\Pr(j|g)$ and $\Pr(j|i)$ can not be derived directly from the election results. In general the following formula for the elements of the transition matrix is:

$$\Pr(j|i) = \Pr(i,j)/\Pr(i), \quad (2.2)$$

The joint probabilities $\Pr(i,j)$ can in general not be observed when using aggregated data. They must be estimated using empirical data or, as in survey analysis, by using recall data to estimate the underlying (latent) transition matrix.

A mathematical approach to estimating the underlying transition matrix is linear regression, which is an intuitive and simple technique, but it may produce unrealistic results when it is applied in practice. The simple technique of linear regression can be expressed as a quadratic programming problem by adding more restrictions, giving more realistic results. Latent Markov theory and latent class models make it possible to combine latent variables with the information contained in observed

variables in the estimation of voting transitions. Closely related to the latent class approach is loglinear modeling. These mathematical techniques and the survey methodology are the basis for this study.

2.2 Benchmark methods

Instead of estimating transitions, we may simply fill in (conceptually defensible) values in the transition matrix and assess the fit of such an ad hoc or benchmark model. Two methods in this study are used purely as a benchmark. The first method is the independency model or so called perfect mobility model. This means that the vote of the constituent at the second instant is totally independent from its vote at the first instant. This can be represented by this formula:

$$\Pr(j|i) = \Pr(j). \quad (2.3)$$

Formula (2.3) states that the probability of voting for party j given party choice i in the previous election is equal to the probability that there will be voted on party j . The information that the voter first voted for party i is not taken into consideration.

The second method is called the 100% stayers model. This means that voters are perfectly loyal to their party. A voter who votes for party j at the first instant will vote for exactly the same party on the second instant, leading to the following formula:

$$\Pr(j|i) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases} \quad (2.4)$$

Formula (2.4) states that voters will simply repeat their initial choice i . Because of lack of a ‘golden’ measurement method it is interesting to use these extremes for comparison and to evaluate the performance of other models.

2.3 Survey methodology

There are various research bureaus all aiming to explain what the Dutch voter is going to do during coming elections. The Dutch Parliamentary Election Studies is the largest survey carried out and is specialized in describing and understanding Dutch voting behavior. The Dutch Election Survey (*Nederlands Kiezers Onderzoek*) is carried out every national election. The last survey has been carried out by Statistics Netherlands. Similar surveys are carried out in many democracies to gain insight into the most important democratic process which is voting. In the Dutch Election Survey (NKO) respondents are asked what their most recent voting preference is as well as their preference at the previous election. Using these data the

real transitions between parties back and forth can be expressed in a transition matrix. For this study the book written about the Dutch elections of 2006 "Een verdeeld Electoraat De Tweede Kamerverkiezingen van 2006" using the NKO research, by Aarts, Van der Kolk & Rosema, (2007) is relevant. The transition matrix can be found in Aarts *et al.* (2007: pp. 224).

The added value of surveys is that it is a method of observation able to offer a general capability in describing the characteristics of a large population. Since measurement of the variables is on the individual level, surveys are also able to analyse on micro level of voting preferences between two elections. Transitions can therefore be analyzed in both directions. It is possible to see for instance who voted in the first election on the PvdA and the second election of the CDA. But we can also see who voted for the CDA in the first election and PvdA in the second election. We are able to see this way how many voters switched between two parties and visualize more than just the total nett transitions between two parties, which means that we can only see the party who got the most support during the last elections. Then we only see the nett transition going from the PvdA and CDA. Surveys are in a way flexible because they are able to collect a wide range of information. It is therefore possible to correlate between socio-demographic properties, attitudes, values, beliefs and (past) voting behavior, allowing for possible explanatory factors of certain behavior.

The problem with this kind of data is that recall data is often very unreliable (Weir, 1975). There are several effects making their accounts unreliable (Voogt, 2004; Keller & ten Cate, 1977; Upton, 1977):

- Respondents do not recall their previous vote,
- respondents do not want to be fickle, so there is a bias towards answering the same choice of party between the previous and current election,
- respondents answer incorrectly and name the party for which they voted in the municipality, provincial or European elections,
- respondents say that they have voted but in fact didn't vote in the previous election, wanting to give a sociable accepted answer,
- respondents want to belong to the winning side and answer incorrectly by giving a party who won in the elections.

Furthermore, there are several effects that provide a considerable bias in the survey results. It is however not easy to determine exactly what the size of this bias is. Voogt, (2004) has done research into the Dutch National Voting Survey. According

to his research the most important effects are non-responsive bias, response bias and stimulus effect (Voogt, 2004). Non-responsive bias is the most important effect and is responsible for an underrepresentation of certain groups within the population. Extensive literature on this problem shows that in general non-respondents are higher educated, younger, more often single with overrepresentation in the urban areas (see for a list of published literature Voogt, 2004: pp. 35). The response bias is also called ‘answer and memory effects’ and reflects the above mentioned reasons for unreliable results (Bethlehem & Kersten, 1986). The last bias (stimulus effect) occurs especially in panel research, where respondents are at least interviewed twice (Greenwald *et al.*, 1987). People that where normally not planning to cast their vote will vote because they have participated in the part of the election research that occurred before the elections. For a further review on the strength and weaknesses of surveys, see Voogt, (2004). Because of lack of a ‘golden measurement method’ it is not easy to determine exactly what the impact of these causes for bias are. Therefore we need to rely on methods that can only partly correct these errors, making the determination of the transition matrix on the basis of surveys not straightforward, which can lead to difficulties.

The participating respondents are asked which party they have voted for in the current national elections and in the previous elections. The cross table of these two recollection variables divided by the row totals yields the requested transition matrix, the so called conditional probabilities $\Pr(j|i)$. Just as in Aarts *et al.* (2007) the sample will be multiplied by a weight factor in order to obtain the most truthful representation of the Dutch Population². To test whether the NKO transition matrix represents reality, the estimated values are applied in the right side of formula (2.1) where subsequently the election results of the second instant are predicted per municipality from the results of the first instant.

An interesting addition into overcoming bias in survey research is presented by Barbosa and Goldstein (2000). Their contribution is useful in longitudinal survey research. They present a method, extending the Goldstein method, which is a multilevel time series model, to include discrete variables for normally distributed responses. This means that the answers given by the respondents follow a normal distribution. This methodology is better capable to estimate the true proportion of movers and stayers and take strategic voting into consideration (Barbosa & Goldstein, 2000).

² The values of this transition matrix differ slightly from the values in the book of Aarts *et al.* (2007) because the matrix in this paper uses the correctly rounded values and the row totals add up to 1. This is not the case with the matrix presented in the book of Aarts *et al.*

2.4 Model based methodology

2.4.1 Linear regression

Keller and ten Cate introduced an interesting approach to estimate voting transitions. They used linear regression to estimate voting transitions within the Netherlands. Their method can be viewed as a sample in which the sample size and population size coincide. Because it is not possible to ask the voter what he or she has voted in the previous election, results are only available on an aggregate level, which is the municipality level. In 1977 there were 842 municipalities in the Netherlands. The property that the election results per municipality are spread over multiple parties is used in the following way: When one party loses and another party wins in a municipality they assume that an overflow occurs between those two parties. The transition probabilities are considered to be constant for all municipalities. They compare their results to the largest survey at that time, the Intromart survey, and the actual voting results, and conclude that their results are relatively good in comparison. They note however, that some probabilities are extremely high, and that in comparison to the Intromart survey there are a lot of probabilities of zero where they are not expected (Keller & ten Cate, 1977).

The basic method of Keller and ten Cate is the linear probability model

$$\Pr(j) = \sum_i \Pr(i) \times \Pr(j|i) + \mathcal{E}(j) \quad (2.5)$$

Where in their case, $\Pr(i)$ is number of votes on party i in 1972, $\Pr(j)$ is the number of votes on party j in 1977, $\Pr(j|i)$ denote the transition probabilities, and $\mathcal{E}(j)$ is random interference term or error. One can estimate the transition matrix using Ordinary Least Squares, that is, by minimizing the squared discrepancies between the actual and estimated values, which is equivalent to minimization of

$$\sum_j \sum_g (p_{jg} - c_j - \sum_i p_{ig} \hat{p}_{ji})^2 \quad (\text{Oosterhoff \& Vaart, 2003, pp. 40}).$$

However, Keller and Tencate note that the population changes between two elections. The voters who have stayed at home are an additional category and the new/resigned voters are added as an extra party of origin. This adds errors into the calculation. They force equality in the total number of votes per municipality but the estimated probabilities $\Pr(i,j)$ are sometimes out-of-bounds, meaning that they are not necessary always between 0 and 1. The problem with this method is that negative probabilities and probabilities larger then one can be obtained (Keller & ten Cate, 1977, pp. 4). With the constrains that the estimates have to be between 0 and

1, and the restriction that the transition probabilities need to sum up to 1 the problem becomes a quadratic programming problem.

2.4.2 Quadratic programming

As mentioned, the estimates of $\Pr(i,j)$ is sometimes out-of-bound. Several authors have adapted the remedy of constrained regression to correct for this problem (Telser, 1963). Quadratic programming is (also) one of the offered solutions (see e.g. Judge & Takayama, 1966; Irwin & Meeler, 1969; McCarthy & Ryan, 1977 and Keller & ten Cate, 1977). The critique on using this method comes from Cleave et al. (1995) who states that this method avoids the formulation of a suitable model Cleave et al., 1995, pp. 5). The authors however fail to offer an alternative solution or more in depth critique, therefore providing little basis for discussion.

Quadratic programming forces restrictions on the objective function.

$$\min_{\hat{\beta}} \left(\left\| \Pr(i)^T \Pr(i) \hat{\beta} - \Pr(i)^T \Pr(j) \right\|^2 \right)$$

Under:

$$\hat{\beta} \geq 0$$

The restrictions are:

1. $\sum_{i=1}^{10} \Pr(i, j) = \Pr(j)$ and $\sum_{j=1}^{10} \Pr(i, j) = \Pr(i)$.
2. $\sum_{j=1}^m \Pr(j|i) = 1$.

Where:

- i is party preference on the first moment (for example 2003),
- j is party preference on the second moment (for example 2006),
- m is total number of political parties,
- $\hat{\beta}$ produces the required transition coefficients under restrictions 1 and 2.

The first restriction states that the marginals of $\Pr(i,j)$ must add up to the observed marginals (this means the election results). The second restriction states that the probabilities to vote for party j given i must add up to 1. Remember that this was not the case with linear regression.

The key difference with normal linear regression is that the matrix with the results of 2003 and the matrix with the results of 2006 are put together into one vector with length 100. $\hat{\beta}$ contains the coefficients over which the restrictions of the objective function are enforced. The only restriction for $\hat{\beta}$ is that it should be equal or larger than 0. This means that all small quadratic programming problems were described per party in one big vector. Additional restrictions could therefore be enforced. $\hat{\beta}$ is the matrix containing the $\Pr(i,j)$'s with the transition probabilities put into a vector with length 100. Thus by reformulating OLS one creates a quadratic programming problem with the goal of minimizing the mean square error (Tijms & Ridder, 2003). Both the statistical software programs S-plus and R can be used to produce actual computations of linear quadratic programming.

2.4.3 Latent Class Analysis

The traditional Markov model is not appropriate to predict long-term social data such as voting behavior (Van de Pol & Langheime, 2004, pp. 2). Latent Class Analysis however, is very suitable for analysis of longitudinal data (Langheime & Pol, 1990; Van de Pol & Langheime, 2004; Vermunt et al., 1999), and suited to estimate voting transitions because of its ability to use nonobservable variables. The underlying model for latent Markov chains is the mover-stayer model, which has been extended over time and will be presented in this section. The first occurrence of the Mover-Stayer model was in 1955. The article by Goodman (1961) presents an approach that has become well known as the mover-stayer model. It was introduced in a job-switching context (Blumen *et al.*, 1955). It is a special case of the mixture Markov model (Poulsen, 1982). The very basic underlying structure of this model is that there are movers and stayers. In the population there are people who vote for the same party in successive elections and people who switch to another party with probability $p(j)$, regardless of the previous election. It provides an elegant explanation for the apparent lack of correlation between party switching. Goodman used the notion of semi-independence (independence between elections) to estimate the elections of 1959 to 1974 (Goodman, 1968). His results showed that for short term analysis the Mover-Stayer model was indeed an interesting way to visualize voting transitions. In the long term it is not suitable, since in contrast to the short term, voters do change their voting preference over a long period of time. Therefore, the estimated proportion of stayers is too large (Upton, 1977, pp. 2). Further elaborations on extensions of the mover-stayer model see section 7.4 of Appendix A.

In order to obtain further improvement of the mover-stayer model, we can incorporate measurement error into the model. This can be done using latent or indirectly measured Markov models. Latent Markov Models are well suited for individual change description in categorical data. The key element of this model is that it views observed change partly as measurement error of latent change (indirectly measured change). In this model there is no time-constant latent variable because it is assumed that there is no unobserved latent heterogeneity (Van de Pol & Langheime, 2004, pp. 4). Methodology such as parametric, nonparametric and semiparametric models is only suited for continuous observations. Because this problem involves discrete data and exact transition times are not available, they are often referred to as panel data (Cook *et al.*, 2002).

Van de Pol and Langheime (1990) describe a mixed Markov latent class model which was originally developed by Wiggins (1973) and Poulsen (1982). The model is able to describe turnover rates between two or more panel waves, which is the case with voting data. In this model there are more latent classes than observed categories and is only identifiable when enough turnover is observed (Van de Pol & Langheime, 1990, pp. 29).

The model used in this study is based on the work of Clogg and Goodman (1984). It is called Latent Class Analysis (LCA). This method searches for the most reasonable values of the conditional probabilities $\Pr(j|i)$, given the election results on both moments. The parameters can be estimated using maximum likelihood. This function can be formulated for the complete table $\Pr(i, j, g)$, the number of votes on party i and j in municipality g for at least one of both moments. The maximum likelihood estimate is the parameter value for which the observed data has the largest probability. This parameter is given by the likelihood function with its maximum value (Agresti, 1996, pp. 8-9). Because of the analytical infeasibility of direct calculation of the Maximum Likelihood (ML) this iterative procedure can be used. Here it is not necessary to calculate the Hessian matrices.

The likelihood function (L) is given by:

$$L = \prod_{ijg} \Pr(i, j, g)^{f_{ijg}} = \prod_g \Pr(g)^{f(g)} \prod_i \Pr(i|g)^{f(i|g)} \prod_j \Pr(j|i, g)^{f(j|i, g)} \quad (2.7)$$

Where f corresponds with frequencies. f_{ijg} are the election results for every municipality. $f_{(i|g)}$ are the election results for party i on the first instant given municipality g . $f_{(j|i, g)}$ are the election results for party j on the second instant given voting preference on the first instant in municipality g .

This can be estimated using the Expectation Maximization algorithm (EM), for which PANMARK was developed. Because of the implementation of heterogeneity in the model, LCA will be more realistic than the mover-stayer variants (Van de Pol & Langheime, 1990, pp. 9; Vermunt *et al.*, 2006, pp. 5).

The 1977 paper by Dempster *et al.* introduces the generalized EM algorithm. It is an iterative computation of maximum-likelihood estimates in case the data is incomplete (Dempster *et al.*, 1977, pp.1). It is based on a relatively simple idea in order to deal with incomplete data problems (Dempster *et al.*, 1977):

- Replace missing values with estimated values
- Estimate the parameters
- Estimate missing values again under the precondition that the new parameter estimate is correct
- Estimate the parameters again
- Repeat parameter estimation until convergence

In the E step the conditional expectation of a complete-data log-likelihood function of the missing data is given based on the observed data and on the current estimated expectation of the missing parameters. In the M-step the parameters are updated so that the expectation is maximized. Maximizing the lower bound in each step is often an easier method than direct maximization of the log-likelihood function. This lower bound always increases because of the M-step in the algorithm, ensuring convergence to a solution. However, the problem is that this algorithm sometimes converges to local optima and no guarantee is given that it converges to the global optimum. (Dempster *et al.*, 1977).

Because of the fact that the complete table $\Pr(i, j, g)$, is not observed, it must be calculated. By creating a temporary transition matrix on the basis of the election results on the first moment, one performs the E-step of the algorithm (Little & Rubin, 1987). The M-step is done by dividing $\sum_g \Pr(i, j, g)$ by $\Pr(i)$ (see formula (2.2)) and creating a new estimation of the transition matrix. The optimization of the ML occurs by adjusting the (temporary) complete table $\Pr(i, j, g)$ alternately to the elections of the first moment and the second moment. This is the start of the M-step, which occurs before the estimation of a new transition matrix.

LCA that uses the EM algorithm solves the problem of unrealistic estimations and is different from the ordinary least squares estimates, which is a method that minimizes the error, whereas this method maximizes the estimates. Latent Class Analysis and Iterative Proportional Fitting (IPF) are both methods to get to a maximum likelihood

estimate. The underlying algorithm however differs. LCA is based on the EM algorithm and IPF is based on the IPF algorithm.

2.4.4 Iterative Proportional Fitting

Iterative proportional fitting (IPF) is a mathematical procedure which is originally developed to combine information from two or more datasets (Bishop *et al.*, 1975). In this case we use information from two instants; the elections of 2003 and 2006. In a first step, the probabilistic model is expressed in terms of e.g., Formula (2.1). This model contains unknown probability table, to be precise, the transitions $\Pr(j|i)$. The IPF estimating approach proceeds by first, filling in values for these transitions, and then improve these estimates in a stepwise procedure that continuously updates the estimates. In our case the initial election results are randomly chosen to obtain the initial values. The input is the estimate given in the case of the benchmark model of total independence:

$$\hat{f}_{ijg}^0 = n \times \frac{f_{g++}}{n} \frac{f_{+i+}}{n} \frac{f_{++j}}{n} \quad (2.9)$$

The next step is to optimize the initial estimate:

$$\hat{f}_{ijg}^{t+1} = \frac{\hat{f}_{ijg}^t f_{gi+}}{\hat{f}_{gi+}^t} \quad \hat{f}_{ijg}^{t+2} = \frac{\hat{f}_{ijg}^{t+1} f_{g+j}}{\hat{f}_{g+j}^{t+1}} \quad \hat{f}_{ijg}^{t+3} = \frac{\hat{f}_{ijg}^{t+2} f_{+ij}}{\hat{f}_{+ij}^{t+2}}$$

Dividing by n can be done at the end of the calculation to get the probabilities of \hat{f} .

Where the f 's are obviously the observed marginal entries and the \hat{f} 's are the estimated marginal entries. The iterative procedure is terminated once the marginal totals agree closely with the observed marginals and is smaller than some prespecified (small) criterion. Under certain conditions of regularity, the IPF updates converge to a ML solution.

2.4.5 Combinations

Another interesting notion is to combine the usefulness of survey research with real election data, economic variables and demographic variables. Little research has been done to truly incorporate these information sources into one model. Mostly, model outcomes and survey research are compared, where survey data are used as a sort of benchmark. It remains difficult to estimate true individual behavior and inferences are still necessary.

King proposes to insert aggregated level data (at the lowest level possible) into the American version of the Dutch NKO research (King, 1996). Egmond *et al.*, (1998) however show, in their research on political participation in which they consider the period of 1971 until 1994, that most contextual variables do not explain much of the variation in the case of turnout levels. Furthermore, they conclude that most variation and explanatory power arises from the context of the elections themselves (Egmond *et al.*, 1998). This may also be true for the electoral behavior and it possibly restricts the potential of contextual variables.

An interesting alternative view on using aggregated and survey information was implemented by Thomsen (2004). In this paper he describes a multiparty situation from Denmark using a conditional logit model as well as survey information. Thomsen only describes the general voting behavior on the national level with a utility model (Thomsen, 2004). This paper continues on his 1987 paper in which he introduces his nonlinear estimator. He treats consecutive elections as symmetric and states that they are a result of one common latent factor which he calls party identification (Thomsen, 1987). In his 2004 paper he inserts party sympathy and issue voting to the mix and adds more explanatory power and accuracy to his model.

