Developments in Statistics and Methodology
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 9, Ljubljana: FDV 1993

Ecological Inference and Slovene Voting Data

Vasja Vehovar[1]

Abstract

In ecological inference aggregated data is used to infer about individual behavior. In this article some general problems are discussed in the case of two rounds of the 1990 presidential elections in Slovenia. The Slovene voting data has been aggregated at the level of precincts consisting of few hundred voters. Another attractive feature of the data is the fact that the two rounds of elections were only two weeks apart. The results of the regression approach and the logit approach are compared and evaluated with the estimates from the survey data.

1 Introduction

The Slovene voting data is composed of the administrative data from five recent elections performed in 1990 after the introduction of a multiparty political system: parliamentary elections, community elections, presidential elections (two rounds) and a plebiscite for independence. Compared with some other countries the number of elections is relatively small, however, even this data could help answer some interesting questions:

i) How do the voters from the first round of presidential elections behave in the second round?

ii) In which party were the voters more inclined not to vote for the independence of Slovenia?

iii) In community elections, to what extent did the voters remain loyal to the party chosen during parliamentary elections?

[1] Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 61109 Ljubljana, Slovenia

iv) How did voters of different parties behave during presidential elections?
v) What are the sociodemographic characteristics of the voters who have chosen a certain party or a certain presidential candidate?

As the behavior of an individual voter is not known it is impossible to give an exact answer to these questions. However, there are at least two approximate answers. Firstly, it can be inferred from a sample survey and, secondly, from the administrative electoral data aggregated at a particular level. Each of these approaches has specific advantages and disadvantages. Here, only latter approach is being discused in the following sections. Namely, in section (2) the data is briefly described. After that, the ecological inference (3) and its problems are exposed (4) with special attention to the regression (5) and the logit approach (6). Finally, the empirical results are presented (7) in a discussion (8) and conclusions are made (9).

2 Slovene voting data

There are about a million and a half eligible voters within 4000 precincts[1] in Slovenia. The precincts are relatively small and can be further aggregated into 62 communities[2]. All elections, except the plebiscite (December 1990), were held in April, 1990, on two occasions only:

- 8th April - presidential elections (1st round), general parliament elections;
- 22nd April - presidential election (2nd round), local (community) elections.

As the elections were performed in a short period of time, the problems of individuals who had moved or died were almost negligible. The same is true for new voters, that is, the voters who were eligible to vote for the first time during a particular election. Similarly, due to the short time gap between the elections, there is an advantage of relatively stable voting units.

Only the first (i) question from section 1 is being answered here, that is, how the voters from the first round of presidential elections behave in the second round. However, the methodology would remain the same when answering the other four questions.

Alternatively, the problem can be expressed with the question marks in the following summary table:

[1] The size of a precinct is relatively small - only few hundred voters. Thus, there is much more information in the aggregated data compared to the situation when only higher levels of aggregation are available.

[2] The data on 62 communities is complete; however, on the precinct level some communities are missing due to technical problems. Thus, the analysis was performed with three quoters, two thirds and one half of the precincts.

Table 1: TWO ROUNDS OF PRESIDENTIAL ELECTIONS IN SLOVENIA, APRIL 1990 (THOUSANDS OF VOTERS)

I. II.	KUČAN	PUČNIK	OTHER	TOTAL
KUČAN	?	?	?	583
PUČNIK	?	?	?	322
DEMŠAR	?	?	?	126
KRAMBERGER	?	?	?	224
OTHER	?	?	?	268
TOTAL	657	464	358	1.480

Of course, filling in the true values would be equal to knowing the voting behavior of the individual voter. As already mentioned, this information is not known, so the ecological inference from the aggregated data will be used, that is, the data in election outcomes within 4000 precincts and 62 communities.

3 Ecological inference

The individual data is often aggregated on a spatial[3] level: precinct, community, town, region, country. The correlation on these aggregates is called the ecological correlation. Similarly, the inference from the aggregated to the individual level data is called the ecological inference. In the individual correlation the variables are descriptive attributes of individuals, such as income, voting for a certain option etc. But in the ecological correlation, the statistical object is a group of persons, such as a community or a precinct. Therefore, variables are being dealt with such as the percentage of voters or the percentage of rich people within certain area aggregates. The main reason for using aggregated data to infer about individual behavior lies is the fact that one is often interested in individual behavior when only aggregates are available. This is especially true in social sciences. Furthermore, this problem occurs when using census data aggregated at particular level (a set of households, enumeration districts, communities). Thus, the use of ecological inference is relatively common. The following are some typical examples:

- There is a positive correlation at the community level between the percentage of voters for the labour party X and the percentage of rich voters. *Therefore, the rich voters are inclined to vote for party X.*

- The more school-girls there are in a class, the better the average mark. *So, the girls are doing better than the boys.*

[3] "Ecos" means space, area.

