Estimating error rates in an administrative register and survey questions using a latent class model

DL Oberski*

Tilburg University, The Netherlands

Abstract

Administrative register data are increasingly used in a number of countries, including the Netherlands, Denmark, Sweden, Norway, the United Kingdom, Israel, Germany, and New Zealand, to replace or supplement the census, and are thought to provide a cost-effective opportunity for longitudinal full-population data analysis in the social sciences. They are also frequently used as "validation data" to study measurement error in survey questions. In spite of quality control procedures, however, there are strong indications that administrative register data can themselves contain considerable measurement error. Moreover, typically the error process does not conform to classical measurement error models. Such errors negate the potential usefulness of administrative data, making it essential to evaluate their extent. This chapter discusses latent variable modeling as a way to estimate measurement error in administrative data by combining error-prone administrative data with an error-prone survey. To demonstrate the approach, a latent class model is applied to linked register-survey residence data from the municipality of Amsterdam.

1 Introduction

Administrative data obtained from government registers provide a wealth of potential for the social sciences (Entwisle and Elias, 2013). Collected during the normal course of public administration, for example to tax, keep track of car ownership, pay welfare benefits, or send out calls to vote in elections (Wallgren and Wallgren, 2007), the variables available in registers can be used by survey researchers as direct variables of interest, or as auxiliary variables in survey sampling, nonresponse adjustments, or validation studies. Such registers are often available longitudinally and for the entire population – a highly attractive combination for researchers, especially when linked to purpose-designed surveys.

*Thanks are due to Robert Selten from the *O+S* research service of the municipality of Amsterdam for providing the tabular data and for clarifying the design of the Sportmonitor 2013 survey, and to Brady West and Clyde Tucker for their useful comments. This work was supported by the Netherlands Organization for Scientific Research (NWO) [Veni grant number 451-14-017].

While administrative registers have many redeeming qualities for survey researchers, the fact that they have been collected for administration and not research can be a disadvantage. In particular, administrative registers may contain considerable measurement errors, including definition, reporting, timing, processing, editing, linkage, and coverage errors (Bakker, 2009; Groen, 2012). For example, Gomez and Glaser (2006) found that a staggering 83.3% of Native Americans and 29.9% of Hispanics were misclassified as a different race in U.S. registers, and epidemiologists have long preferred "gold standard" autopsy data to hospital cause-of-death registers (e.g. Maudsley and Williams, 1996). Since measurement error is well-known to severely distort analyses of substantive interest (Fuller, 1987; Carroll et al., 2006), estimating the extent of such errors is essential to making administrative register data useful for social research.

Some validation studies of administrative registers exist, but often use surveys as the gold standard (e.g. Gomez and Glaser, 2006; Groen, 2012). At the same time, survey methodologists use administrative registers as a gold standard to validate surveys (e.g. Kreuter et al., 2010; West and Olson, 2010). It therefore seems likely that both sources are error-prone. But how can we estimate the extent of measurement errors in both administrative register data and survey answers when neither are perfect?

This chapter demonstrates one approach to doing so: latent variable modeling. As an example it estimates the amount of classification error in survey and administrative measures of the neighborhood of residence, an important modeling and adjustment variable. From a combination of two survey measures of the neighborhood and one administrative register measure, a latent class model is built that accounts for survey mode effects as well as local dependencies between the survey measures. This enables estimation of error rates in all three measures without assuming any one of them to be perfect in advance.

Latent variable modeling has been applied to estimating measurement error in continuous administrative variables by Bakker (2012), Scholtus and Bakker (2013) and Oberski et al. (2013a) using "multitrait-multimethod" (MTMM) designs. Pavlopoulos and Vermunt (2013) used a longitudinal latent class (hidden Markov) approach to estimate misclassification in employment status registers. This previous work has not investigated, however, how misclassification may be estimated in nominal administrative and survey variables without longitudinal information. Moreover, the present application accounts for mode of data collection effects in the survey measures. Modeling these mode effects allows for the identification of both random classification errors and method effects without the need for multiple "traits" (true values). It could therefore be termed a "single-trait-multi-method" (STMM) approach.