In this study two ways are employed to combine the NKO-survey results with the Quadratic Programming model. The QP-model is used because it is relatively easy to incorporate more restrictions into the goal function. Those restrictions are added using a very straightforward confidence interval. We impute the NKO values into the QP-model and allow some freedom for the model to estimate the values, but only within a Confidence Interval (CI):

$$CI := \left[\bar{x} - 2 \times S_x, \bar{x} + 2 \times S_x \right] \quad (2.10)$$

The Sample Standard Deviation (S_x) is then:

$$S_x := \frac{s}{\sqrt{n}} \quad (2.11)$$

Where:

- $s = \sqrt{\frac{1}{N-1} \times \sum_i^N (x_i - \bar{x})^2}$

- $n=459$ municipalities

(Oosterhoff & van der Vaart, 2003).

In an attempt to obtain interesting combinations between the NKO and model based approaches (Quadratic programming was used in this case) two combination models have been created:

- Replace values on the diagonal within the CI with NKO-values
- Replace zero's within the CI with NKO-values

The reason for the first model was to replace unrealistic values on the diagonal. For example, values of one are not sensible and can be corrected using the NKO-values of the diagonal. This also influences the other values in the matrix, because all rows have to add up to 1. The other motive to combine the two approaches is to try to replace possible unrealistic zeros in the matrix (second model). Because of the fact that model based approaches are only able to estimate the nett-transitions, the NKO presented some starting values for these zeros based on the NKO. Again this has an influence on the other values as well, because of the constraints. In the end, constrained models will always produce less optimal solutions than unconstrained models.

2.4.6 Short summary

Table 2 shows that most analyses are based on the analogy of Keller & ten Cate (1977). The new elements in this study are that new models are compared to each other. Furthermore, when using estimations from the second moment which are calculated using the election results from the first moment, one can compare these values to the real election results and determine the fit of these estimates. The models can be evaluated using this methodology and this is another addition to existing research on this topic. Also, restricting the model using combinations of survey and model results have been applied to try to increase the validity of the models.

Table 2. Comparison table

	Weighted survey	Linear regression	Quadratic Programming	Maximum Likelihood*
NKO	X			
Keller & ten Cate		X	X	X
Marginal results equal				X
Combo 1	X		X	
Combo 2	X		X	

* Maximum Likelihood holds both LCA and IPF

3. Application in the Dutch territory

3.1 Introduction

The Netherlands have a multi-party system with proportional representation, which means that entrance for smaller parties is relatively easy. This infers a higher volatility in the number of parties in the Dutch system. The Netherlands is a quite interesting case because in the last decennia more and more voters have not been faithful to one political party, making it harder to predict what voters will do in the coming election. Quin *et al.* (1999) even state that the Netherlands is a critical case because of a substantial amount of research into Dutch voting behavior. The Netherlands has a system of proportional representation without the problem of district factors. In addition there are varying degrees of support by Dutch voters in the last years (Quinn *et al.*, 1999, pp. 3).

This study can offer an interesting contribution to the explanation of the rise and fall of political parties and of voting behavior of the electorate. For example, the rise of Pim Fortuyn and the more recent victory of right wing Geert Wilders was a surprise to many.

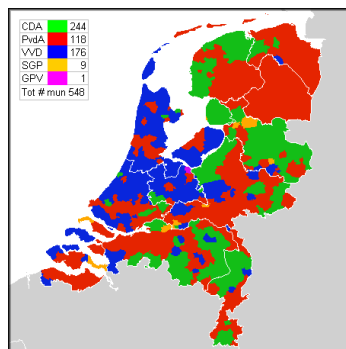


Figure 1a. Dutch election results 1998

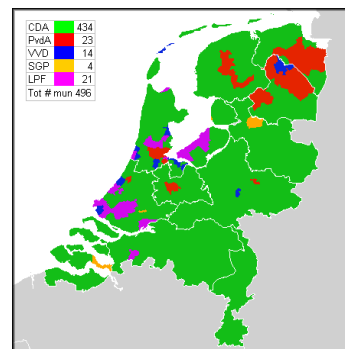


Figure 1b. Dutch election results 2002

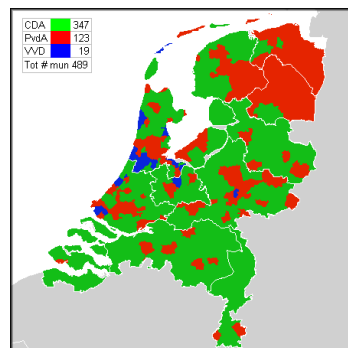


Figure 1c. Dutch election results 2003

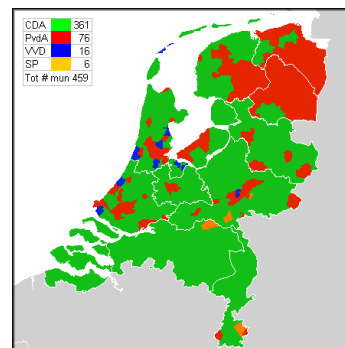


Figure 1d. Dutch election results 2006

(Kiesraad.nl, 2008)

Figures 1a through 1d show that on a macro level the election results are still relatively stable during the last few elections. Studies in several fields have indicated that people are still relatively loyal to their political party (see e.g. Shachar & Shamir, 1996). The Christian-democrats (CDA) continuously remained the largest party in most of the municipalities. The legend shows that the number of municipalities has dropped from 548 in 1998 to 459 in 2006. Although transitions between parties have become more volatile in the last decade, they are still reasonably stable. More than 50% of the voters still decide what to vote several months before the elections, but this number had dropped from 71% in 1977 to just over 50% in 2003 (NKO, 1977-2003). The defeat of the so called purple parties (government parties PvdA, D'66 and VVD) in 2002 is well visible on the map. Also the defeat and recovery of the socio-democrats (PvdA) in the last two elections is evidently present. Despite the net changes, which have been large in the cases of Lijst Pim Fortuyn in 2002 and 2003 and the Socialist Party in 2006, many voters remain stable in their voting behavior (see also figure 3).

Table 3. Election results 2003 and 2006 in percentage

Party	Election results (%)				Change 1998-2002	Change 2002-2003	Change 2003-2006
	1998	2002	2003	2006			
Christian Democrats (CDA)	18.37%	27.93%	28,6 %	26,51%	+9.56%	+0.69%	-2,13%
Labour Party (PvdA)	28.98%	15.11%	27,26%	21,19%	-13.87%	+12.15%	-6,07%
Socialist Party (SP)	3.53%	5.90%	6.32%	16,58%	+2.37%	+0.42%	+10,26 %
Liberal Party (VVD)	24.69%	15.44%	17.91%	14,67%	-9.25%	+2.47%	-3,25%
Group for Freedom (PVV)	-	-	-	5,98 %	-	-	-
Green Left (GL)	7.27%	6.95%	5,14 %	4,60%	-0.32%	-1.82%	-0,54%
Christian Union (CU)	-	2.54%	2,12 %	3,97%	-0.76% *	-0.42%	+1,85 %
Democrats 66 (D66)	8.99%	5.10%	4,07%	1,96 %	-3.89%	-1.03%	-2,11%
Party for animals (PvdD)	-	-	0,49 %	1,83 %	-	-	+1,34%
Christian Reformed Party (SGP)	1.78%	1.72%	1,56 %	1,56 %	-0.06%	-0.16%	0,00%
List Pim Fortuyn (LPF)	-	17.00%	5,70 %	0,21 %	-	-11.30%	-4,93%
Reformed Political Union (GPV)	1.26%	-	-	-	-	-	-
Reformational Political Federation (RPF)	2.03%	-	-	-	-	-	-
Liveable Netherlands (LN)	-	1.61%	0.40%	-	-	-1.21%	-
Other	3.10%	0.71%	1.31%	1.00%	-2.39%	+2.02%	-0.31%

*GPV and RPF are added together to make this comparison

(Source: CBS.nl:Statline)

Table 3 distinctly shows the shifts between parties. As already mentioned, you can see the largest shifts in support with the PvdA in 2002, the LPF in 2003 and the SP in 2006. The traditional Christian party CU shows a steady growth, whereas other parties such the GL and D'66 show a steady decline. Despite some temporarily and interesting changes, the established parties remain well represented in the Dutch parliament.

3.2 Available data

The data needed for this type of analysis are the election results on the lowest level available which in this case is the level of municipalities. For all municipalities the frequencies of all party choices are known. A complicating factor is that the municipalities have been redistributed over time. Between 1998 and 2006 the number of municipalities has decreased from 548 to 459, making some data processing necessary. The solution to this complication was to sum up the results of all constituting parts of redistributed municipalities, so results will be comparable. The results from overseas voting have been merged into “Den Haag Postal voting”, with a separate code, and are treated as one municipality.

There is another complication for comparing results. Some parties only exist in one or a couple of consecutive election periods. For example, the List Pim Fortuyn (LPF) was only popular during 2002 and 2003. From the NKO results one can easily see where the votes from the LPF have gone: to the Partij voor de Vrijheid (PVV). Most PVV votes originate from the former followers of the LPF, so the PVV has in a sense replaced the LPF in 2006.

Not all parties are large enough to be taken into consideration. They compete only in a limited number of constituencies, or do not meet the electoral threshold to be elected into parliament. With so few votes it is not possible to visualize the voting transitions of these parties separately. We combined these votes in the “remainder” category. We consider 10 parties, i.e. $m_i = m_j = 10$. The reason for this selection of parties was based on comparability and on the number of seats gained in the parliament. In section 7.1 of appendix A the parties of the elections of 1998, 2002, 2003 and 2006 are summed up.

Another anomaly in the data was that on the Dutch islands turnout levels were above 100% in some election years. This can be explained by the fact that a lot of people visit these islands for vacation and use their so called voting pass to vote in another municipality than the one they are officially registered in. In this way it is possible that more people cast their vote than there are residents in that municipality. Interestingly enough, those values above 100% have occurred in 1998 and 2002 only, when elections were held in May.

An interesting issue is whether data on a more disaggregated level, such as the election results on the municipality level, will provide better results than election results on the national level only. The municipality level data does have advantages (Park, 2004: pp. 14):

- The aggregation problem is less severely present on this lower level.

- More observational units (instead of just the country or provincial election results) gives larger variances and therefore better precision of the estimates.

He shows that the aggregation bias does actually decrease. However, he states that the bias introduced by non-linearity of the results may even increase, especially when a linear model such as the linear regression model is assumed (Park, 2004).

Every election period new voters enter the registry and others leave the registry due to death, emigration etc. This is carefully registered and has influence on the results. Due to the large amount of work that would have been involved, these concerns have not been taken into consideration in this study. In every election a small number of votes are not valid for some reason. These votes have not been taken into account because they are not included in the calculation of the electoral threshold. (See section 4.9).

The electronic availability of Dutch election results on municipality level is still rather limited. Only the results of the last four elections in 1998, 2002, 2003 and 2006 are electronically available. We had to abandon our plans to compare the 1972/1977 research of Keller & Ten Cate with the more recent election data, because unfortunately the authors did not have the data anymore. The Dutch parliament does have all election results on paper, though. Via the “Image and Sound” service (*beeld en geluid*), it is also possible to get election results, but again not in a suitable format. Fortunately, the voting council (*kiesraad*) is currently creating an electronic archive of all elections on municipality level. This is planned to be finished sometime next year, which is too late for this study. Because it was too time-intensive to enter all the data ourselves it was decided not to pursue this any further.

3.3 Fit measures

In order to correctly compare all models to each other some methods are necessary which will be introduced. To summarize the discrepancies between predictions and actual election results we use the well known Pearson goodness-of-fit chi-square (GFX^2). Its formula is:

$$GFX^2 = \sum_j \sum_g \frac{(\widehat{\Pr(j|g)} - \Pr(j|g))^2}{\Pr(j|g)} \quad (4.1)$$

This formula calculates the relative quadratic differences between observed and expected values, summed over parties, j , and municipalities, g (Long, 1997). Lower values indicate a better fit of the model. However, this value is not evaluated as

such, but compared to the value that is obtained under the assumption that subsequent election results are independent in each community.

The log likelihood ratio statistics was also calculated as a control statistic to verify the found goodness-of-fit statistics. The statistic is calculated as:

$$LRX^2 = 2 \sum_j \sum_g \log\left(\frac{\widehat{\Pr(j|g)}}{\Pr(j|g)}\right) \quad (4.2)$$

This statistic is analogous to the goodness-of-fit statistics. Its behavior is only asymptotically different from the GFX^2 . When the values of the GFX^2 and the LRX^2 differ to greatly one should be very cautious about the validity of the results. (Bishop, Fienberg and Holland, 1975)

To be better able to visualize the difference between the benchmark model of total independence and the other models, the difference of two chi-squares is also calculated. This gives an indication of the improvement in fit of the other models and is called $DLRX^2$.

The difference in degrees of freedom (df) in comparison to the benchmark model is presented as Ddf. For the independence model $9+9=18$ parameters need to be estimated, but for the NKO matrix the number of parameters is $10 \times 9 = 90$. Therefore, Ddf between the two is 72. Using the same analogy, the rest of the parameters of the other models can also be calculated.

McFadden's pseudo R^2 is used to show the fit of all models. It is calculated by taking by dividing the chi square of the model with the chi-square of the independence model.

The formula for the McFadden pseudo- R^2 is (McFadden, 1973, pp. 121):

$$Ps-R^2 = \frac{(l_m - l_0)}{(l_{\max} - l_0)} = 1 - \frac{l_m}{l_0} \quad (4.3)$$

Where l_{\max} is the perfect fit, l_m is the chi-square of the full model for example LCA and l_0 is the chi-square of the benchmark model of independence.

Larger values of the R^2 indicate a better fit of the involved model. When $R^2=1$ then, on the basis of the transition matrix used, the election results from the second

moment on the municipality level can be predicted perfectly from the election results of the first moment³.

³ Veall & Zimmermann (1996) have written a survey on the various forms of pseudo- R^2 in representing model fit. Their main conclusion is that there is no obvious criterion to decide which specific pseudo- R^2 is best. The McFadden pseudo- R^2 is the most commonly used pseudo- R^2 . Cameron & Windmeijer (1993) have made a generalization to cover a wider variety of situations. According to Vaell & Zimmermann the pseudo- R^2 is conceptually close to the R^2 value that can be calculated from OLS. This is a recommendation, because the OLS pseudo- R is traditionally used as a goodness-of-fit measure for linear models like regression. Because there is no consensus on what pseudo- R^2 is best in a certain situation, one must be careful when using, comparing and interpreting results. We note that a different pseudo- R^2 can produce other results and brings about other interpretations of the values. For example the McFadden pseudo- R^2 can produce a value of .25 on the same calculation whereas the McKelvey-Zavoina R^2 produces a value of .5.

4. Results

In this chapter the results of the analyses are presented. First, there is a section on the goodness-of-fit results. In the next sections all models will be discussed separately. The following section contains the discussion on the 100% stayer model, the NKO model and LCA model for the elections of 2003-2006. For all elections different transition matrices are estimated, and for the 2003-2006 election the NKO model and LCA model are presented and reviewed. The chapter ends with a section on seating distributions and the model performances on that account. The other transition matrices and analysis of 1998-2002 and 2002-2003 are presented in Appendix A, section 7.2 and 7.3.

4.1 Goodness of fit statistics

Table 4. Goodness of fit statistics for all models 1998-2002

Model	GFX ²	LRX ²	DLRX ²	Ddf	P	McFadden Pseudo-R ²
Independency	18427358	8057151		-		-
100% Stayers	288695202	12499000	-4441849	-9	<.0001	-14.6667
NKO	478995.8	465193.5	7591957.5	81	<.0001	0.974006
LPM	300381.6	284876.9	7772274.1	81	<.0001	0.988832
LCA	205794.6	201438.4	7855712.6	81	<.0001	0.983699
IPF	1771672	1652404	6404747	81	<.0001	0.903856
Combo	321557.4	318446.1	7738704.9	81	<.0001	0.98255
Combo 2	321557.4	318446.1	7738704.9	81	<.0001	0.98255

Source: (Authors own calculations)

Table 5. Goodness of fit statistics for all models 2002-2003

Model	GFX ²	LRX ²	DLRX ²	Ddf	P	McFadden Pseudo-R ²
Independency	3776480	3580063		-		-
100% Stayers	2418201	2012710	1567353	-9	<.0001	0.3596679
NKO	393554.2	405704.9	3174358.1	81	<.0001	0.895788
LCA	157278.7	154362.7	3425700.3	81	<.0001	0.958353
LPM	98510.9	98410.4	3481652.6	81	<.0001	0.973915
IPF	1999259	1840727	1739336	81	<.0001	0.4706024
Combo	223587.7	221840.4	3358222.6	81	<.0001	0.9407947
Combo 2	223587.7	221840.4	3358222.6	81	<.0001	0.9407947

Source: (Authors own calculations)

Table 6. Goodness of fit statistics for all models 2003-2006

Model	GFX ²	LRX ²	DLRX ²	Ddf	P	McFadden Pseudo-R ²
Independency	4548314	3290676		-		
100% Stayers	2643817	1956575	1334101	-9	<.0001	0.4187
NKO	362460.2	342284.8	2948391.2	81	<.0001	0.9203
LCA	146497.4	139716.3	3150959.7	81	<.0001	0.9678
LPM	178298.9	172427.8	3118248.2	81	<.0001	0.9608
IPF	1856685	1715659	1575017	81	<.0001	0.5918
Combo	238607.4	223775.1	3066900.9	81	<.0001	0.9475
Combo 2	181484.6	175740.6	3114935.4	81	<.0001	0.9601

Source: (Authors own calculations)

In tables 4, 5 and 6 the goodness of fit statistics are shown for the elections 1998-2002, 2002-2003 and 2003-2006. The tables show the results of the benchmark models of independence and 100% stayers. Furthermore, the results of the NKO model are shown. The model based approaches of LCA, LPM=QP and IPF are also presented. Both combination models are presented as combo (replacing values on the diagonal) and combo 2 (replacing zero values).

All elections can be estimated quite well with all models. The elections of 1998-2002 have the highest values in terms of model fit. Only the 100% stayer model performs relatively poorly. The combination models in 1998-2002 and 2002-2003 do not outperform each other and produce identical results. This can be explained by the fact that the values of the NKO-matrix are closer to zero in 1998-2002 and 2002-2003 than in 2003-2006, and for these periods the confidence interval includes the value zero.

The big difference in results between LCA and IPF can be explained from the fact that the IPF algorithm has much heavier constraints on the calculation. The model based approach of LCA only slightly outperforms the QP approach. This is because LCA uses log-estimates to maximize the likelihood, whereas QP uses squared estimates to minimize the error. Log-estimates are better able to deal with values near 0, therefore giving slightly better results. The model based approaches show significant improvements to the NKO (except for IPF), and of course also to the independence and 100% stayer approaches.

The model that performs best is that of LCA, with a pseudo- R^2 of .98, .96 and .97. Given the goodness-of-fit statistics, this model is able to estimate most of the transitions and outperforms the NKO results. This is to be expected because LCA, in contrast to a survey, is designed to optimize the results.

4.2 Benchmark methods

The independence model and the 100% stayer model give an indication of the performance of the other models. They both produce a transition matrix that is calculated using the basic formula (2.1). However, it is not very useful to present these matrices here. The independency model produces a matrix with different columns but with identical rows, corresponding with total independency between elections. This is done on the basis of an initial marginal probability for the first election. The 100% stayer is a very simple matrix, which is the identity matrix with just ones on the diagonal.

Since there is no golden rule how to measure the success and interpret the results, these models are used as benchmarks. The values of the pseudo- R^2 are plotted on the Dutch map⁴. In this way deviations and other interesting facts become visible. Since the independency model is used to calculate the pseudo- R^2 , this model cannot be plotted in this way.

⁴ Remember that in formula 3 for 1998-2002 it is assumed that SGP=LPF and for 2003-2006 it is assumed that LPF=PVV. This only happens using this model.