- The higher the percentage of the unemployed in a region, the higher the criminal rate. *So, the unemployed are inclined to commit crimes.*

- In cities with high air pollution disease Y is found more often. *Therefore, the polluted air affects a person with disease Y.*

It can be noticed from the above examples that the first sentence refers to the correlation among the aggregates, but the second one (*in italics*) switches to the individual level. It can be easily shown that all these conclusions might be wrong. For example, in the first case with rich voters and party X (let's say, the labour party) it is possible that the vote shares for party X are higher in communities where the percentage of people employed in industry is also high. And due to the industry, these communities are also richer with greater percentages of rich voters. On the other hand it is plausible that within each community the rich people themselves are inclined to vote for the conservative party and not for party X. Therefore, in spite of positive correlation between richness and the vote share for party X on the community level, on the individual level the correlation might be just the opposite.

4 Ecological fallacy

Robinson (1951) was among the first - at least in social sciences - to warn about the danger of such inference. He also named an impressive amount of literature where this kind of inference was extensively used. He called the failure of such an inference an ecological fallacy.

The ecological fallacy will be illustrated with numerical example consisting of four precincts with 200, 340, 333, 400 voters. At the time 1 there are two voting options X and X', and at the time 2 there are options Y and Y'. Two different assumptions are considered (CASE 1 and CASE 2) as to the behavior on the individual level (Table 2). For example, number 80 in the first precinct for CASE 1 means there are 80 voters who vote both for option X in the first election and for option Y in the second. But in CASE 2 there are only 30 voters with this kind of behavior.

Table 2: AN EXAMPLE OF ECOLOGICAL FALLACY - CASE 1 AND CASE 2

PRECINCT	CASE 1			CASE 2			PERCENTAGES	
							%X	%Y
		Y	Y'		Y	Y'		
PRECINCT 1	X	80	20	X	30	70	50	50
	X'	20	80	X'	70	30		
		Y	Y		Y	Y		
PRECINCT 2	X	120	30	X	70	80	44	56
	X'	70	120	X'	120	70		
		Y	Y'		Y	Y'		
PRECINCT 3	X	80	20	X	45	55	30	70
	X'	153	80	X'	188	45		
		Y	Y'		Y	Y'		
PRECINCT 4	X	80	20	X	30	70	20	80
	X'	320	80	X'	370	30		
		Y	Y'		Y	Y'		
<u>SUMMARY TABLE</u>	X	360	90	X	175	275	33	67
	X'	563	360	X'	748	175		

It can be observed that the margins are identical (20+80=30+70=100,...) in both cases as are the marginal percentages %X and %Y. Thus, the ecological correlation - that is, the Pearson correlation coefficient between variables %X and %Y - is equal in both cases: r_e=r(%X,%Y)=-1. However, the individual Pearson correlation coefficients[4] calculated from the summary tables are r_1=0.2 in the first case and r_2=-0.4 in the second. Alternatively, measured with the Youle coefficient, there are associations Q_1=0.4 and Q_2=-0.7. Obviously, with the ecological correlation r_e=-1 at the aggregated level there are amazingly different situations at the individual level. On the other hand, the methods for the ecological inference - based on the same percentages %X, %Y - would generally give the following solution[5] (CASE 3) for the individual summary table:

Table 2A: SUMMARY TABLE - CASE 3

	Y	Y'
X	0	450
X'	923	0

[4] In table 2x2 the Pearson coefficient is equal to Cramer's coefficient V.

[5] There could be some minor differences among the methods (both zeros could be replaced with some small numbers), but all the methods would infer that almost all the voters of the option X move to the option Y' in the second elections.

Thus, it is obvious that the ecological inference could be extremely risky. In fact, the aggregated data allows different solutions at the individual level, so some additional information or some additional assumption (implicit or explicit) is required for the ecological inference.

Robinson (1951) established that the link between the ecological (r_e) and individual (r) correlation[6] is the following:

$$r_e = k_1 * r - k_2 * r_w,$$

where r_w is the weighted average of the individual correlation within area units and k_1, k_2 are constants dependent on the clustering effect and the coefficient of variation. It can be clearly seen that the individual (r) and ecological correlation (r_e) need not be the same. It can be further shown that the ecological correlation is generally higher than the individual correlation. The discrepancy tends to increase with higher level of aggregation, that is, with a smaller number of (larger) area units. The conclusion is straightforward: the ecological correlation can not be a substitute for the individual correlation[7].