The following section describes the data on neighborhood of residence obtained from a survey and an important Dutch official administrative register. Section 3 then details the latent class model built to estimate classification error rates in these measures. The resulting estimates are described in Section 4, after which Section 5 concludes.

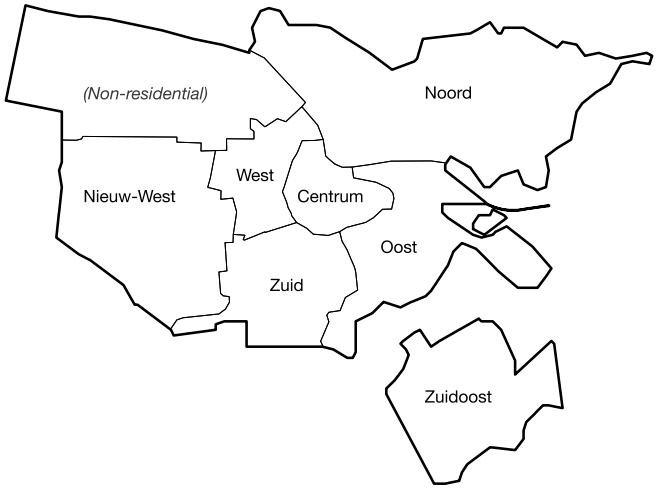

Figure 1: Neighborhoods of residence ("stadsdeel") in Amsterdam, The Netherlands.

2 Administrative and survey measures of neighborhood

The Dutch municipal register is one of the bases of the Dutch official statistical system, as well as a crucial source of information for various Dutch government services. It is collected at the municipal level but used at the national level as input to construct the virtual census, draw samples for official, academic, and commercial probability surveys, and to divide welfare transfers and decide local public policies. The GBA's accuracy is therefore under heavy scrutiny, as evidenced by the 2007 official instruction by Parliament to "improve accuracy to at least 99%"[1]. This instruction has prompted a series of audits, in which investigating officers with special powers are trained by local governments to investigate and visit addresses until the accurate values for the corresponding GBA record can be verified with certainty, a very costly effort.

Our target variable of interest is the neighborhood of residence or "stadsdeel", a key variable in data analysis for public policies, for various "big data" services linking neighborhood data to other official statistics, and for sample survey design. Amsterdam contains eight such neighborhoods, of which one is the port, a non-residential area. These are shown in Figure 1. To give the reader an impression of these areas, some descriptive statistics are provided in Table 2. For example, Zuid is the most affluent neighborhood and Zuidoost the least at the time of writing. The municipality's expensive audit estimated the inaccuracy of the neighborhood of residence to be around 1.7%; one point of interest is whether this estimate could have been arrived at more cheaply.

As it happens, neighborhood of residence was also measured in a survey on a subsample of the population. In March and April of 2013, the municipality of Amsterdam, The Netherlands, performed its biannual survey on sports participation, the "Sportmonitor"[2], among

[1] Dutch Parliament, Motion 31 200 VII-34.

[2] `https://www.amsterdam.nl/publish/pages/422998/sportmonitor_2013_def.pdf`

Table 1: Some official statistics on neighborhoods of Amsterdam (2011–2014). Source: http://www.ois.amsterdam.nl/feiten-en-cijfers/stadsdelen.