Figure 2a. 100% stayer 98-02

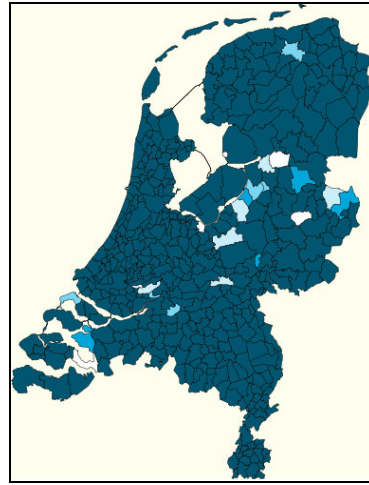


Figure 2b. 100% stayer 02-03

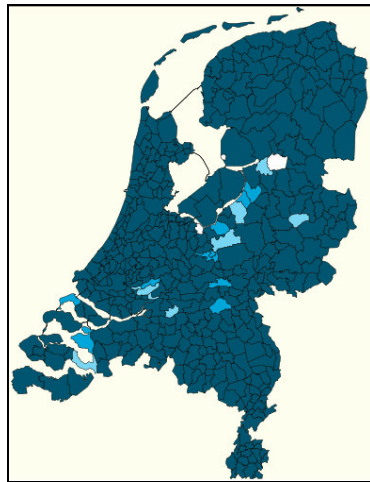


Figure 2c 100% stayer model 03-06

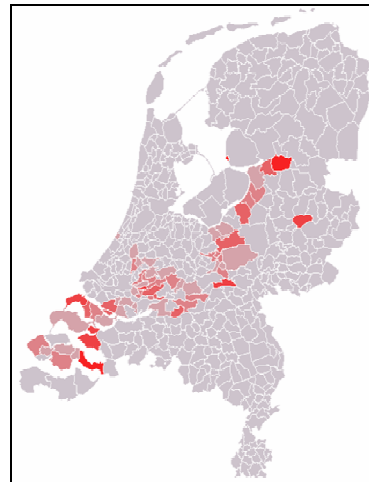


Figure 2d. Dutch biblebelt

(Kiesraad.nl, 2008)

Legend 1998-2002:	$R^2 < 0.9598$	$R^2 < 0.9702$	$R^2 < 0.9778$	$R^2 < 0.9839$	$R^2 < 1$
Legend 2002-2003:	$R^2 < 0.8376$	$R^2 < 0.8632$	$R^2 < 0.8873$	$R^2 < 0.9116$	$R^2 < 1$
Legend 2003-2006:	$R^2 < 0.8668$	$R^2 < 0.9245$	$R^2 < 0.9464$	$R^2 < 0.9622$	$R^2 < 1$

The legends show the pseudo-R statistics per municipality. The lighter the color the higher the corresponding pseudo-R and fit. They are based on the 20th, 40th, 60th and 80th percentile of the NKO pseudo-R values. These values are used to show the performance in terms of colors of all models. The first figure shows the fit of the 100% stayer model in the elections of 1998-2002, 2002-2003 and 2003-2006. As can be seen in figure 2a, the fit of the 100% stayer model in one of the most volatile elections in Dutch history is exceptionally bad.

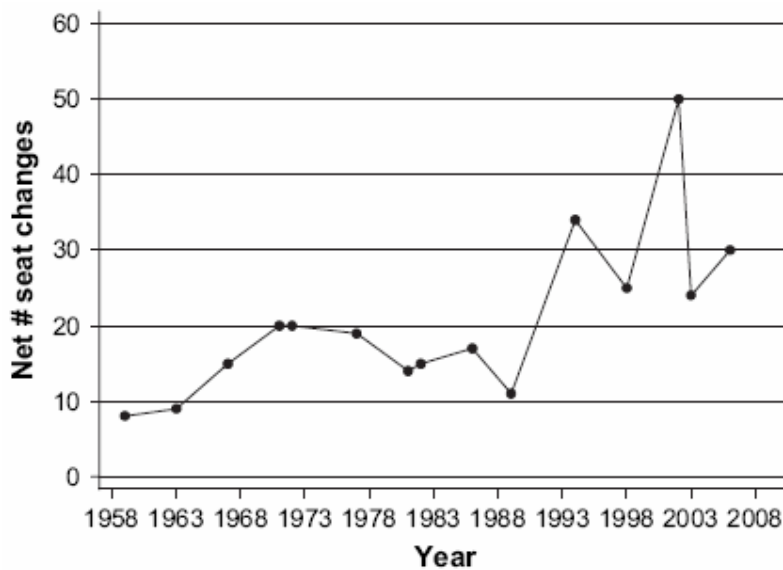


Figure 3. Volatility in seats for all Dutch parliamentary elections from 1959-2006 (Aarts & van der Kolk, 2007 pp. 837)

From figure 3 it is evident that the net number of changes in seats was the highest in the 2002 elections. Clearly the 100% stayer model is not applicable there. In the 2002-2003 and the 2003-2006 elections the 100% stayer model is better able to predict voting behavior from the election results of respectively 2002 and 2003. In 2003 we find very good predictions in Staphorst, Urk and Bunschoten (McF. pseudo $R^2 > .95$). The so called biblebelt is very well visible and shows that citizens in these municipalities are also very loyal voters (McF. pseudo $R^2 > .81$). The Dutch biblebelt is roughly found in the Zeeuwse islands, throughout the river area of the provinces South-Holland, Utrecht, Gelderland and partly through Northern-Brabant (Werkendam and Wijk en Aalburg), and in the most northern part of Overijssel. Especially places as Staphorst, Genemuiden, Nieuw Lekkerland, Elspeet, Opheusden, Kesteren, Barneveld, Ederveen, Oudorp, Tholen, Arnhem, Meliskerke, Aagtekerke, Yserke and Krabbendijke are in the middle of this area. Some other places lay outside this area but in these municipalities there is also a high concentration of reformed people. These municipalities are Urk, Rijssen, Scheveningen and Katwijk (Wikipedia: Bijbelgordel, 2008).

4.3 NKO

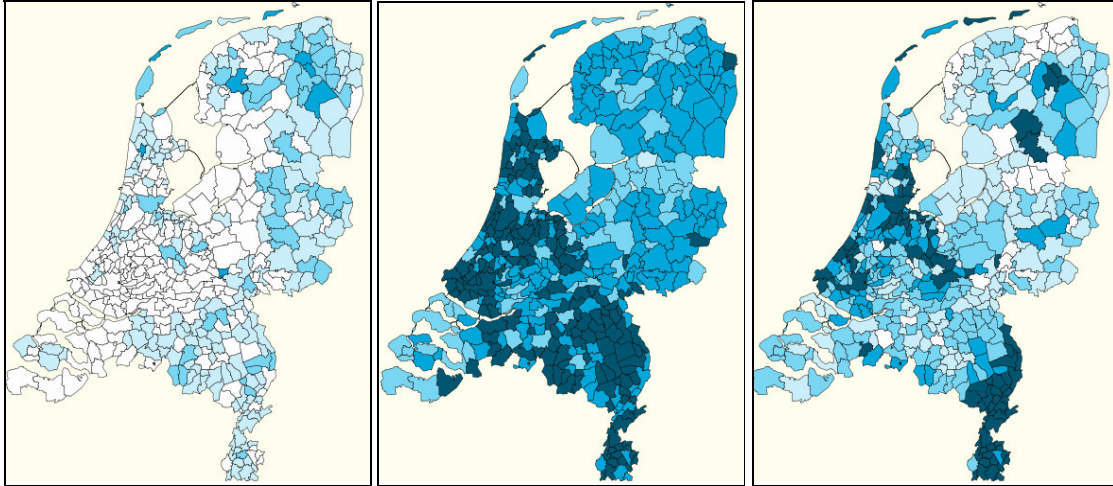


Figure 4a. NKO model 98-02

Figure 4b. NKO model 02-03

Figure 4c. NKO model 03-06

Legend: $R^2 < 0.8650$ $R^2 < 0.9163$ $R^2 < 0.9522$ $R^2 < 0.9728$ $R^2 < 1$

The legend used to compare the model performance through the three consecutive elections is also based on the 20th, 40th, 60th and 80th percentile of the NKO pseudo- R results. This time they have been averaged out over the three elections to create a comparable situation. The map of the Netherlands is a lot lighter when the NKO transition matrix is used compared to the 100% stayer model. Especially in Zeeland, Northern-Brabant and Overijssel the results are much better. But in the Randstad, large parts of Northern and Southern-Netherlands the results are still poor, making it interesting to look at the model based approaches to see if they can perform better.

4.4 Quadratic Programming

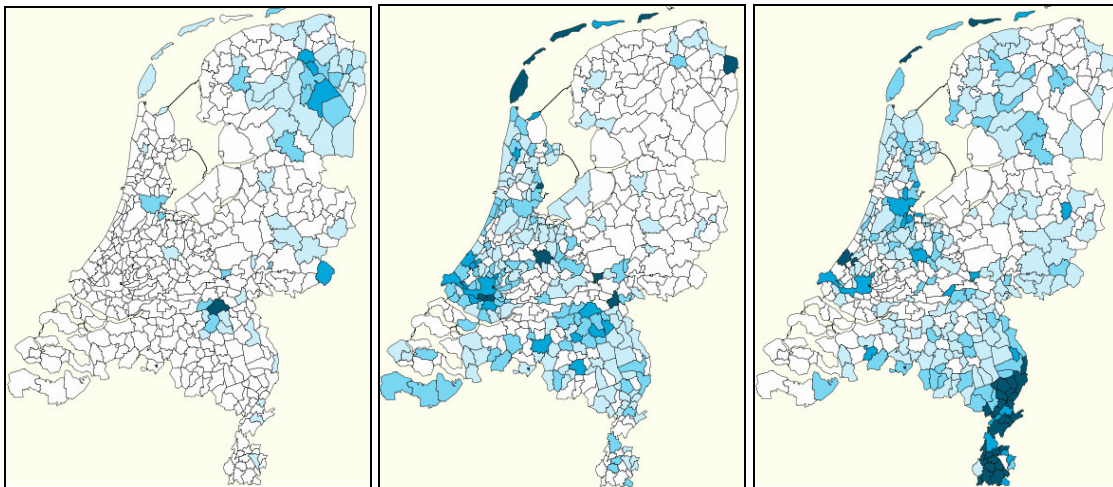


Figure 5a. LPM model 98-02

Figure 5b. LPM model 02-03

Figure 5c. LPM model 03-06

Legend: $R^2 < 0.8650$ $R^2 < 0.9163$ $R^2 < 0.9522$ $R^2 < 0.9728$ $R^2 < 1$

The LPM-model shows much more lighter areas for all elections using the same legend as the NKO-model. Still, the Randstad remains a general problem. There are no real examples to be found in which this model performs noticeably worse than other models or specific areas where this model is not able to estimate correct values.

4.5 Latent Class Analysis

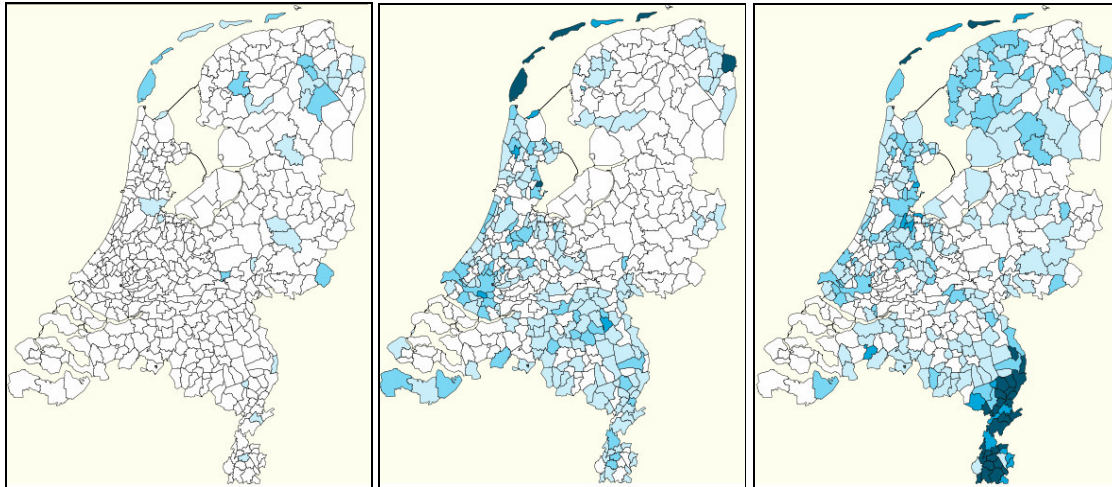


Figure 6a. LCA model 98-02

Figure 6b. LCA model 02-03

Figure 6c. LCA model 03-06

Legend: $R^2 < 0.8650$ $R^2 < 0.9163$ $R^2 < 0.9522$ $R^2 < 0.9728$ $R^2 < 1$

With the LCA model the map has even more lighter areas than with the LPM-model. Even this model does not handle the islands very well, as was also observed with all the other models.

4.6 Iterative Proportional Fitting

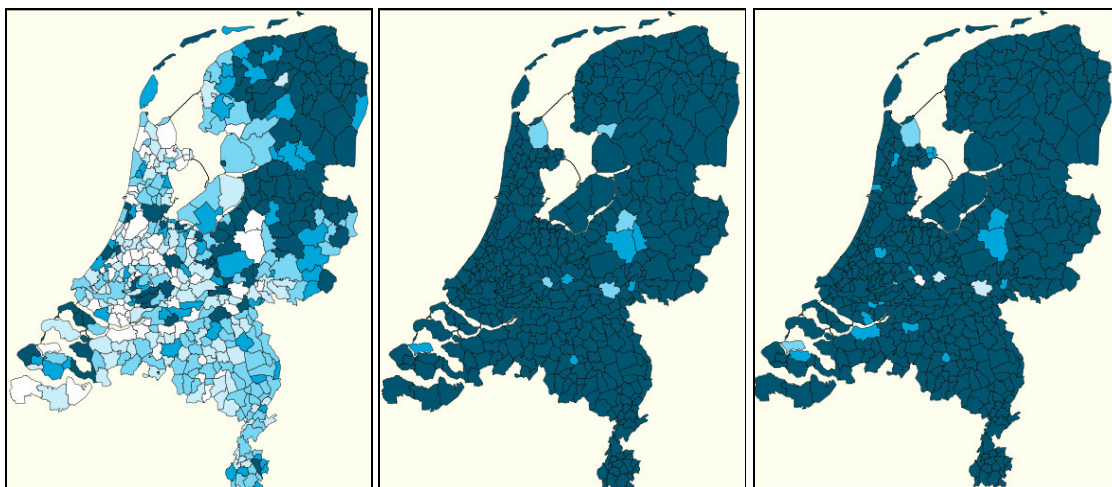


Figure 7a. IPF model 98-02

Figure 7b. IPF model 02-03

Figure 7c. IPF model 03-06

Legend: $R^2 < 0.8650$ $R^2 < 0.9163$ $R^2 < 0.9522$ $R^2 < 0.9728$ $R^2 < 1$

It is relatively difficult to see interesting facts on these maps for the IPF-model because of its general mediocre fit. It is quite obvious that some municipalities are colored light in all elections. These municipalities are Overbetuwe, Lemsterland, Vianen, Wieringermeer, Epe, Apeldoorn and Northern-Beveland. The performance of IPF is relatively bad because the model estimated under IPF (Formula (2.1)) is much more restricted than the other models discussed so far.

4.7 Combinations

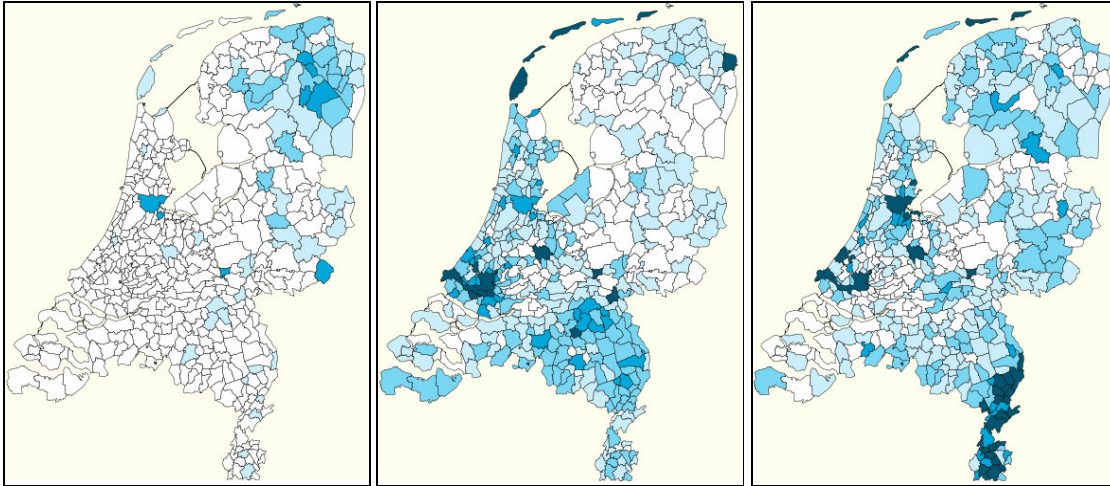


Figure 8a. Combo model 98-02

Figure 8b. Combo model 02-03

Figure 8c. Combo model 03-06

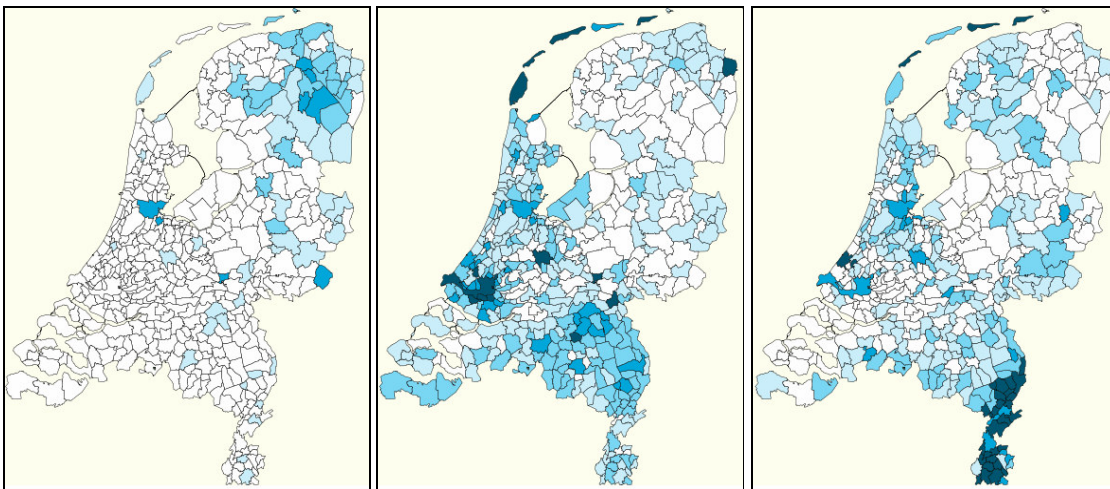


Figure 9a. Combo2 model 98-02

Figure 9b. Combo2 model 02-03

Figure 9c. Combo2 model 03-06

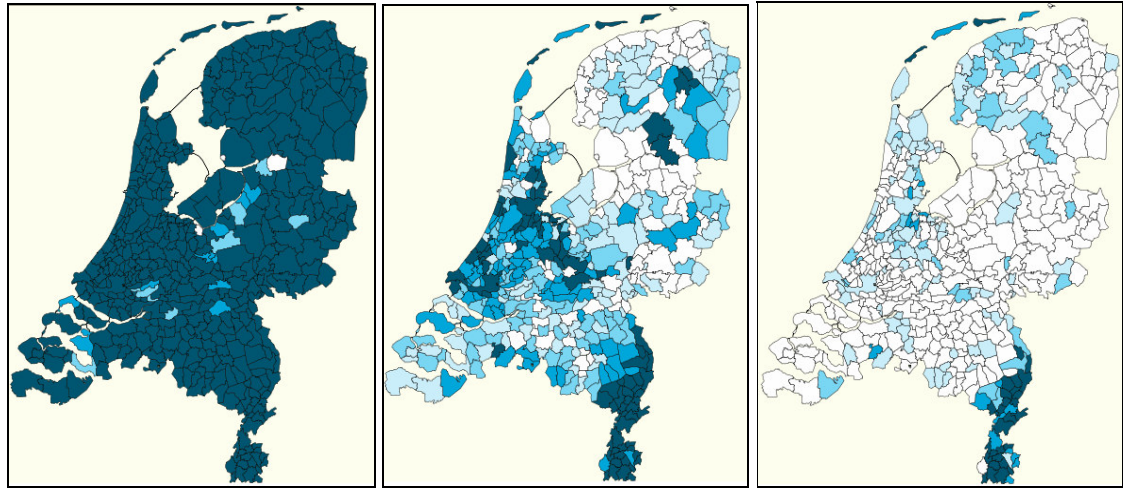
Legend: $R^2 < 0.8650$ $R^2 < 0.9163$ $R^2 < 0.9522$ $R^2 < 0.9728$ $R^2 < 1$

The pictures of the combination model only differ in 2006. The patterns in these figures show the general line of the LPM-model, with slightly worse fits.