5 The regression approach

After the pessimism of Robinson's article, Goodman (1953) proposed the regression approach instead of the correlation one criticized by Robinson. The idea of the regression approach is very simple. The following shows the data on k=1,2..K area units, that is k=1,2..K tables with known margins X_{jk}, Y_{ik} and unknown cell values T_{ijk}:

Table 3: STRUCTURE OF THE AGGREGATED DATA

variable at time 2 (Y)

	1	...	j	...	J	total
variable at time 1 (X)	T_{11k}	...	T_{1jk}	...	T_{1Jk}	X_{1k}
	T_{i1k}	...	T_{ijk}	...	T_{iJk}	X_{ik}
	T_{I1k}	...	T_{Ijk}	...	T_{IJk}	X_{Ik}
total	Y_{1k}	...	Y_{jk}	...	Y_{Jk}	n_k

[6] The Pearson product-moment coefficient is used as the measure of the correlation.

[7] However, the situation is much better if - instead of the simple Pearson correlation coefficient - logit correlations are used at the aggregated level and the tetrachoric correlation at the individual level (section 6).

It can be assumed that each area unit has n_k voters. The sum of the area totals gives the sum of all voters N, $N=\Sigma(n_k)$, k=1..K. Within each unit the voters have options i=1..I at time 1 and options j=1..J at the time 2. The margins of the summary table (X_j, Y_i), $X_j=\Sigma X_{jk}$, $Y_j=\Sigma Y_{jk}$ (that is, the national totals) are also known:

Table 3A: SUMMARY TABLE

variable at time 2 (Y)

	1	...	j	...	J	total
variable at time 1 (X)	T_{11}	...	T_{1j}	...	T_{1J}	X_1
	T_{i1}	...	T_{ij}	...	T_{iJ}	X_i
	T_{I1}	...	T_{Ij}	...	T_{IJ}	X_I
total	Y_1	...	Y_j	...	Y_J	N

One wants to estimate the inner cell values of the summary table (Tij), $T_{ij}=\Sigma T_{ijk}$. Let the quantity $p_{ij}=E(T_{ijk}/X_{ik})$ denotes the probability that the voter who selected option i at time 1, also selected option j at time 2. It is assumed that the conditional probabilities $p_{ij}=E(T_{ijk}/X_{ik})$ are the same for all individuals and for all units. Thus, the equations are as follows:

$$Y_{jk} = p_{1j}{*}X_{1k} + p_{2j}{*}X_{2k} + \ldots + p_{ij}{*}X_{ik} + \ldots p_{Ij}{*}X_{Ik},\ j=1,2..,J$$

Of course, there are possible violations of the assumptions needed for the regression (multicolinearity, heteroscedascity, normality). Besides that, the regression coefficients themselves may lay outside the 0-1 interval. Researchers tried to improve this approach by putting some constrains on the regression or by adding some explanatory area variables. However, the success was rather limited. The same is true with the attempt to create homogeneous regions of area units[8]. However, the most serious problem of the regression approach is the misspecification of the individual voting behavior.

[8] The Slovenian example is instructive: the area units (4000 precincts) are small enough, so that the regional analysis within 62 communities (with about 70 precinct) can be done. However, the regression coefficients within communities also lay outside the 0-1 interval. So, in this aspect, the regionalization bring no improvements.

6 The logit approach[9]

The logit approach belongs to the family of latent structure models founded by Lasarsfeld (1950), however, the application to the ecological inference was developed by Thomsen (1987). The logit transformation L(X) for the probability P(X) and its inverse

$$L(X)= \ln(P(X)/(1-P(X)), P(X)= e^{L(X)}/(1+e^{L(X)})$$

is a very common transformation in an effort to keep the probabilities within 0-1 interval. There are some additional advantages of the logits: it can be shown (Thomsen 1987, p. 20) that the odds ratio model (P(X)/(1-P(X)) is much better for modeling the party swing across the area units between the two elections than the simple additive model or multiplicative model.