		% Population	Income (€1000's)	% Unemployed	% ≥ College degree
A.	Centrum	11	30	9	57
E.	West	17	23	11	45
F.	Nieuw-West	18	19	13	24
K.	Zuid	17	31	6	53
M.	Oost	16	24	9	44
N.	Noord	11	19	12	21
T.	Zuidoost	10	18	13	19

[*Survey: direct question*]. **In which neighborhood do you live?**

1. Centrum
2. Westpoort
3. West
4. Nieuw-West
5. Zuid
6. Oost
7. Noord
8. Zuidoost
9. Don't know

[*Survey: postcode*]. **What is your postcode?** _ _ _ _ [_ _]

Figure 2: Formulation of the 2013 Sportmonitor survey questions asking neighborhood of residence. The postcode (ZIP code) is asked at the most detailed level possible.

the population of approximately 711,000 residents aged 6–74. A random sample of 20,579 households was taken from the municipal register, stratified by neighborhood, ethnicity, and age group (< 18 or $18+$). The data collection was a sequential mixed mode design: first, all sampled persons in the households received a personalized letter via regular mail with the official logo of the municipality, requesting participation in an online web survey but leaving the option of returning a paper form, which a small (3%) percentage of respondents returned. A sample of nonrespondents were followed up by CATI if a phone number was available or visited at their residence and interviewed using CAPI otherwise. This led to a mix of 59% web or paper-and-pencil mode, 30% CATI, and 11% CAPI. The final number of respondents was 4,266, making the response rate 4,266/20,579 = 20.7% (AAPOR RR1 and RR2).

Figure 2 shows the two survey questions about neighborhood. The question first directly asks the respondent their neighborhood of residence, while the second asks it indirectly through the postcode. Of note is the fact that the survey asks the postcode at a high level of detail: it is common for surveys to ask only the first four digits rather than all six. This may have contributed to the higher number of missing values in this variable – 7.7% of values are missing in the second question while only 0.9% of values are missing in the first.

Table 2: Cross-tabulation of the observed survey measures with the administrative measure. A: Centrum, E: West, F: Nieuw-West, K: Zuid, M: Oost, N: Noord, T: Zuidoost.

	Survey: direct							Survey: postcode						
	A	E	F	K	M	N	T	A	E	F	K	M	N	T
Register														
A	432	5	1	0	3	0	0	412	1	0	0	0	1	0
E	1	356	11	1	2	0	1	0	351	0	2	2	0	1
F	0	51	341	1	0	2	0	0	0	372	0	0	2	0
K	8	5	3	1669	2	1	8	0	2	1	1536	0	0	1
M	2	5	0	1	402	2	0	0	2	1	0	386	1	0
N	0	0	1	1	1	441	0	0	0	1	1	0	422	0
T	0	0	0	0	0	3	355	0	0	0	0	0	0	340

Combining the survey with the administrative register, we have four observed variables: the direct survey question with the postcode-derived measure shown in Figure 2, the administrative register, and the survey mode. The table of cross-classifications of these variables was provided by the research service of the municipality of Amsterdam. It should be noted here that we did not obtain any postcode values, but only the derived neighborhood[3]. Since each of the residential neighborhoods is estimated to be inhabited by at least 84,000 people, privacy is not a concern for this cross-tabulation. Such concerns did prevent a larger level of detail from being made available, so that it was not possible to include further covariates in the analysis.

Table 2 cross-tabulates the two survey measures (columns) with the administrative register. The table shows that the correspondence among these measures is rather large, as may be expected with this variable. However, it is imperfect. For example, in the direct question 51 respondents (13%) who live in Nieuw-West according to the register claim to live in the more affluent West neighborhood. The postcode-derived measure indicates this as well, so that the most likely explanation here is measurement error in the direct question. However, there are also 3 respondents ($\sim$ 1%) in this group who live in entirely different non-adjoining neighborhoods according to both survey measures; such cases may indicate that the administrative register, too, contains some classification error. Alternatively, this may be a case of consistent mistake-making on the part of the respondent–it is impossible to tell purely from these cross-tables. It is this difficulty of drawing conclusions by inspecting Table 2 that impels us to use a different tool: latent class models.

[3]The translation from postcode to neighborhood was performed by the research service of the municipality. Although we expect errors to be rare, in principle it is possible that this additional step forms a source of classification error. That is, not all classification error in the postcode measure is necessarily misreporting of the postcode–some of it may be due to coding errors.