4.8 Discussion about the elections 2003-2006

For all elections another analysis is made comparing all models to each other for one election. In this section we zoom in on the elections of 2003-2006 and only on the 100% stayer benchmark model, the NKO model and the best performing model based approach of LCA. The other elections and description of all models are described in appendix A section 7.3.

The reason why the simple linear regression analysis has not been presented here is because this model produces results outside the range between 0 and 1. Interestingly enough, most values that were above 1 relate to the islands. Especially Schiermonnikoog sticks out, which had values above 1 twice. Tubbergen is another municipality that pops up two times. All model based approaches have problems to correctly predict the voting transitions for the islands. The explanation for this observation lies within the fact that the islands are popular resorts for holidays. The population of the islands changes considerably depending on the tourists that are on holiday there. Therefore all models have difficulty dealing with this complication.



A similar reasoning applies for the postal voters. Also Limburg is in all models not estimated correctly. Further analysis shows that the support for the PVV in Limburg is systematically underestimated. Taking into account that the founder of the PVV is from this part of the Netherlands, one can easily understand why Limburg deviates from the national pattern.

Transition matrices

All models have a transition matrix, which is different for every model. We zoom in closer on the matrices of the elections of 2003-2006. The other transition matrices of

1998-2002, 2002-2003 and 2003-2006, including clarification can be found in section 7.2 of appendix A. From the benchmark models no transition matrices have been presented (see section 4.2).

Table 7. NKO transition matrix 2003-2006 (diagonal elements bold) (N=2528)*

2003	2006									
	CDA	PVDA	VVD	SP	GL	D66	CU	PVV	Anders	NG
CDA	.7072	.0339	.0578	.0637	.0040	.0040	.0398	.0199	.0100	.0598
PVDA	.0294	.5910	.0157	.2035	.0274	.0078	.0078	.0117	.0137	.0920
VVD	.2303	.0284	.5521	.0347	.0032	.0095	.0063	.0473	.0189	.0694
SP	.0379	.1061		.6970	.0606		.0152	.0379	.0227	.0227
GL	.0316	.0737	.0105	.2526	.4632	.0105	.0421		.0211	.0947
D66	.0814	.1744	.1744	.1512	.1163	.2326	.0233		.0465	
CU	.0222	.0222		.0222			.9111	.0222		
LPF	.0706	.0353	.1765	.1412		.0118		.3647	.0471	.1529
Anders	.0952	.0238	.0238	.0952	.0238		.0714	.0238	.5714	.0714
NG	.0600	.0622	.0267	.0867	.0044	.0044	.0022	.0467	.0111	.6956

Source: (Aarts *et al.*, 2007: pp. 224)

* There is some deviation from the values presented in Aarts *et al.*, (2007) due to rounding errors in the book.

This weighted matrix (correcting for various biases, see section 2.3) shows where the voters of 2003 have gone in 2006. For example, 71% of the voters who voted for the CDA in 2003 also did so in 2006, whereas 3% moved to the PvdA and 6% to the VVD etc. Also, 70% of the SP voters of 2003 stayed with the SP in 2006, but 11% voted for the PvdA instead. On the other hand 22% of the PvdA voters of 2003 voted for the SP in 2006, which is more then they have gained, as is depicted in the cartoon.



(Janssen, 2006)

From the relatively high values on the diagonal it is apparent that most voters were loyal to their party. They are so called stayers. D'66 has lost a lot of support and there are relatively little stayers in this party in comparison to other parties. It becomes also quite visible that a lot of former LPF voters have voted for the PVV in 2006.

Table 8. Transition matrix LCA model 2003-2006 (diagonal elements bold)

2003	2006									
	CDA	PVDA	VVD	SP	GL	D66	CU	PVV	Anders	NG
CDA	.8304		.0028	.0664			.0093	.0400		.0511
PVDA		.7419		.2028			.0154			.0399
VVD	.1679		.7889				.0225			.0206
SP				1.0000						
GL					.8394	.1606				
D66		.1700	.1451		.0748	.2799	.0075		.1979	.1248
CU							1.0000			
LPF		.0047				.0017		.5587	.2224	.2125
Anders							.1453		.8547	
NG		.0158		.1145				.0645		.8051

Source: (Van der Ploeg *et al.*, 2008)

Inspection of the LCA transition matrix shows that most voters are stayers. Particularly the SP, CDA, PvdA, CU, VDD and GL have high stayer rates. Someone who did not vote in 2003 often was also a non-voter in 2006 (81%). This is 6% higher than was estimated with the LPM model. The proportion of non-voters is a disputable piece of information, even with the model based techniques. You can still see that 11% of the non-voters in 2003 did vote for the SP in 2006, explaining partly the growth of the SP. The 2003 electorate of D'66 was fragmented over at least four parties. According to the LCA model, in 2006 over 55% of the previous LPF voters casted their vote in favor of the PVV and a large portion of former LPF voters remained at home. The SP and CU have values of 1 on their diagonal indicating that they have very loyal support.

4.9 Seating distributions 2003-2006

An interesting visualization of the performance of the models can be obtained by calculating the seating distribution in the Dutch second chamber. There are 150 seats to be distributed. The calculation of the seat distribution is as follows:

first, after counting every casted vote, the Quota to get a seat (Q) is calculated. All votes, except the non-valid ballots, are taken into consideration, giving the following formula:

$$Q := \frac{V}{150}$$

Where V is total number of valid votes.

It is possible for parties to register themselves together with another party and form a so called list combination. This has the advantage that the number of votes per combination is counted, and per list the number of times the quota is met is calculated to obtain the total number of seats earned by these parties. Also, with the distribution of the remaining seats, the list combination is viewed as one party. A list combination is only profitable when both parties meet the electoral threshold.

In the Netherlands the electoral threshold is equal to the quota. Meaning that the higher the electoral turnout, the higher the quota and therefore the higher the electoral threshold. First, the number of times that a party meets the quota is calculated resulting in number of whole seats gathered per party. The allocation of these whole seats (S) to the individual parties (I_i) is done as follows:

$$S := \sum_{i=1}^N \frac{I_i}{Q}$$

Where I is the total number of votes for individual party or list combination i , N is the total number of parties that have met the quota and S is whole number of seats.

After this calculation a limited number of seats remains, which are called remaining seats. They are determined by subtracting the number of whole seats from the total of 150 seats. The number of remaining seats (R) varies from 5 to 12 seats. The allocation of these remaining seats is done with the method of greatest average. averages for remaining seats (A) are calculated using the method of greatest average. The averages for the remaining seats (A) are then calculated by dividing the total number of votes per party or list combination by the number of whole seats+1,+2 and +3 (remaining seats):

$$A := \sum_{i=1}^N \sum_{j=1}^3 \frac{I_i}{S_i + j}$$

Where j is remaining number of seats and S_i is the whole number of seats per party.

The whole list of averages is then ranked and the highest R averages get an extra seat. It is thus possible that a party gets multiple remaining seats. The last step is to allocate seats to every individual party within a list combination. For this purpose a quota just for this list combination is calculated. The total number of votes of the list combination is divided by the total number of allocated seats for this list

combination. On the basis of the largest remainders the remaining seats within the list combination are divided.

Table 9 presents the translation of the voting results into the seating distributions of the last four elections. These real results of 2003 and 2006 are compared with the estimates of the models to show another visual fit.

Table 9. Seat distribution 1998-current

Party	1998	2002	2003	2006
Christian Democrats (CDA)	29	43	44	41
Labour Party (PvdA)	45	23	42	33
Socialist Party (SP)	5	9	9	25
Liberal Party (VVD)	38	24	28	22
Group for Freedom (PVV)	-	-	-	9
Green Left (GL)	11	10	8	7
Christian Union (CU)		4	3	6
Democrats 66 (D66)	14	7	6	3
Party for animals (PvdD)	-	-	-	2
Christian Reformed Party (SGP)	3	2	2	2
List Pim Fortuyn (LPF)	-	26	8	-
Reformed Political Union (GPV)	2	-	-	-
Reformational Political Federation (RPF)	3	-	-	-
Livable Netherlands (LN)	-	2	-	-
Total	150	150	150	150

(Source: CBS: statline)

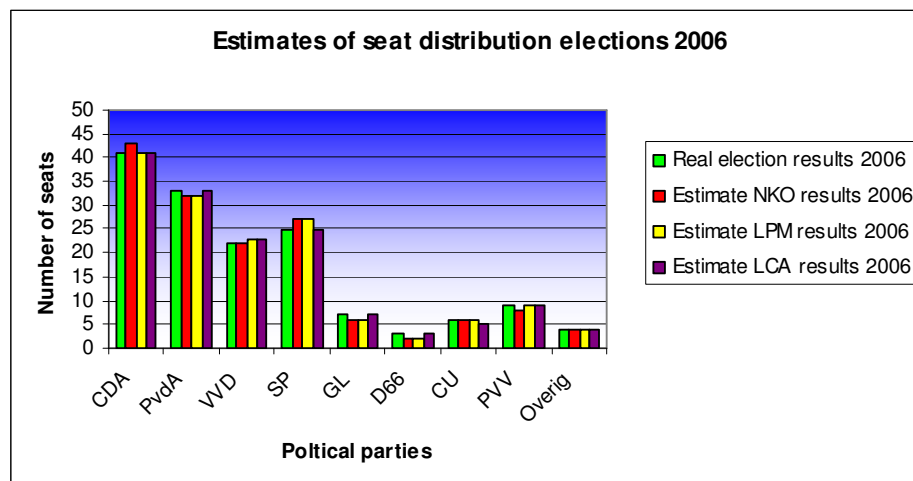


Figure 10. Estimations of seating distributions in 2006

Figure 10 shows the real seating distribution compared to the model based on survey results and the two best performing models of LPM and LCA. It becomes quite clear that the LCA estimations only deviate from the real election results on two occasions. The VVD gets one seat too much, whereas the CU one seat too little. We do note that in order to calculate these estimates the “other” categories was equalized with the real results and not with the estimations. Because in the remainder category all other parties are categorized (see section 7.1 in appendix A), parties who didn’t pass the electoral threshold are also included. Summing up these votes leads to an inaccuracy and to a large number of votes to be taken into consideration for the seating distribution. In order to correctly calculate the seating distribution, only the parties who have passed the electoral threshold should be taken into consideration. In order to not to complicate matters too much in this analysis the remainder category equals the number of seats gathered by the parties that are in the remainder category. In 2006 this is equal to 4 seats because both the SGP and PvdD have gotten 2 seats in the Dutch parliament.

5. Discussion and further research

The main research question that has been central in this paper:

To what extent can voting transitions in the Netherlands on municipality level be accurately described where the necessary transition matrices are lacking?

5.1 Ecological inference

As mentioned in the introduction an important issue with voting data is that it is on an aggregated level. Instead of knowing the voting preferences of every individual we only know the marginal totals. The aim of ecological methodology is to estimate the expected frequencies on the individual level. The research into aggregated statistical analysis has been long withstanding. Sociologists were the first to study this statistical problem (see e.g. Robinson, 1950; Goodman, 1953; Duncan & Davis, 1953). In 1950 heavy criticism by Robinson posed a serious problem for this type of analysis. His analysis showed that correlations on the aggregated level were much higher than the actual correlations on the individual level (Robinson, 1950). Goodman's (1953) article "Ecological Regression and Behavior of Individuals" opposed this notion. With some interesting techniques he showed that in certain circumstances it is possible to make inferences about individual behavior (Goodman, 1953). In the nineties new attention emerged with the development of new models (see e.g. Freedman et al., 1991; Grofman, 1991; King, 1997, Cho, 1998, Cho & Gaines, 2004, King, 2004). For an extensive overview of the methodology until 1995, see Cleave et al., (1995). The breakthrough by Gary King (1997) has given a new impulse to this area of research.

The largest problem with predicting on the individual- or cross-level using aggregated data is known as aggregation bias (Robinson, 1950). When the data is grouped on the dependent or independent variables, group induced correlation occurs. In order to resolve this problem one can assume that for every precinct/municipality etc. the voting rates are constant. This is also known as the constancy assumption and when this holds analysis is straightforward with Ordinary Least Squares (OLS). Parameters are constant but it is not very likely that they are not correlated with any regressor. Because it is not likely, that all voters vote the same regardless of demographical factors (Cho, 1998, pp. 2).

5.1.1 *EI estimator*

Gary King has written two important books on ecological inference in 1997 and 2004. According to King the solution to the ecological inference problem can be found in the random coefficient model. It improves the Goodman estimator and focuses on the very common out-of-bounds problem. To overcome this problem he uses the information from the bounds. The first stage of a two-stage procedure is the determination of the parameters which are the election results of a specific election across precincts/municipalities. He assumes that there is no dependency between the statistical behavior of a demographic group and the precinct they reside in (constancy assumption). In his model parameters vary according to a truncated bivariate normal distribution instead of a normal distribution which is used in OLS. These are conditioned to lie between 0 and 1. He uses the underlying bivariate normal distribution to calculate local estimates, which he then combines to create aggregate level estimations. He does this by addition over the different precincts (King, 1997). Still, this model performs just as poorly as OLS when aggregation bias is present (Cho, 2001). Park writes an interesting paper in 2004 using the Thomsen estimator in a 2 by 2 situation in which he outperforms the EI-estimator, ecological regression (Goodman-estimator) and constrained regression (Park, 2004). The EI model is most often used as a benchmark model in comparison with other methods.

Much of the research on ecological inference focuses on 2 by 2 systems, but recently more methodology has been developed for more general system such as multiparty systems.

Gibson and Cielecka (1995) propose separate OLS models for every party with demographic variables. They treat the separate parties as dependent variables. They use this model to explain the first Polish elections after the communist period. They compare different demographic variables to find correlations with party support and discover satisfactory explanations for the Polish development (Gibson and Cielecka, 1995). This paper is later used by several authors to test their model in a multiparty situation in comparison to this model.

Katz and King (1999) developed a different model for multiparty systems. It is especially suited for district-level aggregate election data. It is also appropriate to incorporate contextual variables. Most of the current literature deals with two party systems, making analysis much easier. In this instance they look specifically at the British elections of 1992 who have three parties that competed in these elections. They introduce a Markov Chain Monte Carlo (MCMC) approach. With this approach it is possible to analyze up to 8 parties. It is also applicable to incorporate characteristics of the aggregated areas. This method is better suited for missing data

then maximum likelihood estimation. The missing data are initially estimated and put into the data. The method of MCMC estimates, using maximum likelihood, the missing data iteratively until reaching the equilibrium distribution and convergence. The MCMC approach is still rather limited and in 2000 this model was extended to incorporate more parties (Katz and King, 1999). Lewis (2003) also uses the same algorithm to estimate voting behavior in a two party system but then over multiple elections. Implementing the MCMC approach like this makes not only analysis across precincts possible as in King (1997), but also across elections within precincts. He finds similar results as King (Lewis, 2003).

In 2001 Honaker, Katz and King extended the MCMC model to more than three parties. They use a full information maximum likelihood (FIML) model to estimate district-level data. This model is equivalent to the Least Squares regression in two party systems and uses the EM algorithm to approximate the results. They compare their new model to their own model and to the Gibson and Cielacka model. They find that their approach is more accurate, faster and can scale up to multiparty systems (Honaker *et al.*, 2001).

Wellhofer (2001), applies the EI-estimator successfully in the multiparty system of Italy for RxC contingency table (Wellhofer, 2001). This is in short an array formed by the intersection of two or more classification variables. The fields in the array are filled with frequencies of observations (Gunst, 2004). Wellhofer uses the survey results of the Italian Voting Survey as benchmark and combines these results with a political analysis.

The problem of ecological inference is an important problem and this study may add in a different way to the discussion about it. It is however not central to this study because we do not aim to make statements about individual behavior only about on voting behavior on municipality level.

5.2 Discussion

The most important assumption of the benchmark model of independence is that there is perfect mobility of a voter between two elections. In this model the probability to vote for a certain party on the second moment is independent of the choice of party on the first moment. Analysis shows that with this model the election results on the level of municipality cannot be predicted using this model. As expected, also the model that assumes that all voters are completely loyal does not reflect reality. Using a transition matrix that can be estimated from NKO data directly improves the model fit considerably. Model based estimation techniques improve these results even further. The results show that on the level of voting

transitions NKO and the model based estimation techniques are very similar. A great advantage of NKO is that it can map the actual mobility of voters. For each respondent it can be shown what party he voted. Therefore, transitions between different parties can be analyzed in both directions. Model based estimation techniques only reflect the nett-transitions between parties. This way, transitions between two parties often diminish, and consequently the transition matrix contains many zeros. This is closely related to the problem of ecological inference (King, 1997, 2004).

Aggregated data are usually easier to obtain than individual data, and can offer valuable indications about individual behavior. Ecological inferences will therefore still be made. The problems of confounding and aggregation bias will probably not be resolved in the near future (Freedman, 1999). Therefore, survey research is still the only way to obtain reliable results on the behavior of individuals.

An advantage of model based research is that it is a cheap technique compared to survey research; municipal results are always available. A second advantage is that with model based estimation techniques the total population of the Netherlands can be observed, while the NKO research is a sample survey in which the voting behavior is extrapolated to the total population. In this way, the model based research made it possible to observe interesting regional behavior. In the estimations it becomes visible what the influence is on the voting behavior of certain politicians in the regions where they come from. Also, the difficulty in explaining transitions on the Dutch islands becomes visible. Other purely election specific elements are visualized better. For instance, in 2006 the voting transitions of Limburg clearly deviated from other provinces.

The results of the model based techniques of Quadratic Programming and Latent Class Analysis are comparable and there is only a minor difference in model fit. The combination models also give a good fit but where slightly worse and did not show more realistic results.

The most important conclusion of this research is that the different methods (sample survey and model based) are in a way complementary. Results of the methods can be confirmed by each other, improving the validity of the results. Furthermore, this research shows that the assumption that there is a single transition matrix that is the basis of the voting behavior of all Dutch voters is not sufficiently sustainable. Future research will have to be done to investigate the influence of fine tuning of the models for obtaining the transition matrices. Possible directions of research are: firstly, using different matrices for different regions or for different degrees of urbanization. Secondly, other options such as Self Organizing Maps and Stochastic

global search methods, such as using survival of the fittest within genetic algorithms, may prove to be valuable optimization techniques. Self Organizing Maps is a technique within artificial neural networks. This technique can be used to make logical groups of municipalities which can be used to make better a division in regions and to make multiple transition matrices. These techniques are in the field of Artificial Intelligence which is a different field of research. Thirdly, it could be interesting to define clusters on the basis of electoral support for a certain influential party. This technique may be more applicable for 2 or 3 party systems, but it would also be possible to be applied in multi-party systems.