The basic idea of the logit approach will now be presented. For that reason, the example with only two options in each election will be used. An important advantage of the logit approach is the fact that it uses a consistent theory for the behavior of the individual - the well known Rasch model. Thus, one should start with a model for the probability of voter j choosing option (party) t:

$$P(X_{jt}=1/\theta_j,w_t) = e^{\theta j+w_t}/(1+e^{\theta j+w_t}).$$

Here, θ_j is the position of the individual (subject) j on the latent scale and w_t is the position of the party (stimuli) t on the same scale. The specific objectivity, which is a basic property of the Rasch model, has a particular meaning: voting behavior (transition probabilities) is independent of the individual behavior. In the same way, after aggregation, the voting behavior is independent of the area unit.

Similarly, through some derivations, a model[10] for the probability of selecting option X at time 1 and option Y at time 2 could be expressed. It is reasonable to assume the party's position as fixed except for one dimension - the party's popularity ($ß_o$,α_o). It is also assumed that individuals can be located in the same latent space θ:

$$P(X/\theta)=e^{\alpha_o+\alpha\theta}/(1+e^{\alpha_o+\alpha\theta}),\quad P(Y/\theta)=e^{ß_o+ß\theta}/(1+e^{ß_o+ß\theta}), \qquad (*)$$

where $ß_o$ and ß are the attributes of party Y. Similarly, α_o and α_θ are the attributes of party X. Furthermore, it is assumed that within each area unit, the latent variable θ, with dimension of R, is distributed normally:

$$\theta=N_R(\mu,\sigma).$$

[9] This presentation follows Thomsen (1987), ch.3.

[10] It is called "the model for distribution analysis" (Thomsen 1987, p. 132).

It is also assumed that mean values μ are distributed across K area units normally:

$$\mu=N_R(0,I).$$

With these assumptions it can be shown through some derivations - using logit and probit transformation - that the probability of the transition from party X at time 1 to party Y at time 2 can be expressed as an integral function F(.) of P(X),P(Y) and τ_t:

$$P(Y/X)=F(P(X),P(Y),\tau_t),$$

where P(X) and P(Y) stand for the marginal distributions across area units - that is, the percentage of voters for option X and Y - which are all known. Here τ_t is the total individual correlation between variables P(X) and P(Y) which are - as has been seen in (*) - functions of the latent variable q. Quantity τ_t is also called the tetrachoric correlation. This correlation assumes that binary categorical variables - what is available in the data - come from dichotomization of the continuous latent variables and the tetrachoric correlation is the correlation of these two latent variables.

The key step in the logit approach is the substitution of the τ_t with the Pearson correlation coefficient r_{logit} calculated from the logit transformations of the known margins P(X), P(Y) of the area units. For that substitution to be valid, some additional assumptions are necessary:

i) *Homogeneity of units*: The area units should be analyzed in the homogeneous regions.

ii) *Isomorphism of the latent variable*: All the components of the latent variable should have the same structure, that is:

$$Var(\theta_1)/Var(\mu_1)=Var(\theta_2)/Var(\mu_2)=...=Var(\theta_R)/Var(\mu_R)=C.$$

In other words, the ratio of the within area unit variability (Var(θ_i)) to the between-area unit variability (Var(μ_i)) should be constant. Moreover, this ratio should be high for the logit approach to be valid without additional complication. This means that only a minority of the variation should be the result of the differences between the area units.

Under these two assumptions the tetrachoric correlation, τ_t, can be replaced with r_{logit}, the correlation on logit transformations. However, it is known from the theory that the tetrachoric correlation τ_t can be approximated with the Youle coefficient Q, therefore:

$$r_{logit} \equiv \tau_t \approx Q = (ad-bc)/(ad+bc).$$

Thus, the procedure is relatively simple; one should calculate logit transformation for the marginal values P(X) and P(Y), which are known for every area unit, and then calculate the Pearson correlation coefficient based on these values. The value of r_{logit} is used as the coefficient Q, which enable straightforward solution for the unknown cells (a,b,c,d) of the 2x2 summary table.

The basic idea of the logit approach described here also holds for more than two options in each election. In this case linear multiple logit model for individual choice is used which is a straightforward generalization of the binary choice model. However, there are some serious complications in the computations of the cell values. An iteration algorithm is needed to overcome that and one of the categories should be selected as the basic one. By default, this is the category of the voters who have abstained.

Besides the regression and logit approach, there are also other techniques which offer some promising results: the model of maximal entropy (Johnston, Pattie 1991), the aggregated compound multinominal model (Cleave, Browne, Payne 1991) and the component model (Kohlsche 1991).

7 Results

The results of the logit approach were obtained with software ECOL (Thomsen 1991). The results of the regression approach were obtained with an ordinary least square regression.

a) The logit approach

In Table 4 (also table 10 column ECOL 2) the transition probabilities[11] can be observed which were obtained by the logit approach when precincts were used as the area unit. The basic discrepancy with other models lies in the loyalty of the voters who abstained. Here, in the logit approach, only 65% of the voters who had abstained for the first round abstained for the second round, too.