3 A latent class model for neighborhood of residence

From the three observed measures and the survey mode indicator, our goal is to estimate the classification error in each of the measures. Since none of the three measures are perfect, a latent variable model is used in which the unobserved ("latent") true neighborhood is measured with error by the three indicators. Such models in which the latent variable is categorical are known as "latent class models" in the literature (e.g. Hagenaars and McCutcheon, 2002) and have been applied to the estimation of classification error in survey data by Lazarsfeld (1950), McCutcheon (1987) and, more recently, Biemer (2011) and Oberski (2013).

The standard latent class model assumes one single latent class variable, S, say, conditionally upon which the observed indicators y_1, y_2, and y_3 are independent. For example, in our seven-neighborhood case, a seven-class model might be formulated:

$$P(y_1, y_2, y_3) = \sum_{s=1}^{7} P(y_1|S)P(y_2|S)P(y_3|S)P(S). \tag{1}$$

Responding "Don't know" or not providing a postcode can simply be modeled as separate categories of the observed variables. However, this standard model implies that the two survey measures are independent of each other given the latent class, $P(y_2, y_3|S) = P(y_2|S)P(y_3|S)$. This "local independence" assumption will be violated when respondents who do not answer the direct survey question also do not (wish to) provide their postcode; or when the classification error differs over survey mode. For a more general discussion of possible reasons for local dependence, see Biemer (2011, Sec 5.2).

In order to account for possible local dependence between the survey measures, we allow for a "method factor": a latent variable that represents an individual tendency to answer in a certain way, independent of the true neighborhood of residence. Note that this means the "method factor", instead of being a systematic bias, is a stochastic effect shared by the two survey questions. For example, this unobserved variable may represent a person's tendency to keep their neighborhood secret from the researcher. A category of the method factor could also represent a tendency to report living in a particular neighborhood, such as Zuid, regardless of whether one actually lives there. To investigate the effect of survey mode, we also include the survey mode chosen as a covariate in the model. Since mode is not randomized, it may correlate with the true neighborhood of residence. We model this dependency as a direct effect from survey mode to true neighborhood[4]. There is no reason to suppose, however, that survey mode affects the measurement of the administrative register, an assumption that allows the identification of the method factor. The resulting model is shown in Figure 3.

From Figure 3, the likelihood of the observed measures y_1, y_2, and y_3 given the fixed "survey mode" covariate x is

$$P(y_1, y_2, y_3|x) = \sum_{s=1}^{7} \sum_{m=1}^{2} P(y_1|S)P(y_2|S, M, x)P(y_3|S, M, x)P(S|x)P(M|x), \tag{2}$$

[4]Note that this formulation implies the same set of conditional independencies (d-separation) as the model in which the effect $x \rightarrow S$ is reversed to $x \leftarrow S$, and is therefore equivalent to that model. However, the current formulation allows x to be considered fixed and is therefore more convenient computationally.

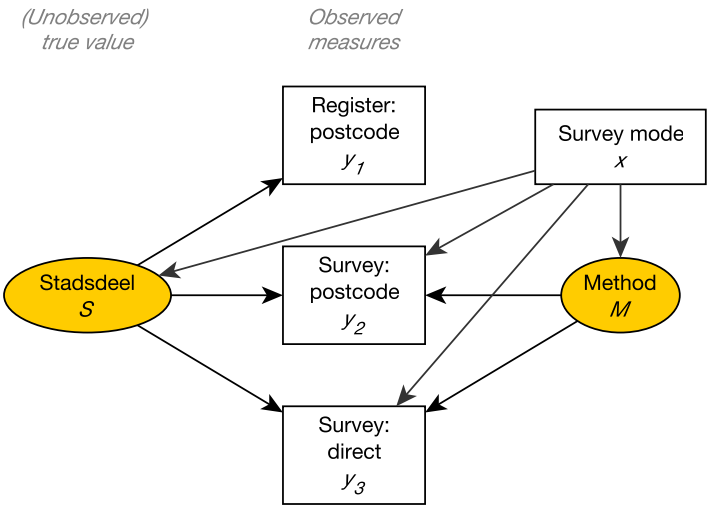

Figure 3: Latent class measurement model for the three measurements of neighborhood of residence. Latent variables are shown in shaded ovals, while observed variables are shown as rectangles.