6. Bibliography

- Aarts, K., H. van der Kolk & M. Rosema (2007). *Een verdeeld Electoraat De Tweede Kamerverkiezingen van 2006*, Spectrum.
- Aarts, K. & H. van der Kolk (2007). The parliamentary election in the Netherlands, 22 November 2006, Notes on Recent Elections/Electoral Studies **26** : 797-837.
- Agresti, A. (1996). *An introduction to categorical data analysis*, John Wiley & Sons, New York.
- Barbosa, M. F. & H. Goldstein (2000). Discrete Response Multilevel Models for Repeated Measures : An Application to Voting Intentions Data, *Quality & Quantity* **34**: 323-33.
- Bishop, Y.M.M., S.E. Fienberg & P.W. Holland (1975). *Discrete multivariate analysis. Theory and practice*. Cambridge: MIT Press.
- Betlehem J.G. & H.M.P. Kersten (1986). *Werken met Nonresponse (Working with Nonresponse)*. Voorburg: Centraal Bureau voor de Statistiek.
- Blumen, J., M. Kogan, & P. J. McCarthy (1955). *The Industrial Mobility of Labor as a Probabilities Process*, Cornell Studies of Industrial and Labor Relations (6). Ithaca, New York: Cornell University Press.
- Cameron, A.C. & F.A.G. Windmeijer (1993). Deviance Based R-Squared Measures of Goodness of Fit for Generalized Linear Models, Working Papers in Economics and Econometrics.
- CBS: Statline (2008). Overheid en politiek, Verkiezingen en politiek, Tweede Kamerverkiezingen, URL: <http://statline.cbs.nl/StatWeb/dome/?LA=NL> (Last visited on: 04 June 2008).
- Cho, W. K. T. (1998). If the assumptions fits...:a comment on the King ecological inference solution, *Political Analysis* **7**: 143–163.
- Cho, W. K. T. & A. H. Yoon (2001). Strange Bedfellows: Politics, Courts, and Statistics: Statistical Expert Testimony in Voting Rights Cases, *Cornell Journal of Law and Public Policy*, **102**: 237-64.
- Cho, W. K. T. & Gaines, B. J. (2004). The limits of ecological inference: The case of split-ticket voting, *American Journal of Political Science*, **48(1)**: 152–171.

- Cleave, N., P.J. Brown and C.D. Payne (1995), Evaluation of Methods for Ecological Inference, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **158**: 55-72.
- Clogg, C.C. & L.A. Goodman (1985). Simultaneous latent structure analysis in several groups. In: N.B. Tuma, ed., *Sociological Methodology 1985*. San Francisco, Jossey-Bass, :81-11.
- Cook, R.J., J. D. Kalbfleisch & G. Y. Yi (2002). A Generalized Mover-Stayer Model for Panel Data. *Biostatistics*, **3**: 407-420.
- Dempster, A. P., N.M. Laird & D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological*, **39**: 1 - 37.
- Duncan, O. D. and Davis, B. (1953). An alternative to ecological correlation. *American Sociological Review*, **18**: 665–666.
- Egmond, van M., N.D. de Graaf & C. van der Eijk (1998). Electoral participation in the Netherlands: Individual and contextual influences, *European Journal of Political Research* **34**: 281-300.
- Freedman, D. A., Klein, S. P., Sacks, J., Smyth, C. A., & C. G. Everett (1991). Ecological regression and voting rights (with discussion). *Evaluation Review*, **15**: 673–816.
- Freedman, D.A. (1999). Ecological Inference and the Ecological Fallacy, *International Encyclopedia of the Social & Behavioral Sciences*, Ed. Smelser N.J. & P.B. Baltes Technical Report, **549(6)** :4027-403.
- Gibson, J. & A. Cielecka (1995). Economic Influences on the Political Support for Market Reforms in Post-Communist Transitions: Some Evidence from the 1993 Polish Parliamentary Elections, *Europe-Asia Studies*, **47(5)**: 765-785.
- Greenwald, A., C., Carnot, R. Beach & B. Young (1987). Increasing voting behavior by asking people if they expect to vote, *Journal of Applied Psychology*, **72**: 315-318.
- Goodman, L. (1953). Ecological regressions and behavior of individuals. *American Sociological Review*, **18**: 663–666.
- Goodman, L. A. (1961). Statistical methods for the mover-stayer model, *Journal of the American Statistical Association*, **56(296)**: 841-868.

- Goodman, L. A., (1968). The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing values, *Journal American Statistical Association*, **63**: 1091-1131.
- Grofman, B. (1991). Statistics without substance: A critique of Freedman et al. and Clark and Morrison, *Evaluation Review*, **15(6)**: 746–769.
- Gunst, M.C.M. (2004). *Statistical Models*, Vrije Universiteit Amsterdam.
- Honaker, J., J.N. Katz & G. King (2001). An Improved Statistical Model for Multiparty Electoral Data, Working Papers 1111, California Institute of Technology, Division of the Humanities and Social Sciences.
- Irwin, G. A. & D. A. Meeter (1969). Building voter transition models from aggregate data, *Midwest Journal Political Science*, **13**: 545-566.
- Janssen, T. (2006). “Wouter Bos-PvdA”, URL: <http://www.tomjanssen.net/> (Last visited on: 23 April, 2008).
- Judge, G. G. & Takayama, T. (1966). Inequality restrictions in regression analysis. *Journal American Statistical Association*, **61**: 166-181.
- Kampen, J. & M. Swyngedouw (2000). The ordinal controversy revisited, *Quality and Quantity* **34(1)**: 87–102.
- Katz, J.N. & G. King (1999). A Statistical Model for Multiparty Electoral Data, *The American Political Science Review*, **93(1)**: 15-32.
- Kiesraad.nl (2008). Kiesraad.nl – Verkiezingsuitslagen, URL: <http://www.verkiezingsuitslagen.nl> (Last visited on: 17 January 2008).
- Keller, W.J. & A. ten Cate (1977). De verschuiving van de kiezersvoorkeur, *Economische Statistische Berichten* 26-10-1977.
- King, G., (1997). *A Solution to the Ecological Inference Problem. Reconstructing Individual Behavior from Aggregate Data*, Cambridge MA, Cambridge University Press.
- King, G., O. Rosene & M. Tanner (2004). *Ecological Inference. New Methodological Strategies*, Cambridge University Press, Cambridge.
- King, G.O. (1996). “Elections and the National Election Studies”, Paper prepared for the National Election Studies, Congressional Elections Conference.
- Langeheime, R. & F. van de Pol (1990). Discrete Time Mixed Markov Latent Class Models, *Sociological Methodology*, **20**: 213-247.

- Lewis, J.B. (2003). Extending King's Ecological Inference Model to Multiple Elections using Markov Chain Monte Carlo, Chapter in Gary King, Ori Rosen, and Martin Tanner, Eds. *Ecological Inference: New Methodological Strategies*. Cambridge: Cambridge University Press. 2004.
- Little, R.J.A & D.B. Rubin (1987). *Statistical inference with missing data*. New York: Wiley.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, LA: SAGE.
- McCarthy, C. & T. M. Ryan (1977). Estimation of voter transition probabilities from the British general elections of 1974, *Journal of the Royal Statistical Society Series A* **140**: 78-85.
- McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior, in Zarembka, P. (ed.), *Frontiers in Econometrics*, pp.105-142, Academic Press, New York.
- Nationaal Kiezersonderzoek (2002/2003). Dutch Parliamentary Election Study 2002/2003 (NKO 2002/2003), DPES 2002/2003.
- Oosterhoff, J. & A.W. van der Vaart (2003). *Algemene Statistiek*, Vrije Universiteit Amsterdam.
- Park, W-H. (2004). "Estimation of Voter Transition Rates and Ecological Inference", Presented at the 2003 Midwest Political Science Association Annual Conference. Chicago, April 2003.
- Ploeg, C.E. van der, F. van de Pol en J.K. Kampen, (2008), De verschuivingen van de partijvoorkeur van de Nederlandse kiezers tussen de nationale verkiezingen van 2003 en 2006, Offered for publication in: Het Nationaal Kiezersonderzoek 2006: opzet, uitvoering en resultaten, Schmeets, H. (ed.), CBS Nederland.
- Pol, van de F. & R. Langeheime (2004). *Latent Transition and Markov Models*, Statistics Netherlands.
- Poulsen. C. S. (1982). "Latent Structure Analysis with Choice Modeling Applications." Ph.D. diss., Wharton School, University of Pennsylvania.
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, **15**: 351-357.
- Shachar, R. & M. Shamir, (1996), Estimating Vote Persistence Sources without Panel Data, *Political Analysis*, **6**: 107-124.

- Staatsblad (2003). "Wet op het Centraal Bureau voor de Statistiek" van 20 november 2003, :516.
- Thomsen, S.R. (1987). *Danish Elections 1920–79: A Logit Approach to Ecological Analysis and Inference*. Aarhus: Politica.
- Thomsen, S.R. (2004). "Individual voting behavior in Denmark 1998-2001", Paper presented at the annual meeting of The Midwest Political Science Association, Chicago, April 2004.
- Tijms, H.C. & A.A.N. Ridder, (2003), *Mathematisch Programmeren*, Vrije Universiteit Amsterdam.
- Quinn, K.V., A.D. Martin & A.B. Whitford (1999). Voter Choice in Multi-Party Democracies: A Test of Competing Theories and Models, *American Journal of Political Science*, **43(4)**: 1231-1247.
- Upton, G.J.G. (1977). A Memory Model for Voting Transitions in British Elections, *Journal of the Royal Statistical Society Series A*, **140**: 86-94.
- Veall, M.R. & K.L. Zimmermann (1996). Pseudo- R^2 Measures for Some Common Limited Dependent Variable Models, *Sonderforschungsbereich*, **386**, paper 18, Institut für Statistik.
- Vermunt, J.K., R. Langeheime & U. Bockenholt (1999). Discrete-time discrete-state latent Markov models with time-constant and time varying covariates, *Journal of Educational and Behavioral Statistics*, **24**: 178-205.
- Voogt, R. (2004). *I'm not interested- Nonresponsive bias, response bias and stimulus effects in election research*, Universiteit van Amsterdam.
- Weir, B.T. (1975). The Distortion of Voter Recall, *American Journal of Political Science*, **14(1)**: 53-62.
- Wellhofer, E. S. (2001). Party realignment and voter transition in Italy, 1987-1996, *Comparative Political Studies*, **34(2)**: 156-86.
- Wiggins, L. M. (1955) 1973. Panel Analysis, Latent Probability Models for Attitude and Behavior Processes, *Elsevier*, Amsterdam.
- Wikipedia: Bijbelgordel (2008). Bijbelgordel (Bible belt) URL: <http://nl.wikipedia.org/wiki/Bijbelgordel> (last updated: 3 may 2008).

7. Appendix A

7.1 Analysis of parties

For 1998:

Cristian Democrats (CDA), Labour Party (PvdA), Liberal Party (VVD), Socialist Party (SP), Green Left (GL), Democrats 66 (D66), Reformed Political Union and Reformational Political Federation (GPV and RPF=CU), Christian Reformed Party (SGP), Other (Centre Democrats (CD), General Old People Union/ Union 55+ (AOV/U55+), Netherlands Mobile (NMob), Seniors 2000 (S2000), Dutch Middle Class Party (NMP), The Greens (Groenen), Nature law Party (NWP), Catholic Political Party (KPP), Free Indian Party (VIP), New United Old People Union (NSOV), New Communist Party (NCPN), Idealists/You (IdeA/JIJ), The Voters collective (KColl)), Not voted. Total is 10 parties.

For 2002:

Christian Democrats (CDA), Labour Party (PvdA), Liberal Party (VVD), Socialist Party (SP), Green Left (GL), Democrats 66 (D66), Christian Union (CU), List Pim Fortuyn (LPF), Other (Christian Reformed Party (SGP), Livable Netherlands (LN), United Seniors Party (VSP), Free Indian Party & Old People Union (VIP/OU), Durable Netherlands (DN), Party of the Future (PvdT), New Centre Party (NMP), Republican Peoples Party (RVP)), Not voted. Total is 10 parties.

For 2003:

Christian Democrats (CDA), Labour Party (PvdA), Liberal Party (VVD), Socialist Party (SP), Green Left (GL), Democrats 66 (D66), Christian Union (CU), List Pim Fortuyn (LPF), Other (Party for Animals (PVDD), Christian Reformed Party (SGP), Alliance of Renewal and Democracy (AVD), Conservatives.nl (Conservatieven), Durable Nederland (DN), Livable Netherlands (LN), List Veldhoen (Veldhoen), New Communist Party (NCPN), Party of the Future (PvdT), List Ratelband (Ratelb), Progressive Integration Party (VIP)), Not voted. Total is 10 parties.

For 2006:

Cristian Democrats (CDA), Labour Party (PvdA), Liberal Party (VVD), Socialist Party (SP), Green Left (GL), Democrats 66 (D66), Cristian Union (CU), List Pim Fortuyn (LPF), Other (Party for Animals (PVDD), Christian Reformed Party (SGP), List Five Fortuyn (Fortuyn=LPF), Netherlands Transparent (NT), OneNL (EenNL), List Poortman, Party for the Netherlands (PVN), Continuous Direct Democratic Party (CDDP), Liberal Democratic Party (LDP), United Senior Party (VSP), Ad Bos

Collective (Ad Bos), Green Free Internet Party (GVIP), List Potmis=Islam Democrats (ID), Tamara's Open Party (TOP), Solid Multicultural Party (SMP), LRV-Seat in Parliament (LRVP)), Not voted. Total is 10 parties.

7.2 Transition matrices 1998-2002, 2003-2006

All transition matrices from all models from the elections 1998-2002 and 2002-2003 are presented below.

1998-2002

Table 10 . Transition matrix NKO model 1998-2002 (diagonal elements bold)

1998	2002									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	.7602	.0045	.0317	.0090	.0090	.0136	.0136	.0860	.0090	.0633
PVDA	.1329	.4860	.0455	.0559	.0734	.0245		.0874	.0105	.0839
VVD	.1752	.0109	.5109	.0109	.0036	.0109	.0036	.2482	.0255	
SP	.0250	.0750		.5000	.0250			.2500		.1250
GL	.0400	.0800	.0267	.1333	.5067	.0533		.1067		.0533
D66	.1511	.1223	.1007	.0576	.0719	.2158		.0576	.0072	.2158
CU	.3409		.0227	.0227			.5455	.0682		
SGP	.0625								.9375	
Anders	.1628	.0698	.1395	.0465	.0930	.0233	.0233	.1395	.0930	.2093
NG	.0863	.0288	.0144	.0288	.0144	.0072	.0072	.1475	.0036	.6619

Source: (Nationaal Kiezersonderzoek, 2002/2003)

Table 11. Transition matrix LR model 1998-2002 (diagonal elements bold)

1998	2002									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	1.3062	-.1981	.0431	.0258	.0238	.0147	-.0053	-.3159	.0093	.0025
PVDA	.0404	.6233	.0189	-.0102	.0199	-.0007	.0111	.2130	.0234	.0438
VVD	.2844	-.1525	.6047	-.0381	.0108	.0342	-.0164	.6647	.0646	-.0789
SP	.3817	-.4568	-.0864	.6706	.1180	.0306	-.0102	.5405	.0026	-.0463
GL	-.1002	.1678	-.0383	.4966	1.0296	.0684	.0157	-.2921	.0924	.0318
D66	-.1055	.4809	.4405	.2192	.1002	.5442	.0222	-1.1844	-.1227	-.0203
CU	.0333	.2889	-.0260	-.0150	.0115	-.0233	.8810	.1450	.0012	.0431
SGP	.1885	-.1185	-.1463	-.0151	-.0404	-.0370	-.0637	.1089	1.0155	-.1511
Anders	.3863	-.2061	-.8131	.4308	-.5399	-.2416	.0711	.7814	.4833	.5230
NG	-.0708	.0544	.0042	-.0553	-.0175	-.0012	-.0083	.2678	-.0236	.7840

Source: (Authors own calculations)

Table 12. Transition matrix QP model 1998-2002 (diagonal elements bold)

1998	2002									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	1.0000									
PVDA	.0821	.5237	.025	.1069	.0615	.044	.0165	.0261	.0716	.0427
VVD	.2004		.4768					.30000	.0229	
SP										1.0000
GL				.3093	.6907					
D66			.5143		.0312	.4545				
CU	.3398						.6602			
SGP							.0179	.0289	.9531	
Anders								1.0000		
NG	.1332			.0175				.2003		.6489

Source: (Authors own calculations)

Table 13. Transition matrix LCA model 1998-2002 (diagonal elements bold)

1998	2002									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	1.0000									
PVDA	.1668	.5291	.0917	.0314	.0279	.0230	.0014		.0242	.1046
VVD	.2045		.4997					.2733	.0226	
SP				.7349	.0830	.0074				.1747
GL		.1543		.1522	.6930		.0005			
D66		.0007	.1831	.1706	.1465	.4992				
CU	.1920						.7821		.0259	
SGP	.0548							.0363	.9089	
Anders								.6984	.3016	
NG	.0163			.0066		.0089	.0023	.2585	.0101	.6973

Source: (Private communication from F. van de Pol)

Table 14. Transition matrix IPF model 1998-2002 (diagonal elements bold)

1998	2002									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	.2765	.1078	.1252	.0396	.0446	.0360	.0242	.1295	.0375	.1789
PVDA	.2591	.1146	.1273	.0411	.0465	.0372	.0223	.1316	.0366	.1837
VVD	.2594	.1097	.1341	.0404	.0462	.0383	.0217	.1346	.0374	.1782
SP	.2571	.1130	.1250	.0439	.0472	.0372	.0202	.1328	.0341	.1895
GL	.2566	.1134	.1299	.0420	.0486	.0386	.0210	.1322	.0346	.1831
D66	.2565	.1121	.1338	.0415	.0479	.0391	.0213	.1331	.0355	.1792
CU	.2696	.1090	.1191	.0362	.0412	.0336	.0403	.1256	.0555	.1699
SGP	.2682	.1018	.1175	.0330	.0359	.0311	.0415	.1306	.0761	.1644
Anders	.2587	.1104	.1282	.0410	.0462	.0371	.0214	.1348	.0370	.1853
NG	.2593	.1110	.1261	.0410	.0461	.0369	.0219	.1332	.0368	.1878

Source: (Authors own calculations)

Table 15. Transition matrix Combination 1 model 1998-2002 (diagonal elements bold)

1998	2002									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	.8806		.0199	.0080		.0059	.0411	.0174	.0272	
PVDA	.1511	.5236	.0074	.0758	.0724	.038		.0227	.0539	.0551
VVD	.2095		.4430			.0279		.2991	.0205	
SP				.4140						.5860
GL				.3108	.6210	.0681				
D66			.6274	.0061	.0445	.3220				
CU	.5060						.4940			
SGP							.0889		.9111	
Anders								.9797	.0203	
NG	.1284							.1958		.6758

Source: (Nationaal Kiezersonderzoek, 2002/2003 and authors own calculations)

Table 16. Transition matrix Combination 2 model 1998-2002 (diagonal elements bold)

1998	2002									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	.8806		.0199	.0080		.0059	.0411	.0174	.0272	
PVDA	.1511	.5236	.0074	.0758	.0724	.038		.0227	.0539	.0551
VVD	.2095		.4430			.0279		.2991	.0205	
SP				.4140						.5860
GL				.3108	.6210	.0681				
D66			.6274	.0061	.0445	.3220				
CU	.5060						.4940			
SGP							.0889		.9111	
Anders								.9797	.0203	
NG	.1284							.1958		.6758

Source: (Nationaal Kiezersonderzoek, 2002/2003 and authors own calculations)

As can be seen from these transition matrices SGP and LPF occupy the same position in the matrix. This leads to a very low value or zero value on the diagonal for the transition between SGP and LPF. Minor problems arise with the stayer model, which assumes that every voter who has voted SGP before will vote for this

party again. However since the LPF is a new party in 2002 the stayer model cannot incorporate this in any way leading to an enormous misfit.

All models show a big loss for the PvdA, which is consistent with the election results. It is interesting to see that the other winner of the elections in 2002, the CDA has a lower value on the diagonal of the NKO-matrix in comparison to the values on the diagonal of the model based matrices. This can again be explained by the ability of survey research to present transitions between parties and not only the nett-transitions as is the case with model based approaches. Because it is difficult to present the transitions differently one must derive the support for the LPF from the other category and not-voted in the previous election category. It is often said that Pim Fortuyn attracted non-voters to vote and this is indeed visible in the transition matrices and in the turn-out levels of that election year.