The summary Table 10 also indicates there are some differences when using communities as the level of analysis (Table 10 columns ECOL 0 and 1) or when separately analyzing the precincts within each community - regionalization (Table 10 column ECOL 3).

Table 4: TRANSITION PROBABILITIES - THE LOGIT APPROACH (ECOL 2)

I. II.	KUČAN	PUČNIK	ABSTAIN	OTHER	TOTAL
KUČAN	89.0	2.0	8.7	0.1	100.0
PUČNIK	9.1	76.7	14.0	0.2	100.0
DEMŠAR	44.7	31.5	23.9	0.2	100.0
KRAMBERGER	10.4	59.7	29.5	0.4	100.0
ABSTAIN	21.7	12.8	65.1	0.4	100.0

[11] The category of OTHER includes the voters who were eligible in only one election (movers, new voters, deaths). They amount for less than 0.2% of all voters, so they are negligible in our analysis.

b) Regression approach

Table 5 shows[12] the results obtained with the regression approach (also table 10 column REG 4). As some of the corresponding coefficients lay outside the 0-1 interval, the transition probabilities for these values were corrected with an arbitrary expert model[13]. However, these are only minor changes what can be seen when compared with raw regression coefficients multiplied by 100 (table 10 column REG 3). Table 5 implies that the general conclusions are basically the same as in the case of the logit approach, however, some noticeable differences can be observed. The biggest discrepancy is the loyalty of the abstaining voters - 97% compared to 65% in the logit approach.

Table 5: TRANSITION PROBABILITIES - THE REGRESSION APPROACH (REG 4)

I. II.	KUČAN	PUČNIK	ABSTAIN	TOTAL
KUČAN	93.9	1.0	5.1	100.0
PUČNIK	1.0	97.6	1.5	100.0
DEMŠAR	62.0	30.5	7.5	100.0
KRAMBERGER	25.2	46.2	28.6	100.0
ABSTAIN	2.1	1.1	96.7	100.0

c) Evaluation from surveys

It is only through the survey results that the ecological inference can be evaluated. Of course, the surveys themselves are also subject to severe methodological problems: attrition, memory effect, nonresponse and measurement problems. It is especially difficult to deal with abstaining voters as they are often linked to nonrespondents. In our case, there is an additional disadvantage: there has been no specific survey made to estimate the transition probabilities.

c1) Slovene Public Opinion (SPO). Slovene Public Opinion is a regular survey using standardized methodology carried out by the Institute of Social Sciences. The data used were obtained from the second mail follow-up[14] survey (March 1990) of the face-to-face survey of January, 1990. The response rate was relatively good and there was no discrepancy in any control variable among 1265 person included in the second follow-up and the initial sample of 2148 persons. There is, however, a serious problem in the second round as there is no "won't vote (=abstain)" option (Table 6).

[12] Regression results are presented without the category of nonvoters, which is in our case - as mentioned - almost negligible.

[13] The corrections were made by Gunter Ogris, IFES, Vienna.

[14] The questions: "Who would you vote for in the presidential election if the election were today?" and "It looks like Kučan and Pučnik will come into the second round. If that happens, who will you vote for then?"

Table 6: SLOVENE PUBLIC OPINION DATA (RAW DATA)

I. II.	KUČAN	PUČNIK	DON'T KNOW	TOTAL
KUČAN	582	4	2	588
PUČNIK	5	296	0	301
DEMŠAR	39	23	19	81
KRAMBERGER	55	71	43	169
WON'T VOTE	5	3	16	24
DON'T KNOW	20	15	67	102
TOTAL	706	412	147	1265

To establish the comparison, the "don't know" and "won't vote" options were combined to the category "abstain". This can be justified with the fact that in the second round these voters have the same behavior which can be deduced from nearly identical raw percentages (calculated from Table 6) for the categories "don't know" and "won't vote". After combining the two, the table was adjusted (raked) to the true election outcome. In Table 7 (also Table 10 column SPO 2) the transition probabilities have been calculated from the adjusted table. Some noticeable differences can be seen when compared to the ecological results. For example, 80.1% of the abstaining voters from the first round also abstain in the second round. In Table 10, column SPO 1 the transition probabilities can be found before the process of ranking.

Table 7: TRANSITION PROBABILITIES - SLOVENE PUBLIC OPINION (ADJUSTED - SPO 2)

I. II.	KUČAN	PUČNIK	ABSTAIN	TOTAL
KUČAN	98.0	1.0	1.0	100.0
PUČNIK	1.4	98.0	0.3	100.0
DEMŠAR	33.8	26.3	39.9	100.0
KRAMBERGER	22.0	36.9	41.1	100.0
ABSTAIN	9.7	9.7	80.1	100.0

c2) Telephone survey. There was a small telephone survey performed by a private agency Varianta (Table 8). It was done in the time between the two elections[15]. For reason of comparability, the categories "don't know", "refuse to answer", "won't (didn't) vote" were combined to the category "abstain". The justification is more or less the same as in the case of SPO.

Table 8: TELEPHONE SURVEY (RAW DATA)

I. II.	KUČAN	PUČNIK	ABST.	DON'T K	REFUS.	TOTAL
KUČAN	125	0	1	8	1	135
PUČNIK	3	48	0	9	0	60
DEMŠAR	11	9	1	10	1	32
KRAMBERGER	4	9	0	4	0	17
ABSTAIN	7	1	7	4	0	19
REFUS.	7	3	0	27	5	42
TOTAL	157	70	9	62	7	305

After adjusting (raking) the raw data to the election outcome, the transition probabilities were calculated in table 9 (also in table 10 column TEL 2). In column TEL 1 (Table 10) there are transition probabilities calculated before the adjustment process. It can be observed that the transition probabilities are somehow closer to the logit result than to the regression result.

Table 9: TRANSITION PROBABILITIES - TELEPHONE SURVEY (ADJUSTED - TEL 2)

I. II.	KUČAN	PUČNIK	ABSTAIN	TOTAL
KUČAN	91	1	8	100
PUČNIK	4	83	13	100
DEMŠAR	30	33	37	100
KRAMBERGER	21	59	21	100
ABSTAIN	22	9	68	100

8 Discussion

In Table 10, a comparison of all transition probabilities can be observed. Some differences can be noticed when using regionalization (ECOL 3) or when using communities as the unit of analysis (ECOL 0 and ECOL 1). The same is true for the regression approach (REG 1 and REG 2), where only raw regression coefficients (multiplied by 100) are presented. However, there is no substantial difference within each method. Similarly, there is not much difference - with few exceptions - between raw (SPO1, TEL1) and adjusted (SPO2, TEL2) data in survey results. We can also

[15] The questions were as follows: "Can you tell us who you voted for in the first round of the presidential elections?" and "Who will you vote for in the second round?".

observe the sensitivity of the ecological analysis to the level of aggregation (community level: REG 1,2, ECOL 0,1,6, precinct level: REG 3,4 ECOL 2,3,4,5), regionalization (ECOL 3,5) and the use of incomplete data (data from 62 communities: REG 2, ECOL 0, data from 38 communities: REG 1,3, ECOL 1,2,3, data from 32 communities: ECOL 4,5,6). All the difference can be observed in Table 10.

Table 10: COMPARISON OF TRANSITION PROBABILITIES (%)

		REG				ECOL							SPO		TEL.	
		1	2	3	4	6	0	1	2	3	4	5	1	2	1	2
KRAMB	-> KUČAN	23	35	25	25	7	14	9	10	14	8	11	33	22	24	21
KRAMB	-> PUČNIK	24	29	44	46	63	56	63	60	53	58	53	42	37	53	59
KRAMB	-> ABSTAIN	47	38	32	29	30	30	26	30	33	33	36	25	41	24	21
DEMŠAR	-> KUČAN	80	75	62	62	37	33	37	45	49	46	51	48	34	34	30
DEMŠAR	-> PUČNIK	51	47	30	31	27	36	33	31	30	28	28	28	26	28	33
DEMŠAR	-> ABSTAIN	-45	-34	8	8	35	31	30	24	21	26	21	24	40	38	37
PUČNIK	-> KUČAN	-7	-9	-10	1	3	4	5	9	9	7	9	2	1	5	4
PUČNIK	-> PUČNIK	125	119	108	98	87	88	83	77	81	81	81	98	98	80	83
PUČNIK	-> ABSTAIN	-20	-11	2	2	10	10	12	14	11	12	11	0	0	15	13
KUČAN	-> KUČAN	101	100	98	94	97	96	97	89	88	92	89	99	98	93	91
KUČAN	-> PUČNIK	-6	-4	-4	1	0	0	0	2	3	1	3	1	1	0	1
KUČAN	-> ABSTAIN	5	5	6	5	3	4	3	9	9	7	8	0	1	7	8
ABST	-> KUČAN	-15	-12	2	2	18	20	18	22	20	19	18	20	10	23	23
ABST	-> PUČNIK	-12	-10	-3	1	4	4	4	13	13	11	13	14	10	7	9
ABST	-> ABSTAIN	131	125	100	97	77	75	77	65	67	70	67	66	80	71	68