Something odd presents itself in the QP-matrix with the SP, which I cannot explain. Also both combination matrices are the same in this case. This can be explained by the fact that the optimization algorithm tries to find an optimal solution given the constraints. In this case these constraints are not restrictive enough.

2002-2003

Table 17. Transition matrix NKO model 2002-2003 (diagonal elements bold)

2002	2003									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	.7647	.0623	.0623	.0035	.0069	.0069	.0069	.0069	.0104	.0692
PVDA	.0415	.8238	.0104	.0259	.0207	.0104				.0674
VVD	.0476	.0595	.7321	.0060	.0060	.0060		.0060		.1369
SP		.1719		.6719	.0469	.0469				.0625
GL	.0135	.3649		.1351	.4324	.0135	.0135			.0270
D66	.0566	.2830	.0566		.0377	.5660				
CU	.2188	.0625			.0625		.6563			
LPF	.1227	.1104	.1411	.0123	.0123	.0061		.2822	.0307	.2822
Anders	.1333	.1000	.0667					.0667	.5667	.0667
NG	.0842	.0632	.0526	.0158	.0158	.0053	.0053	.0211	.0158	.7211

Source: (Nationaal kiezersonderzoek, 2002/2003)

Table 18. Transition matrix LR model 2002-2003 (diagonal elements bold)

2002	2003									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	1.0338	.1171	-.0098	.0487	-.0362	-.0322	.0031	-.0418	-.0222	-.0747
PVDA	-.1405	1.4444	-.0942	-.0654	-.0079	-.0087	.0005	.0579	.0098	-.2459
VVD	.1601	.0283	1.1449	-.0028	-.0225	.0043	-.0016	-.1260	-.0546	-.0150
SP	.0799	.0748	.0028	1.0777	-.0099	-.0360	-.0075	-.0916	-.0269	-.1552
GL	.1284	.1669	-.0243	.0520	.7543	.0567	.0226	-.0270	-.0386	.1491
D66	-.4574	-.3375	.1099	-.2421	.2289	.9003	-.0207	.0795	.1073	.3610
CU	.0114	-.0587	.0412	-.0050	.0311	.0210	.8825	.0597	.0273	.0809
LPF	-.0526	.0774	.1590	.0279	.0042	.0185	.0100	.5654	.0145	.2131
Anders	.2460	-.0983	.0305	-.0666	.0393	.0145	-.0463	-.0371	.9143	-.0312
NG	.0185	.0605	-.0410	.0164	-.0005	-.0046	-.0048	-.0449	.0028	.9753

Source: (Authors own calculations)

Table 19. Transition matrix QP model 2002-2003 (diagonal elements bold)

2002	2003									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	1.0000									
PVDA		1.0000								
VVD			1.0000							
SP		.3637		.566						.0703
GL		.7137			.2863					
D66					.3237	.6763				
CU	.1433	.0828					.7738			
LPF	.0348		.1685	.1059	.0623	.0304	.0068	.3222		.2692
Anders	.2206		.0101				.0127	.0117	.7449	
NG		.2334		.0434						.7232

Source: (Authors own calculations)

Table 20. Transition matrix LCA model 2002-2003 (diagonal elements bold)

2002	2003									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	.9960			.0040						
PVDA		1.0000								
VVD			1.0000							
SP		.0970		.9020						.0010
GL		.2940			.6640					.0420
D66		.1010			.0960	.8040				
CU	.0480	.0930			.0080		.8510			
LPF		.2120	.1570	.0090	.0040	.0020	.0010	.3430		.2720
Anders	.2500		.0220						.7280	
NG		.1990		.0300						.7700

Source: (Private communication from F. van de Pol)

Table 21. Transition matrix IPF model 2002-2003 (diagonal elements bold)

2002	2003									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	.2900	.1978	.1486	.0442	.0327	.0285	.0198	.0424	.0285	.1675
PVDA	.2680	.2130	.1498	.0454	.0344	.0295	.0193	.0426	.0270	.1711
VVD	.2738	.2018	.1587	.0444	.0343	.0308	.0179	.0437	.0269	.1677
SP	.2709	.2073	.1510	.0473	.0348	.0300	.0175	.0430	.0253	.1728
GL	.2694	.2084	.1525	.0463	.0358	.0307	.0177	.0426	.0247	.1718
D66	.2709	.2048	.1566	.0454	.0352	.0312	.0176	.0433	.0255	.1694
CU	.2891	.1972	.1400	.0398	.0311	.0267	.0332	.0411	.0434	.1586
LPF	.2749	.2013	.1534	.0447	.0334	.0295	.0182	.0447	.0282	.1716
Anders	.2833	.1920	.1460	.0404	.0304	.0270	.0268	.0437	.0482	.1621
NG	.2738	.2059	.1478	.0458	.0336	.0288	.0182	.0433	.0268	.1759

Source: (Authors own calculations)

Table 22. Transition matrix Combination 1 model 2002-2003 (diagonal elements bold)

2002	2003									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	.8858	.0157	.0179	.0349	.0153	.0036	.0145	.0107	.0015	
PVDA		.9462		.0173	.0192					.0173
VVD			.8529		.0672	.0800				
SP		.4086		.5914						
GL		.6553			.3447					
D66		.013	.4296		.0748	.4825				
CU	.3460						.5755		.0785	
LPF	.2273		.154	.0329				.3037		.2821
Anders	.2370	.0146	.0117				.0534		.6832	
NG		.2518		.0291						.719

Source: (Nationaal Kiezersonderzoek, 2002/2003 and authors own calculations)

Table 23. Transition matrix Combination 2 model 2002-2003 (diagonal elements bold)

2002	2003									
	CDA	PVDA	VVD	SP	GL	D66	CU	LPF	Anders	NG
CDA	.8858	.0157	.0179	.0349	.0153	.0036	.0145	.0107	.0015	
PVDA		.9462		.0173	.0192					.0173
VVD			.8529		.0672	.08				
SP		.4086		.5914						
GL		.6553			.3447					
D66		.013	.4296		.0748	.4825				
CU	.3460						.5755		.0785	
LPF	.2273		.1540	.0329				.3037		.2821
Anders	.2370	.0146	.0117				.0534		.6832	
NG		.2518		.0291						.719

Source: (Nationaal Kiezersonderzoek, 2002/2003 and authors own calculations)

Quite obvious in these transition matrices is the apparent loss of the LPF. It is also visible that the CDA has remained stable and the PvdA has recaptured a lot of support. Because of the short period of time between the elections it seems that there are even more stayers then normally and this is visible on the high values on the

diagonal in the NKO-matrix but also in the model based matrices. It is clear that in all the matrices the purple parties who have lost so heavily in the previous elections have regained part of their support that they had lost to the LPF. Again both combination matrices are the same, the reason for this effect is given in the previous paragraph.

2003-2006

For 2003-2006 are the transition matrices of the NKO and LCA already represented in chapter 4.

Table 24. Transition matrix LR model 2003-2006 (diagonal elements bold)

2003	2006									
	CDA	PVDA	VVD	SP	GL	D66	CU	PVV	Anders	NG
CDA	.8430	-.0318	-.0022	.0765	-.0162	-.0075	.0172	.0564	-.0172	.0027
PVDA	-.0748	.7416	-.0224	.2252	-.0044	-.0184	.0109	-.0164	.0052	.0128
VVD	.2548	.0850	.8329	.2721	-.0945	-.0447	.0368	.1887	-.0164	.1924
SP	.1417	-.2749	-.1320	1.6291	-.1073	-.0930	-.0578	-.0123	-.0922	.0544
GL	.3436	.2122	.2608	.0798	.9923	.2024	.0934	.1932	.0411	.3042
D66	-.2413	-.1210	.0920	-1.361	.5959	.4550	-.0819	-.7486	.2387	-.4962
CU	.0849	.1137	-.0278	-.1720	-.0847	.0257	1.3768	-.1866	.0361	.2518
LPF	.2513	-.1456	-.0593	-.0533	-.0825	.0699	.0198	.2922	.2566	.1648
Anders	-.0549	-.0465	.0181	-.1415	.0910	-.0069	.0861	-.0259	1.0380	-.1088
NG	-.1050	.1060	.0050	-.0085	.0146	.0318	-.0158	.0946	.0098	.8106

Source: (Authors own calculations)

Table 25. Transition matrix QP model 2003-2006 (diagonal elements bold)

2003	2006									
	CDA	PVDA	VVD	SP	GL	D66	CU	PVV	Anders	NG
CDA	.8396		.0014	.0318			.0316	.0392	.0016	.0547
PVDA		.7627		.1822			.0122			.0429
VVD	.1496		.8218						.0286	
SP				1.0000						
GL		.0439		.1366	.5604					.2591
D66					.3551	.4029	.0764		.1656	
CU							1.0000			
LPF								.5608	.1138	.3254
Anders							.0993		.9007	
NG				.1811				.0761		.7428

Source: (Authors own calculations)

Table 26. Transition matrix IPF model 2003-2006 (diagonal elements bold)

2003	2006									
	CDA	PVDA	VVD	SP	GL	D66	CU	PVV	Anders	NG
CDA	.2668	.1511	.1227	.1237	.0281	.0113	.0368	.0480	.0407	.1708
PVDA	.2477	.1635	.1229	.1285	.0300	.0120	.0352	.0472	.0384	.1745
VVD	.2549	.1545	.1316	.1238	.0302	.0125	.0340	.0477	.0398	.1709
SP	.2515	.1590	.1228	.1308	.0300	.0121	.0327	.0480	.0368	.1762
GL	.2490	.1599	.1268	.1275	.0319	.0129	.0340	.0471	.0375	.1733
D66	.2509	.1572	.1314	.1252	.0317	.0131	.0336	.0472	.0385	.1714
CU	.2626	.1534	.1155	.1139	.0273	.0108	.0555	.0430	.0541	.1637
LPF	.2534	.1545	.1267	.1250	.0290	.0119	.0346	.0494	.0411	.1743
Anders	.2614	.1454	.1185	.1120	.0256	.0105	.0502	.0465	.0657	.1642
NG	.2507	.1582	.1221	.1285	.0292	.0118	.0338	.0492	.0385	.1778

Source: (Authors own calculations)

Table 27. Transition matrix Combination 1 model 2003-2006 (diagonal elements bold)

2003	2006									
	CDA	PVDA	VVD	SP	GL	D66	CU	PVV	Anders	NG
CDA	.8267		.0176	.0346			.0223	.0421	.0072	.0495
PVDA		.7072		.2112	.0019		.0182			.0615
VVD	.1678		.6680		.0175	.0427	.0141		.0899	
SP				.8154						.1846
GL		.3809		.0058	.5765	.0369				
D66			.628		.2327	.1393				
CU							.8374		.1626	
LPF								.4749	.0387	.4864
Anders	.0338		.0116				.2668		.6877	
NG				.2161				.093		.6909

Source: (Aarts *et al.* 2007: pp. 224 and authors own calculations)

Table 28. Transition matrix Combination 2 model 2003-2006 (diagonal elements bold)

2003	2006									
	CDA	PVDA	VVD	SP	GL	D66	CU	PVV	Anders	NG
CDA	.8380		.0008	.0295			.0328	.0404	.0006	.058
PVDA		.7529		.1965			.0134			.0372
VVD	.1505		.7825						.0669	
SP		.0245		.9755						
GL				.0178	.5978	.0164	.0054			.3626
D66		.0811	.0811	.0579	.312	.3841	.0588		.025	
CU							1.0000			
LPF			.0867	.0514				.5333	.0903	.2383
Anders	.0115			.0115			.0896		.8873	
NG				.1729				.0811		.7460

Source: (Aarts *et al.* 2007: pp. 224 and authors own calculations)

The Linear Regression transition matrix shows a lot of results below 0 and above 1. The value of this matrix is therefore not very high, because of the illogical transitions.

Inspection of the LPM transition matrix shows that most voters are stayers. Particularly the SP, CDA, PvdA, CU, VDD and GL have high stayer rates. Someone who did not vote in 2003 often was also a non-voter in 2006 (74%). But you can still see that 18% of the non-voters in 2003 did vote for the SP in 2006, explaining partly the growth of the SP. The 2003 electorate of D'66 was fragmented over at least four parties. According to the LPM model, in 2006 over 56% of the previous LPF voters casted their vote in favor of the PVV and a large portion of former LPF voters remained at home. The SP and CU have values of 1 on their diagonal indicating that they have very loyal support.

Because of the heavy restrictions on the IPF model a very different transition matrix is presented. The marginal results need to be equal in this approach, which is a very severe restriction. This model uses the initial matrix as start matrix. It is interesting to observe the low values on the diagonal. One can see that these low values on the diagonal are not realistic, leading to a mediocre fit.

As shown with the first combination matrix, this combination matrix contains a lot of zero's. The values of 1 have disappeared because of the confidence interval restriction. Because of the other restriction of the rows summing up to 1, some other zeros in the matrix have disappeared. The second combination matrix does contain less zero's than the first combination matrix but still does contain some zero values that does not seem likely. It is quite logical that the values within this matrix are still calculated as close as possible near the optimal solution, again producing zero's. Still, no transitions between the PvdA and CDA are detected, which doesn't seem likely at all.

7.3 Comparison all models per election year

1998-2002



Figure 11a. 100% stayer model 98-02

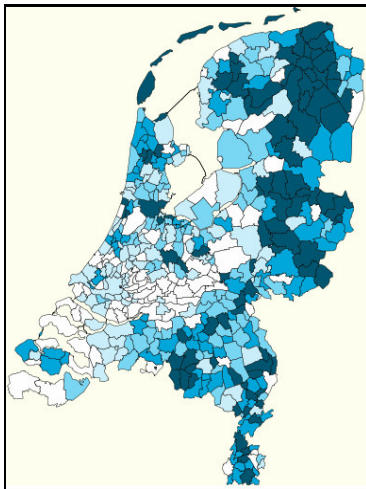


Figure 11b. NKO model 98-02

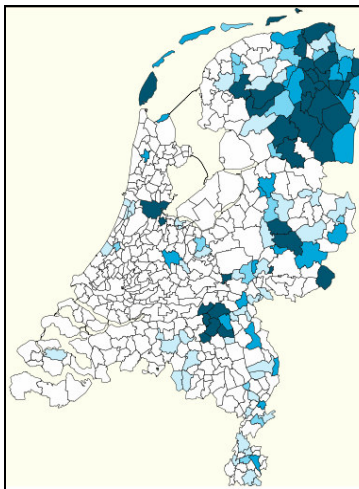


Figure 11c. LPM model 98-02

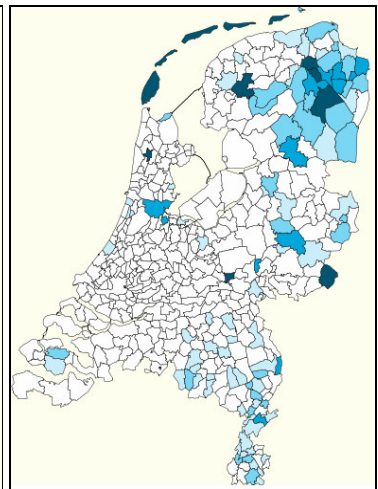


Figure 11d. LCA model 98-02

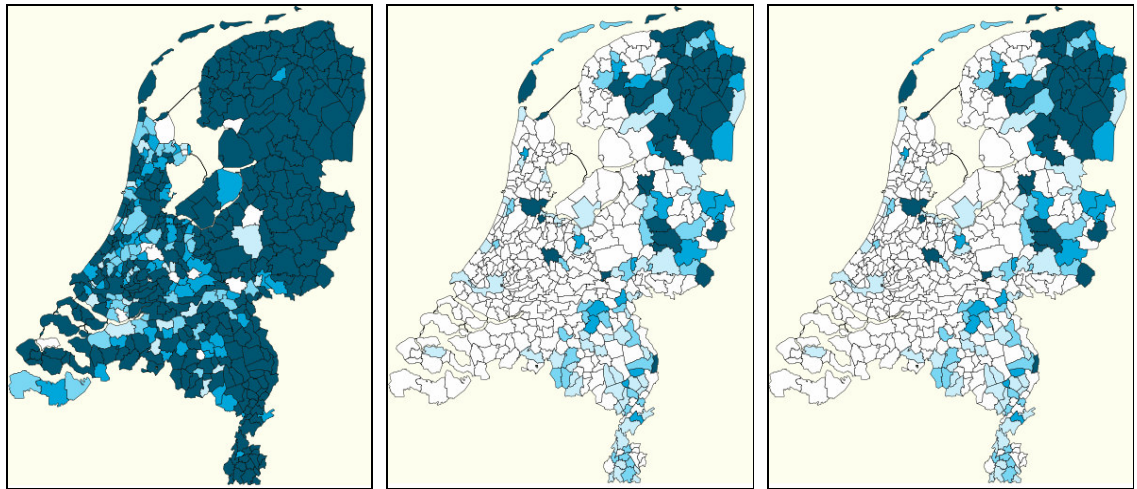


Figure 11e. IPF model 98-02

Figure 11f. Combo model 98-02

Figure 11g. Combo2 model 98-02

Legend 1998-2002: $R^2 < 0.9598$ $R^2 < 0.9702$ $R^2 < 0.9778$ $R^2 < 0.9839$ $R^2 < 1$

These pictures, based on the percentile classification of the NKO 2002, show that the fit of the models is very high. Most municipalities in the LCA-model have a McF. pseudo R^2 between the 0.9839 and 1. The fit of the 100% stayer model is worse than the independence model. All models again seem to have problems with the islands. Also the regions/ provinces Groningen and Overijssel show a bad fit. Since the general fit of the models is so high it is difficult to find an explanation that is statistically significant. The only visible explanation in the results is that the VVD has a slight underestimation in these parts of the Netherlands. Looking at the preferential votes for some important candidates create peaks in the number of votes for a certain party in several municipalities. This leads in most of those municipalities to a slightly lower fit. We can find misfits for the SP in Oss, misfits for the LPF in Rotterdam and misfits for the D'66 in Leiden. This analogy is also applicable in other municipalities.