REG	**1**	-	regression analysis on 38 communities
REG	**2**	-	regression analysis on 62 communities
REG	**3**	-	regression analysis on 2470 precincts
REG	**4**	-	adjusted results from REG 3
ECOL	**6**	-	logit analysis on 32 communities
ECOL	**0**	-	logit analysis on 62 communities
ECOL	**1**	-	logit analysis on 38 communities
ECOL	**2**	-	logit analysis on 2470 precincts
ECOL	**3**	-	logit analysis on 2470 precincts within 38 communities
ECOL	**4**	-	logit analysis on 2291 precincts
ECOL	**5**	-	logit analysis on 2291 precincts within 32 communities
SPO	**1**	-	raw (combined categories) data from SPO survey
SPO	**2**	-	adjusted data from SPO survey
TEL	**1**	-	raw (combined categories) data from telephone survey
TEL	**2**	-	adjusted data from telephone survey

To see the differences between methods more clearly, the distances are calculated between the matrices of different methods:

Table 11: DISTANCES BETWEEN TRANSITION MATRICES

METHOD	REG4	SPO2	TEL2
ECOL3	0.14	0.14	0.06
REG 4		0.09	0.14
SPO 2			0.12

The distances are calculated as the half sum of the absolute difference of the corresponding proportions for the two matrices. The proportions of the total voters are taken within certain cell, not the raw percentages. This distance also stands for the proportion of voters that need to be removed (replaced), to obtain identical distribution. It can be seen that the regression result is closer to the SPO survey result, but the logit result is closer to the telephone survey outcome. Besides that, the differences within the ecological results are noticeable (from 0.03 to 0.08) as are the distances between the adjusted and nonadjusted results of each survey: 0.06 for the SPO survey and 0.03 for the telephone survey.

There may be three straightforward reasons for the discrepancy in the results:

a) The methodology

a1) As mentioned, the ecological inference itself is suffering from the fatal problem of the ecological fallacy. Furthermore, it is difficult to test the assumptions. Besides that, there is also a serious problem in the logit approach concerning the sensitivity toward the selection of the basic category[16]. However, the regression approach has even more serious drawbacks. For example it can be seen that many regression coefficients lay outside the 0-1 interval. Finally, in the survey approach as well, one is faced with severe methodological problems.

a2) Besides making general methodological remarks, it should be specifically mentioned that there are some indices of method effect: the loyalty of voters is overestimated by regression and underestimated by logit method. This conclusion can be supported by the findings of the within-community analysis - the regression and logit analysis were performed for every community - where the same tendency was observed. It should be added that in both approaches, in the regression one in

[16] The "abstain option" was taken as the basic category. However, choosing other options for the basic one gives highly unacceptable results.

particular, there are large differences[17] in transition probabilities when calculated for each community. The problem of creating homogeneous regions is thus waiting for further research.

b) The data

b1) Due to some technical troubles the data on precinct level may be questionable as only two thirds (2470 precincts) of the data has been used for the ecological inference. Later, however, the calculations were made using three quoters of the data (2714 precincts) with no apparent differences in the results. In addition, the marginal proportions fit well into the election result at national level. It should be added that at the community level, the complete data was available.

b2) As no special survey was designed, oversimplifications must be made in combining the categories when adopting the survey results. Besides that, it should be remembered that there was a 30% attrition in the SPO survey. Furthermore, the telephone survey was relatively small, with a 30% telephone noncoverage in Slovenia.

c) The (un)stability of the voting process

Voting behavior and the political parties in Slovenia are extremely unstable and irregular. The mere appearance of such a "strange"[18] candidate as Mr. Kramberger and, moreover, his relative success (one fifth of the vote share), strongly confirms this. There are some indices that ecological inference (and the surveys inference also) work much better in a the stable political environment (the Scandinavian countries, the Austria).

It is the opinion of the author that this reason (c) together with some specific features of both methods (1b) and some violations of the assumptions are the dominant factors for the differences in the results. Nevertheless, the general methodological drawbacks (a1) and shortcomings of the data (b), exist, however, they can not be the main reasons for the discrepancies.

In spite of some differences in the results, the conclusions obtained are relatively promising. Further research should concentrate on the following:

- implementing other methods for ecological analysis (especially the aggregated compound multinominal model),
- testing of the assumptions,
- creating the homogeneous regions,
- close observation of possible interactions in the data.

[17] This implies that there is a lot of variability among area units (communities). Thus, the assumptions for the logit approach may be violated at least when the community is used as the level of analysis.

[18] Mr. Kramberger was a kind of a popular person, a "self-made man", a joker, far from being a professional politician in any sense.

9 Conclusions

a) Some basic conclusions can be made which are the same for all methods. Also, they are highly plausible from the substantive point of view:

• Demšar voters more often chose Kučan in the second round. However, to an even greater extent Kramberger voters selected Pučnik.

• About one third of the Kramberger and also one third of the Demšar voters were abstained in the second round showing much higher amount than the voters of the other two candidates (below 10% abstained in the second round).

The biggest discrepancies are found in the estimations of loyalty options: they are high for the regression and low for the logit approach. The survey results are somewhere in between but closer to the logit approach. There are also some other differences, especially in the transition probabilities for the Demšar voters. The differences within each method are moderate with few exceptions.

b) It may not be said that the survey results are obviously better than the ecological results or that the results from the logit approach are uniformly better than that of the regression. All of the results obtained by different methods should be treated - together with their methodological meaning - as a part of one picture which offers something new about the unknown reality. It should not be forgotten that this is an attempt to discover what is otherwise unknown by the definition - the voting behavior of the individual in the entire nation. It should be stressed (Thomsen 1987, p. 37), that the ecological inference is similar to the time prediction methods, the difference being that the forecasting across the levels and not across time. Failure to create a good forecast should not prevent one from further development of methodology. Instead, the methods should be studied and empirical verification should be made.

References

[1] Aarts K., and Horstman R. (1991): Political change and the electoral geography of the Netherlands. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[2] Berglung S., and Thomsen S. (1990): Modern Political Ecological Analysis. Abo: Abo Academy Press.

[3] Brown P., and Cleave N., Payne C. (1991). Methods of ecological inference: An Empirical Test and a Cautionary Tale. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[4] Brown P., Cleave N., and Payne C. (1991): Evaluation of Methods for Ecological Inference. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[5] Corbetta P. (1991): Estimating Electoral Flows in Italy from Aggregate Data: A test of Some Pre-Conditions. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[6] Ersson S., and Worlund I. (1991): Swedish Communism 1920-1990: An Ecological Analysis. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[7] Forcina A., and Marchetti G.M. (1991): Effects of misspecification in Brown and Payne model. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[8] Goodman L. (1953): Some Alternatives to Ecological Correlations. *American Journal of Sociology*, **64**, 610-625.

[9] Immerfall S. (1991): Macrohistorical Models in Electoral Research: A Fresh look at the Stein-Rokkan Tradition. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[10] Johnston R., and Pattie C. (1991): Dealigment, Spatial Polarization and Economic Voting: Exploration of Recent Trends in British Voting Behavior. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[11] Kohlsche A. (1991): Estimation of Voting Migration and Vote Splitting. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[12] Krauss F. (1991): Comparison of the results of different ecological models concerning voting transitions in the Federal elections 1990. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[13] Lasarsfeld P. (1950): The Logical and Mathematical Foundation of Latent Structure Analysis. In: Stoufler S. (ed.) *Measurement and Prediction. Studies in Social Psychology in World War II.* Princeton: Princeton University Press.

[14] Payne C. (1991): Election Forecasting in the UK 1970-1990. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[15] Rasch G. (1968): *A Mathematical Theory of Objectivity and Its Consequences for Model Construction.* Copenhagen: Institute of Statistics, University of Copenhagen.

[16] Robinson W.S. (1951): Ecological Correlation and the Behavior of Individuals. *American Sociological Review*, **15**, 351-357.

[17] Schadee H. (1991): What to do with transition matrices once you have got them: A case study of elections in Verona and Padova 68-87. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[18] Thomsen S. (1987): Danish Elections 1920-1979. A logit Approach to Ecological Analysis and Inference. Arhus: Politica.

[19] Thomsen S. (1991): ECOL, Software for the ecological inference.

[20] Thomsen S. (1991): Comparative Electoral Dynamics. Paper presented at the ECPR joint session of workshops, Essex, March 1991.

[21] Visser M., and Horstman R. (1991): Religion and Voting Behavior: An Ecological Inquiry. Paper presented at the ECPR joint session of workshops, Essex, March 1991.