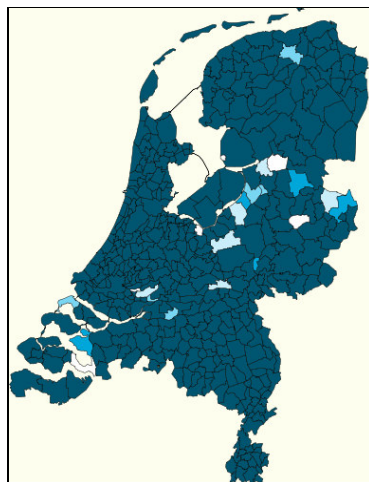


Figure 12a. 100% stayer model 02-03

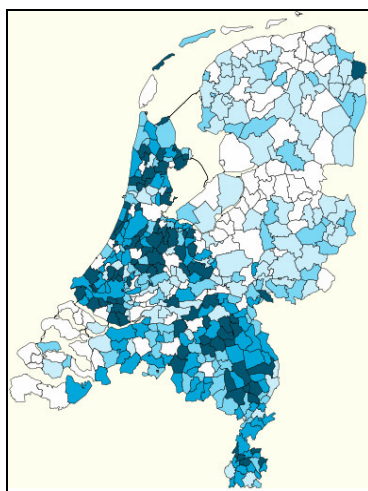


Figure 12b. NKO model 02-03

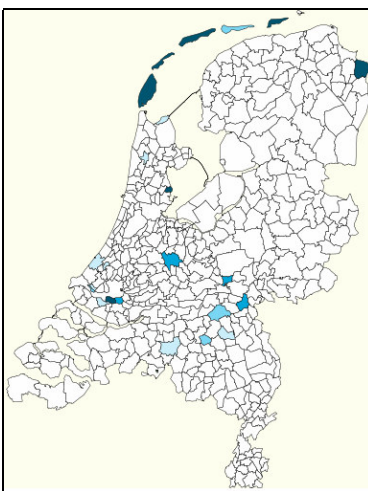


Figure 12c. LPM model 02-03

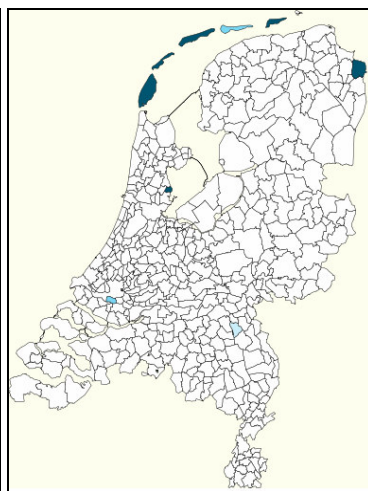


Figure 12d. LCA model 02-03

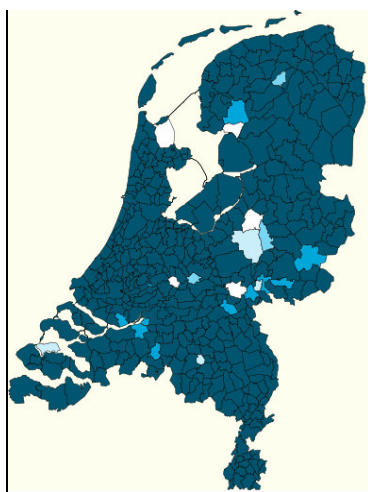


Figure 12e. IPF model 02-03

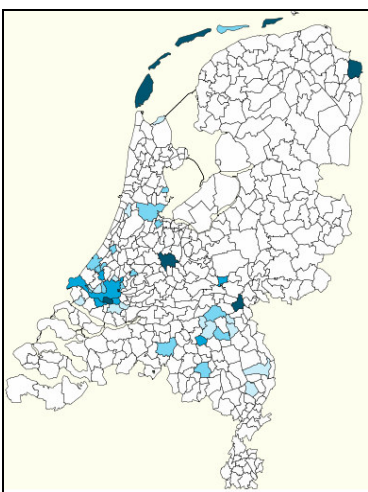


Figure 12f. Combo model 02-03

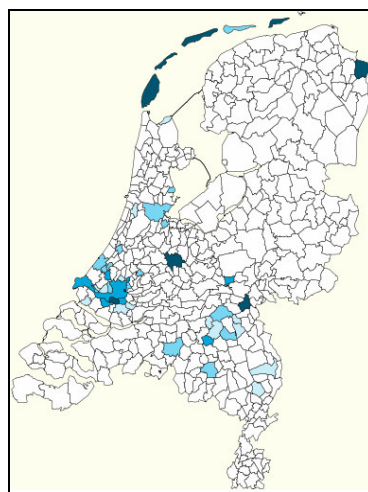







Figure 12g. Combo2 model 02-03

Legend 2002-2003:  $R^2 < 0.8376$  $R^2 < 0.8632$  $R^2 < 0.8873$  $R^2 < 0.9116$  $R^2 < 1$

These pictures show little distinctiveness between the model fits, the estimates of all models are close to each other. The elections were only 17 months apart from each other. Most voters have remained with their parties, leading to a large proportion of stayers (see the transition matrices in section 7.2 of appendix A). Again, the estimates for the islands are not well. Interestingly enough is the municipality of Reiderland in the North of Groningen somewhat worse. Because the New Communist Party (NCPN) didn't participate in 2002 a relatively large proportion of voters voted for the LPF. In 2003 they however switched for a small part back to the NCPN and to other parties leading to a somewhat inferior fit.

2003-2006

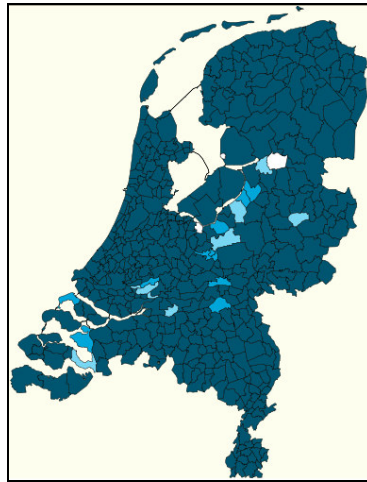


Figure 13a. 100% stayer model

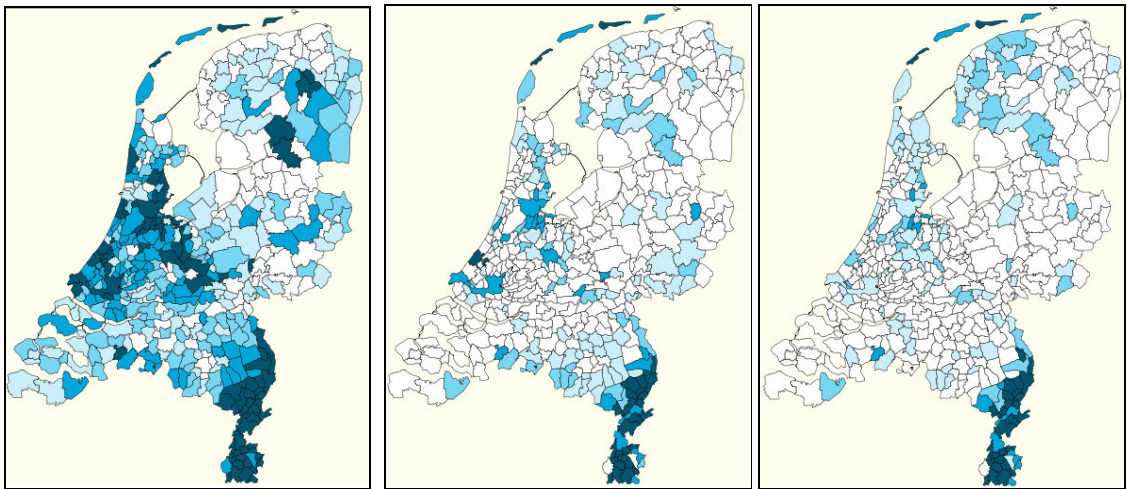


Figure 13b. NK0 model 03-06

Figure 13c. LPM model 03-06

Figure 13d. LCA model 03-06

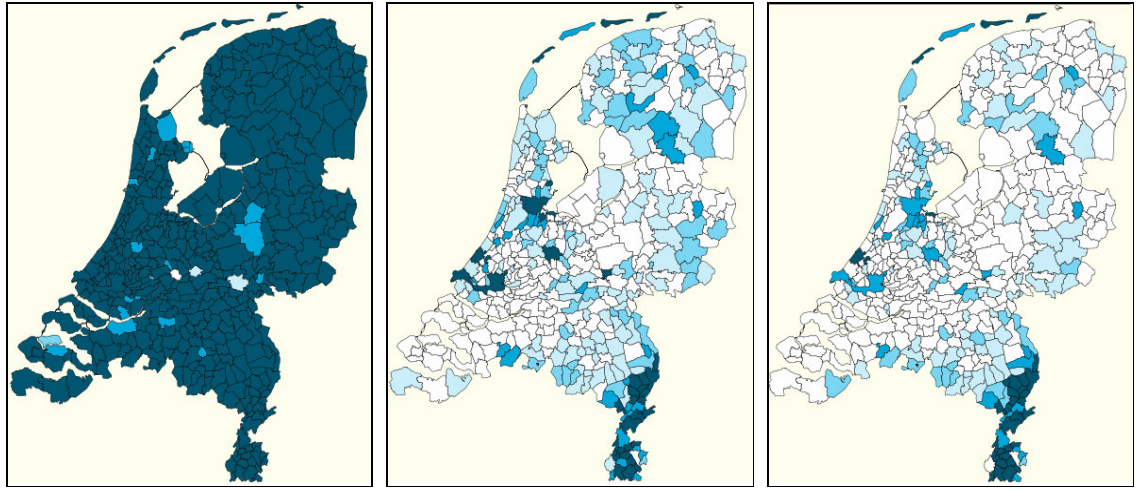


Figure 13e. IPF model 03-06

Figure 13f. Combo model 03-06

Figure 13g. Combo2 model 03-06

Legend 2003-2006: $R^2 < 0.8668$ $R^2 < 0.9245$ $R^2 < 0.9464$ $R^2 < 0.9622$ $R^2 < 1$

Here we can see that including all models doesn't change much in the analysis of 2006, already presented in chapter 4.

7.4 Extensions to the mover-stayer model

There are several extensions of the mover-stayer model. Hawkes multiple transition matrix uses linear regression. This method makes it possible to break electoral districts up into smaller homogenous units. This simple model outperforms the original mover-stayer model (Hawkes, 1969).

Another extension to the mover-stayer model is the memory model by Upton, 1977. Instead of having equal probabilities for every election he introduces constants $a_{k-1}, a_{k-2}, \dots, a_{k-l}$ that are associated with the individual party choice a in the previous l elections. With $x_{ij} = 1$ or 0 the probability of voting by an individual for party j in the k th election is given by $(\sum_{i=1}^l a_i x_{k-i}) p_j$. The attraction of a individual at election k with no previous voting history for party j is presented as p_j . He then incorporates an individuals l own previous voting history into the model. His research does not present definitive results but does present motivating incentives for more studies (Upton, 1977, pp.87).

The generalized mover-stayer model of Cook *et al.* (2002) describes an extension to the latent mover-stayer model. They define a mixture of nested continuous-time Markov processes. Instead of having the stayers stay in their initial steady state, each individual may have one or more absorbing states but when this state is entered no more transitions out of this state take place. However, before this final state is

entered, stayers may make transitions to a number of other states (Cook *et al.*, 2002, pp. 2). Especially the type heterogeneity found in election data can also be incorporated better. Because there are voters that remain in the same state for a very long time, the original mover-stayer is suited. For instance, voters are relative stable in their voting behavior (Egmond *et al.*, 1998). However, they may go through some party choices before settling in their final choice. This can be modeled using the different absorbing states and can therefore be a more suitable model than the original mover-stayer model (Cook *et al.*, 2002, pp. 4).

7.5 Used programs and code

The programs used in this project are various. For the compilation of the datasets and NKO analysis the statistical program SPSS has been employed.

For the calculation of the Latent Class Analysis the program Panmark written by F. van de Pol was used.

For the comparative analysis of the estimation methods the programs R and Matlab were used. The code for the free statistical program R is provided below.

Code R:

```
getwd()
kamer0306n <- as.matrix(read.table("kamer0306n.dat"))
#kamerverkiezingen 2003 en 2006
kamer0203n <- as.matrix(read.table("kamer0203n.dat"))
#kamerverkiezingen 2002 en 2003
kamer9802n <- as.matrix(read.table("kamer9802n.dat"))
#kamerverkiezingen 1998 en 2002
NKO2006 <- as.matrix(read.table("NKO2006.dat")) #NKO transitie matrix
NKO2003 <- as.matrix(read.table("NKO2003.dat")) #NKO transitie matrix
NKO2002 <- as.matrix(read.table("NKO2002.dat")) #NKO transitie matrix
frank2006 <- as.matrix(read.table("frank2006.dat")) #franks matrix
frank2003 <- as.matrix(read.table("frank2003.dat")) #franks matrix
frank2002 <- as.matrix(read.table("frank2002.dat")) #franks matrix
#combo2 <- as.matrix(read.table("combo.dat")) #combo matrix
stayer <- as.matrix(read.table("stayer.dat")) #stayer matrix
pseudo4 <- as.matrix(read.table("PSEUDO4.dat")) #Pseudo-R LPM en
gemcodes

eps <- .Machine$double.eps

code <- pseudo4[,1] #Gemeentecodes
pseudo4 <- pseudo4[,2] #Pseudo-R

gem <- kamer0306n[,1] #Gemeentes genummerd 1-459
gemcode <- kamer0306n[,2] #Gemeentecodes
t <- kamer0306n[,3] #verkiezingen genummerd 1,2
p <- kamer0306n[,4] #Partijen genummerd 1-10
f <- kamer0306n[,5] #Frequentie

#gem <- kamer0203n[,1] #Gemeentes genummerd 1-459
#gemcode <- kamer0203n[,2] #Gemeentecodes
#t <- kamer0203n[,3] #verkiezingen genummerd 1,2
```

```

#p <- kamer0203n[,4] #Partijen genummerd 1-10
#f <- kamer0203n[,5] #Frequentie

#gem <- kamer9802n[,1] #Gemeentes genummerd 1-459
#gemcode <- kamer9802n[,2] #Gemeentecodes
#t <- kamer9802n[,3] #verkiezingen genummerd 1,2
#p <- kamer9802n[,4] #Partijen genummerd 1-10
#f <- kamer9802n[,5] #Frequentie

p.kij <- array(0, c(max(gem),max(p),max(p)))
P.kij <- array(0, c(max(gem),max(p),max(p)))
f.ki <- matrix(0, max(gem),max(p))
f.kj <- matrix(0, max(gem),max(p))
f.ij <- matrix(0, max(p),max(p))
f.k <- array(0, max(gem))
f1.k <- matrix(0, max(gem),1)
f2.k <- matrix(0, max(gem),1)
p2.k <- matrix(0, max(gem),1)
f.i <- array(0, max(p))
f.j <- array(0, max(p))

p.ki <- matrix(0, max(gem),max(p))
p.kj <- matrix(0, max(gem),max(p))
p.ij <- matrix(0, max(p),max(p))
Inip.ij <- matrix(0, max(p),max(p))
P.ki <- matrix(0, max(gem),max(p))
P.kj <- matrix(0, max(gem),max(p))
P.ij <- matrix(0, max(p),max(p))
p.k <- matrix(0, max(gem),1)
p.i <- matrix(0, max(p),1)
p.j <- matrix(0, max(p),1)

for (x in 1:length(gem)) {
  k <- gem[x]
  i <- p[x]
  s <- t[x]
  if (s < 2) f.ki[k, i] <- f[x]
  if (s > 1) f.kj[k, i] <- f[x]} #Marginalen gemeenten x
verkiezing
for (i in 1:max(p)) {
  for (k in 1:max(gem)) {
    f.i[i] <- f.i[i] + f.ki[k, i]
    f1.k[k] <- f1.k[k] + f.ki[k, i]}} #Marginalen
voor eerste verkiezingen
p.i <- f.i/sum(f.i) # Marginale kansen voor eerste verkiezingen

for (j in 1:max(p)) {
  for (k in 1:max(gem)) {
    f.j[j] <- f.j[j] + f.kj[k, j]
    f2.k[k] <- f2.k[k] + f.kj[k, j]}} #Marginalen
voor tweede verkiezingen
p.j <- f.j/sum(f.j) # Marginale kansen voor tweede verkiezingen
p2.k <- f2.k/sum(f2.k) # Marginale kansen voor gemeentes bij tweede
verkiezingen
for (k in 1:max(gem)) {
  for (i in 1:max(p)) {
    p.ki[k,i] <- f.ki[k,i]/f1.k[k]}}
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    p.kj[k,j] <- f.kj[k,j]/f2.k[k]}}

for (i in 1:max(p)) {
  for (j in 1:max(p)) {

```

```

Inip.ij[i,j] <- p.i[j]}} #De initiële TRANSITIEMATRIX!!

for (k in 1:max(gem)) {
  for (i in 1:max(p)) {
    for (j in 1:max(p)) {
      P.kij[k,i,j] <- p2.k[k]*p.i[i]*p.j[j]}}

for (i in 1:max(p)) {
  for (j in 1:max(p)) {
    for (k in 1:max(gem)) {
      P.ij[i,j] <- P.ij[i,j]+P.kij[k,i,j]}}}

#####
#Test met onafhankelijkheid
#####
phat1.kj <- p.ki %% Inip.ij
fhat1.kj <- matrix(0, max(gem),max(p))

pSums.k <- rowSums(phat1.kj)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    fhat1.kj[k,j] <- phat1.kj[k,j]*f2.k[k]/pSums.k[k]}}
LRX2.1 <- 2*sum(f.kj * log(f.kj/fhat1.kj))
Inip.ij

#####
#Test alleen maar stayers
#####
phat2.kj <- p.ki %% stayer
fhat2.kj <- matrix(0, max(gem),max(p))

pSums.k <- rowSums(phat2.kj)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    fhat2.kj[k,j] <- phat2.kj[k,j]*f2.k[k]/pSums.k[k]}}
LRX2.2 <- 2*sum(f.kj * log(f.kj/fhat2.kj))
stayer

#####
#Test met NKO
#####
phat3.kj <- p.ki %% NKO2006
fhat3.kj <- matrix(0, max(gem),max(p))

pSums.k <- rowSums(phat3.kj)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    fhat3.kj[k,j] <- phat3.kj[k,j]*f2.k[k]/pSums.k[k]}}
LRX2.3 <- 2*sum(f.kj * log(f.kj/fhat3.kj))
NKO2006

#####
#Regressie op frequenties (Keller & Ten Cate, 1977) Lineaire regressie
#####
B <- solve(t(f.ki) %% f.ki) %% t(f.ki) %% f.kj
phat4.kj <- p.ki %% B
fhat4.kj <- matrix(0, max(gem),max(p))

pSums.k <- rowSums(phat4.kj)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    fhat4.kj[k,j] <- phat4.kj[k,j]*f2.k[k]/pSums.k[k]}}
LRX2.4 <- 2*sum(f.kj * log(f.kj/fhat4.kj))

```

B

```
#####
#Regressie op frequenties (Keller & Ten Cate, 1977) Kwadratisch
Programmeren
#####
#http://cran.r-project.org/web/packages/quadprog/index.html
#hier functie alle waarden corrigeren zodat de rijen optellen tot 1
#Kwadratisch programmeren oplossing met voorwaarden kansverdeling
library(quadprog)

#Rijen en kolomrestricties tegelijk
#####
meq <- max(p)
XtX <- t(p.ki) %*% p.ki
Dmat <- matrix(0, max(p)^2, max(p)^2)
dvec <- matrix(0, max(p)^2, 1)
Amat <- matrix(0, max(p)^2, max(p))
for(i in 1:max(p))
{
  Xty <- t(p.ki)%*%p.kj[,i]
  range<-(i-1)*max(p)+1:(i*max(p))
  Dmat[range,range]<-XtX %*% XtX
  dvec[range] <- XtX %*% Xty #te minimaliseren vector
}
Amat <- cbind(
  t(matrix(diag(max(p)), max(p), max(p)^2)),

  diag(max(p)^2)
)
bvec <- rbind(
  matrix(1, max(p), 1),
  matrix(0, max(p)^2, 1)
)
sol<-solve.QP(Dmat, dvec, Amat, bvec, meq)
LPM<-matrix(sol$solution, max(p), max(p))

phat5.kj <- p.ki %*% LPM
fhat5.kj <- matrix(0, max(gem),max(p))

pSums.k <- rowSums(phat5.kj)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    fhat5.kj[k,j] <- phat5.kj[k,j]*f2.k[k]/pSums.k[k]}
LRX2.5 <- 2*sum(f.kj * log(f.kj/fhat5.kj))
print(formatC(abs(LPM), dig=4, format="f"), quote=FALSE)

#####
#Test met LCA
#####
phat6.kj <- p.ki %*% frank2006
fhat6.kj <- matrix(0, max(gem),max(p))

pSums.k <- rowSums(phat6.kj)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    fhat6.kj[k,j] <- phat6.kj[k,j]*f2.k[k]/pSums.k[k]}
LRX2.6 <- 2*sum(f.kj * log(f.kj/fhat6.kj))
frank2006

#####
#IPF
```

```
#####
iter <- 0
P.kij.size.k <- dim(P.kij)[1]
P.kij.size.i <- dim(P.kij)[2]
P.kij.size.j <- dim(P.kij)[3]

while (abs(p.ij[1,1] - P.ij[1,1])>10E-6){
  iter <- iter +1
  p.ij <- P.ij

  #Quasi-marginalen bijstellen
  P.ki <- apply(P.kij, c(1,2), sum)

  #Sjtapf einz
  P.kij <- P.kij * ((p.ki/P.ki) %o% array(1, P.kij.size.j ))

  #Quasi-marginalen bijstellen
  P.kj <- apply(P.kij, c(1,3), sum)

  #Sjtapf zwei
  P.kij <- P.kij * aperm((p.kj/P.kj) %o% array(1, P.kij.size.i ),
c(1,3,2))

  #Quasi-marginalen bijstellen
  P.ij <- apply(P.kij, c(2,3), sum)

  #Sjtapf drei
  P.kij <- P.kij * (array(1, P.kij.size.k ) %o% (p.ij/P.ij))

}#End while

P.i <- matrix(0, max(p),1)
for (i in 1:max(p)) {
  for (j in 1:max(p)) {
    P.i[i] <- P.i[i]+P.ij[i,j]}
for (i in 1:max(p)) {
  for (j in 1:max(p)) {
    P.ij[i,j] <- P.ij[i,j]/P.i[i]}

phat7.kj <- p.ki %*% P.ij
fhat7.kj <- matrix(0, max(gem),max(p))

pSums.k <- rowSums(phat7.kj)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    fhat7.kj[k,j] <- phat7.kj[k,j]*f2.k[k]/pSums.k[k]}

LRX2.7 <- 2*sum(f.kj * log(f.kj/fhat7.kj))
P.ij

#####
#COMBO K&T en NKO
#####
#http://cran.r-project.org/web/packages/quadprog/index.html
#hier functie alle waarden corrigeren zodat de rijen optellen tot 1
#Quadratisch programmeren oplossing met voorwaarden kansverdeling
library(quadprog)

#berekening BTI
#####
SD <- sd(NKO2006, na.rm=FALSE)/sqrt(459) #Sx
#Sx s/sqrt(n) s=sqrt 1/N-1 sum_i=1^N (x_i-xstreep)^2
LCFI <- mean(phat3.kj)-2*SD #sample mean -/+2*Sx
```

```

HCFI <- mean(phat3.kj)+2*SD
#marge <- 0.025 # 10%
diagmin <- matrix(diag(NKO2006) - LCFI, max(p), 1)
diagmax <- matrix(diag(NKO2006) + HCFI, max(p), 1)

#nieuwe berekening
meq <- max(p)
XtX <- t(p.ki) %*% p.ki
DiagHelper <- matrix(0, max(p)^2, max(p))
Dmat <- matrix(0, max(p)^2, max(p)^2)
dvec <- matrix(0, max(p)^2, 1)
for(i in 1:max(p))
{
  DiagHelper[(i - 1) * max(p) + i, i] <- 1
  Xty <- t(p.ki)%*%p.kj[,i]
  range<-(i-1)*max(p)+1:(i*max(p))
  Dmat[range,range]<-XtX %*% XtX
  dvec[range] <- XtX %*% Xty #te minimaliseren vector
}
Amat <- cbind(
  t(matrix(diag(max(p)), max(p), max(p)^2)),
  DiagHelper,
  -DiagHelper,
  diag(max(p)^2)
)
bvec <- rbind(
  matrix(1, max(p), 1),
  diagmin,
  -diagmax,
  matrix(0, max(p)^2, 1)
)
sol<-solve.QP(Dmat, dvec, Amat, bvec, meq)
combo <- matrix(sol$solution, max(p), max(p))

phat8.kj <- p.ki %*% combo
fhat8.kj <- matrix(0, max(gem),max(p))

pSums.k <- rowSums(phat8.kj)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    fhat8.kj[k,j] <- phat8.kj[k,j]*f2.k[k]/pSums.k[k]}
LRX2.8 <- 2*sum(f.kj * log(f.kj/fhat8.kj))
print(formatC(abs(combo), dig=4, format="f"), quote=FALSE)

#####
#COMBO 2 0 waarden vervangen met NK0-waarden
#####
SD <- sd(NKO2006, na.rm=FALSE)/sqrt(459) #Sx
#Sx s/sqrt(n) s=sqrt 1/N-1 sum_i=1^N (x_i-xstreep)^2
LCFI <- mean(phat3.kj)-2*SD #sample mean -/+2*Sx
HCFI <- mean(phat3.kj)+2*SD
marge <- 0.025 # 10%
grenswaarde<-1e-4
BTIMIN<-(LPM<grenswaarde)*(NK02006-LCFI)-(LPM>=grenswaarde)*(1E99)
BTIMIN<-matrix(BTIMIN, max(p)^2, 1)
BTIMAX<-(LPM<grenswaarde)*(NK02006+HCFI)+(LPM>=grenswaarde)*(1E99)
BTIMAX<-matrix(BTIMAX, max(p)^2, 1)

#nieuwe berekening
meq <- max(p)
XtX <- t(p.ki) %*% p.ki
Dmat <- matrix(0, max(p)^2, max(p)^2)
dvec <- matrix(0, max(p)^2, 1)

```

```

for(i in 1:max(p))
{
  Xty <- t(p.ki)%*%p.kj[,i]
  range<-(i-1)*max(p)+1:(i*max(p))
  Dmat[range,range]<-XtX %*% XtX
  dvec[range] <- XtX %*% Xty #te minimaliseren vector
}
Amat <- cbind(
  t(matrix(diag(max(p)), max(p), max(p)^2)),
  diag(max(p)^2),
  -diag(max(p)^2),
  diag(max(p)^2)
)
bvec <- rbind(
  matrix(1, max(p), 1),
  BTIMIN,
  -BTIMAX,
  matrix(0, max(p)^2, 1)
)
sol<-solve.QP(Dmat, dvec, Amat, bvec, meq)
combo2 <- matrix(sol$solution, max(p), max(p))

phat9.kj <- p.ki %*% combo2
fhat9.kj <- matrix(0, max(gem),max(p))

pSums.k <- rowSums(phat9.kj)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    fhat9.kj[k,j] <- phat9.kj[k,j]*f2.k[k]/pSums.k[k]}
LRX2.9 <- 2*sum(f.kj * log(f.kj/fhat9.kj))
print(formatC(abs(combo2), dig=4, format="f"), quote=FALSE)

#####
#CHI-SQUARE TEST PER MUNICIPALITY
#####
res1.kj <-matrix(0,max(gem),max(p))
res2.kj <-matrix(0,max(gem),max(p))
res3.kj <-matrix(0,max(gem),max(p))
res4.kj <-matrix(0,max(gem),max(p))
res5.kj <-matrix(0,max(gem),max(p))
res6.kj <-matrix(0,max(gem),max(p))
res7.kj <-matrix(0,max(gem),max(p))
res8.kj <-matrix(0,max(gem),max(p))
res9.kj <-matrix(0,max(gem),max(p))
enen <- matrix(1,10,1)
for (k in 1:max(gem)) {
  for (j in 1:max(p)) {
    res1.kj[k,j] <- ((fhat1.kj[k,j]-f.kj[k,j])^2)/fhat1.kj[k,j]
    res2.kj[k,j] <- ((fhat2.kj[k,j]-f.kj[k,j])^2)/fhat2.kj[k,j]
    res3.kj[k,j] <- ((fhat3.kj[k,j]-f.kj[k,j])^2)/fhat3.kj[k,j]
    res4.kj[k,j] <- ((fhat4.kj[k,j]-f.kj[k,j])^2)/fhat4.kj[k,j]
    res5.kj[k,j] <- ((fhat5.kj[k,j]-f.kj[k,j])^2)/fhat5.kj[k,j]
    res6.kj[k,j] <- ((fhat6.kj[k,j]-f.kj[k,j])^2)/fhat6.kj[k,j]
    res7.kj[k,j] <- ((fhat7.kj[k,j]-f.kj[k,j])^2)/fhat7.kj[k,j]
    res8.kj[k,j] <- ((fhat8.kj[k,j]-f.kj[k,j])^2)/fhat8.kj[k,j]
    res9.kj[k,j] <- ((fhat9.kj[k,j]-f.kj[k,j])^2)/fhat9.kj[k,j]
  }
}
res1.k <- res1.kj%%enen
res2.k <- res2.kj%%enen
res3.k <- res3.kj%%enen
res4.k <- res4.kj%%enen

```

```

res5.k <- res5.kj*%enen
res6.k <- res6.kj*%enen
res7.k <- res7.kj*%enen
res8.k <- res8.kj*%enen
res9.k <- res9.kj*%enen

#####
#Pseudo R per gemeente
#####
PR2.2.k <- 1-res2.k/res1.k
PR2.3.k <- 1-res3.k/res1.k
PR2.4.k <- 1-res4.k/res1.k
PR2.5.k <- 1-res5.k/res1.k
PR2.6.k <- 1-res6.k/res1.k
PR2.7.k <- 1-res7.k/res1.k
PR2.8.k <- 1-res8.k/res1.k
PR2.9.k <- 1-res9.k/res1.k

#####
#Schatting zetelverdeling per model
#####
sumtot <- colSums(f.kj)
sumkiesdeler <- sum(sumtot)1:9)
kiesdeler <- 9838683/150
sumfecht <- colSums(f.kj)
sumfONAF <- colSums(fhat1.kj)
sumfSTAYER <- colSums(fhat2.kj)
sumfNKO <- colSums(fhat3.kj)
sumfLPM <- colSums(fhat5.kj)
sumfLCA <- colSums(fhat6.kj)
sumfIPF <- colSums(fhat7.kj)
sumfCOMBO <- colSums(fhat8.kj)
sumfCOMBO2 <- colSums(fhat9.kj)

sumfONAF
sumfSTAYER
sumfNKO
sumfLPM
sumfLCA
sumfIPF
sumfCOMBO
sumfCOMBO2
sumfecht
#2003-2006
herverdeling<-function(fractioneel)
{
  direct<-floor(fractioneel)
  rest<-sum(fractioneel)-sum(direct)
  #- voor aflopend
  rang<-rank(-c(fractioneel/(direct+1),
    fractioneel/(direct+2)))
  extra<-matrix(0, length(fractioneel), 2)
  extra[rang[1:rest]] <- extra[rang[1:rest]] + 1
  extra<-rowSums(extra)
  direct + extra
}

#####
#STATISTIEKEN
#####
GFX2.1<-sum(res1.k) #Onafhankelijkheid
GFX2.2<-sum(res2.k) #STAYER

```



```

GFX2.3<-sum(res3.k) #NKO
GFX2.4<-sum(res4.k) #LR
GFX2.5<-sum(res5.k) #LP
GFX2.6<-sum(res6.k) #LCA
GFX2.7<-sum(res7.k) #IPF
GFX2.8<-sum(res8.k) #COMBO
GFX2.9<-sum(res9.k) #COMBO2

PR2.2 <- 1-GFX2.2/GFX2.1
PR2.3 <- 1-GFX2.3/GFX2.1
PR2.4 <- 1-GFX2.4/GFX2.1
PR2.5 <- 1-GFX2.5/GFX2.1
PR2.6 <- 1-GFX2.6/GFX2.1
PR2.7 <- 1-GFX2.7/GFX2.1
PR2.8 <- 1-GFX2.8/GFX2.1
PR2.9 <- 1-GFX2.9/GFX2.1

GFX2.1 #Onafhankelijkheid
GFX2.2 #STAYER
GFX2.3 #NKO
GFX2.4 #LR
GFX2.5 #LP
GFX2.6 #LCA
GFX2.7 #IPF
GFX2.8 #COMBO
GFX2.9 #COMBO2

LRX2.1 #Onafhankelijkheid
LRX2.2 #STAYER
LRX2.3 #NKO
LRX2.4 #LR
LRX2.5 #LP
LRX2.6 #LC
LRX2.7 #IPF
LRX2.8 #COMBO
LRX2.9 #COMBO2

PR2.2 #STAYER
PR2.3 #NKO
PR2.4 #LR
PR2.5 #LP
PR2.6 #LC
PR2.7 #IPF
PR2.8 #COMBO
PR2.9 #COMBO2

which(PR2.2.k[,1]>3) #STAYER
which(PR2.3.k[,1]>3) #NKO
which(PR2.4.k[,1]>3) #LR
which(PR2.5.k[,1]>3) #LP
which(PR2.6.k[,1]>3) #LC
which(PR2.7.k[,1]>3) #IPF
which(PR2.8.k[,1]>3) #COMBO
which(PR2.9.k[,1]>3) #COMBO2

```

```
#####
#Misfit
#####
misfitLCA2006 <- (fhat6.kj-f.kj)/fhat6.kj
write.csv(misfitLCA2006, "misfitLCAenEcht2006.txt", row.names=F,
quote=FALSE)

#####
#PLOTS
#####
par(mfrow=c(3,3))
#chi square tegen gemeenten uitgezet
plot(PR2.2.k)
plot(PR2.3.k)
plot(PR2.4.k)
plot(PR2.5.k)
plot(PR2.6.k)
plot(PR2.7.k)
plot(PR2.8.k)
plot(PR2.9.k)

write.csv(res1.k, "ONAF2006.txt", row.names=F, quote=FALSE)
write.csv(res2.k, "STAYER2006.txt", row.names=F, quote=FALSE)
write.csv(res3.k, "NKO2006.txt", row.names=F, quote=FALSE)
write.csv(res4.k, "LR2006.txt", row.names=F, quote=FALSE)
write.csv(res5.k, "LPM2006.txt", row.names=F, quote=FALSE)
write.csv(res6.k, "LCA2006.txt", row.names=F, quote=FALSE)
write.csv(res7.k, "IPF2006.txt", row.names=F, quote=FALSE)
write.csv(res8.k, "COMBO2006.txt", row.names=F, quote=FALSE)
write.csv(res9.k, "COMBO22006.txt", row.names=F, quote=FALSE)

write.csv(PR2.2.k, "pr22006.txt", row.names=F, quote=FALSE)
write.csv(PR2.3.k, "pr32006.txt", row.names=F, quote=FALSE)
write.csv(PR2.4.k, "pr42006.txt", row.names=F, quote=FALSE)
write.csv(PR2.5.k, "pr52006.txt", row.names=F, quote=FALSE)
write.csv(PR2.6.k, "pr62006.txt", row.names=F, quote=FALSE)
write.csv(PR2.7.k, "pr72006.txt", row.names=F, quote=FALSE)
write.csv(PR2.8.k, "pr82006.txt", row.names=F, quote=FALSE)
write.csv(PR2.9.k, "pr92006.txt", row.names=F, quote=FALSE)

#####
#Plaatje Nederland
#####
codenum <- as.numeric(code)
ExportSaeToSvg = function(ids, data, levels, method = "range", bounds =
"", outputFile)
{
  # ids: id-codes van gemeenten
  # data: vector of data of same length as ids
  # levels: the number of levels to quantize into (each level will
correspond to a color)
  # For now this is set to 5 regardless.
  # outputFile: file name for exported svg file
  # (hard coded) templateFile: this is a svg file where the color
codes need replacing,
  # it should contain the small areas of
interest (i.e. 'gemeenten' for now)

  levels = 5

  # Define color codes (table with hex entries) and assign an index
to each small area
```

```

    #colorCodes = c("#F8FE00", "#FED400", "#FEA100", "#FE6200",
"#FF0000")
    colorCodes = c("#005670", "#00A7DA", "#79D7F4", "#CBEFFA",
"#FFFFFF")

    dataCodes = 1:length(data)

    if (method == "range") {
        # Calculate level boundaries
        minData <- min(data)
        maxData <- max(data)
        binSize <- (maxData - minData) / levels
        for (i in 1:levels) {
            minVal = minData + (i-1) * binSize
            maxVal = minVal + binSize
            dataCodes[data >= minVal & data <= maxVal] = i
        }
    }

    if (method == "quantile") {
        q = quantile(data ,c(.20, .40, .60, .80))
    }
    if (method == "manual") {
        q = bounds
    }
    dataCodes[data < q[1]] = 1
    dataCodes[data >= q[1] & data < q[2]] = 2
    dataCodes[data >= q[2] & data < q[3]] = 3
    dataCodes[data >= q[3] & data < q[4]] = 4
    dataCodes[data >= q[4]] = 5

    # Parse template file:
    # extract small area ID: if found, replace it's color code

    # create id strings consisting of 4 characters
    idstr <- as.character(ids)
    idstrlen <- nchar(idstr)
    nzeros <- 4 - idstrlen
    idstr[nzeros == 0] = paste("GM", idstr[nzeros == 0], sep="")
    idstr[nzeros == 1] = paste("GM0", idstr[nzeros == 1], sep="")
    idstr[nzeros == 2] = paste("GM00", idstr[nzeros == 2], sep="")
    idstr[nzeros == 3] = paste("GM000", idstr[nzeros == 3], sep="")

    # read template file
    templateFile = "C:\\Program Files\\R\\R-
2.6.0\\Gemeenten2006grijs.svg"
    conIn <- file(templateFile, "r")
    allLines <- readLines(conIn, -1)
    close(conIn)

    # Temporary code to make all polygons same color (gray)
    # for (i in 1:length(allLines)) {
    #     thisLine <- allLines[i]
    #     if (substring(thisLine, 1, 9) == "<g id=\"GM\"") {
    #         # substring(thisLine, 28, 34) <- "#cccccc"
    #         thisLine = paste(substring(thisLine, 1, 27),
"#cccccc", substring(thisLine, 33, 35), sep = "")
    #         allLines[i] <- thisLine
    #     }
    # }
    # outputFile = "C:\\Program Files\\R\\R-
2.6.0\\Gemeenten2006grijsv2.svg"
    # conOut <- file(outputFile, "w")

```

```

# writeLines(allLines, conOut)
# close(conOut)
# return()
# END - Temporary code to make all polygons same color

# now replace color codes in appropriate lines
range <- 1:length(idstr)
iCountGMlinesNotReplaced <- 0
iCountGMlinesReplaced <- 0
for (i in 1:length(allLines)) {
  thisLine <- allLines[i]
  testCode <- substring(thisLine, 8, 13)
  thisIndex <- range[idstr == testCode]
  if (length(thisIndex) > 0) {
    newCode <- colorCodes[dataCodes[thisIndex]]
    substring(thisLine, 28, 34) <- newCode
    allLines[i] <- thisLine
    iCountGMlinesReplaced <- iCountGMlinesReplaced + 1
  } else {
    if (substring(thisLine, 1, 9) == "<g id=\"GM\"") {
      iCountGMlinesNotReplaced <-
iCountGMlinesNotReplaced +1
    }
  }
}

conOut <- file(outputFile, "w")
writeLines(allLines, conOut)
close(conOut)

#if (method == "range") {
#   return(list(minData, maxData))
#}
invisible()
return(list(iCountGMlinesReplaced,iCountGMlinesNotReplaced ))
}

#ExportSaeToSvg = function(ids, data, levels, method = "range", bounds
= "", outputFile))
#q = c(0.8668, 0.9245, 0.9464, 0.9622)
q = c(0.8650, 0.9163, 0.9522, 0.9728)
outPath = "C:\\Program Files\\R\\R-2.6.0\\"
ExportSaeToSvg(codenum, PR2.2.k, 5, method = "manual", bounds = q,
outputFile = paste(outPath, "stayer2006b.svg", sep = ""))
ExportSaeToSvg(codenum, PR2.3.k, 5, method = "manual", bounds = q,
outputFile = paste(outPath, "NKO2006b.svg", sep = ""))
ExportSaeToSvg(codenum, PR2.4.k, 5, method = "manual", bounds = q,
outputFile = paste(outPath, "LR2006b.svg", sep = ""))
ExportSaeToSvg(codenum, PR2.5.k, 5, method = "manual", bounds = q,
outputFile = paste(outPath, "LPM2006b.svg", sep = ""))
ExportSaeToSvg(codenum, PR2.6.k, 5, method = "manual", bounds = q,
outputFile = paste(outPath, "LCA2006b.svg", sep = ""))
ExportSaeToSvg(codenum, PR2.7.k, 5, method = "manual", bounds = q,
outputFile = paste(outPath, "IPF2006b.svg", sep = ""))
ExportSaeToSvg(codenum, PR2.8.k, 5, method = "manual", bounds = q,
outputFile = paste(outPath, "COMBO2006b.svg", sep = ""))
ExportSaeToSvg(codenum, PR2.9.k, 5, method = "manual", bounds = q,
outputFile = paste(outPath, "COMBO22006b.svg", sep = ""))

```