where we have chosen two categories for the latent method factor. That is, the likelihood is that of a latent class model with two separate latent class variables, one representing the true neighborhood of residence S and the other the respondent's answer tendency in the survey M. Given the latent S, the observed register y_1 is assumed independent of (y_2, y_3, x), while y_2 and y_3 are only mutually independent given S, M, and x. The latent "trait" S and "method" M are assumed independent given the survey mode.

The conditional probabilities on the right-hand side of Equation 2 are parameterized using a logistic regression model. For instance, for the survey postcode question, the conditional probability is modeled as a logit with main and second-order interaction effects from its categorical regressors,

$$P(y_2 = k|S = s, M = m, x) = \frac{\exp(\tau_k + \lambda_{ks} + \lambda_{km} + \gamma_{kx})}{\sum_{k'} \exp(\tau_{k'} + \lambda_{k's} + \lambda_{k'm} + \gamma_{k'x})}, \tag{3}$$

where

- τ_k is a logistic intercept that determines how many respondents choose neighborhood k on variable y_2 overall,
- λ_{ks} is a logistic slope for the true neighborhood. It corresponds to the log-odds of the respondent choosing neighborhood k on y_2 given that the true neighborhood is s,
- λ_{km} is a logistic slope for the method factor. It corresponds to the log-odds increase in claiming to live in neighborhood k on y_2 within method class m, and
- γ_{kx} is a logistic slope for the survey mode. It corresponds to the log-odds increase in claiming to live in neighborhood k on y_2 for survey mode x.

As in multinomial logistic regression, the reference category parameters are set to zero for identification purposes, $\tau_1 = \lambda_{1.} = \gamma_{1x} = 0$; thus, the log-odds increases are relative to these chosen reference categories. It is, in principle, also possible to incorporate higher-order interactions. For instance, the logistic "loading" λ_{ks}, which is directly related to the measurement quality, could be allowed to vary over survey mode by replacing the parameter λ_{ks} by λ_{ksx}, a possibility that can be tested using the data. Even without such an interaction, however, the classification probability in Equation 3 will differ over survey mode simply due to the nonlinear nature of the multinomial logistic regression.

Maximum-likelihood estimation of the model can proceed by directly maximizing the likelihood in Equation 2, or by expectation-maximization (EM), starting the E-step by choosing starting values for the model parameter estimates $\hat{\vartheta}_{t-1}$, and calculating, for each observation, the posterior joint probability of the 7×2 latent variable values given the observation,

$$\hat{P}(S, M|x, y_1, y_2, y_3) = \frac{\hat{P}(y_1, y_2, y_3|x, S, M)\hat{P}(S|x)\hat{P}(M|x)}{\sum_{s=1}^{7}\sum_{m=1}^{2}\hat{P}(y_1, y_2, y_3|x, S, M)\hat{P}(S|x)\hat{P}(M|x)}, \tag{4}$$

where $\hat{P}(\cdot) := P(\cdot|\hat{\vartheta}_{t-1})$. In the maximization step, new parameter estimates are then obtained by standard estimation techniques for multinomial logistic regression, by iterative proportional fitting, or by pseudo-ML, using the posteriors as weights,

$$\hat{\vartheta}_t = \arg\max_{\vartheta} \sum_{i=1}^{n}\sum_{s=1}^{7}\sum_{m=1}^{2}\hat{P}(S, M|x, y_1, y_2, y_3)\ln P(y_1, y_2, y_3|x, S, M, \vartheta). \tag{5}$$

The E-step in Equation 4 and M-step in Equation 5 are then iterated until convergence. In practice it is usually not necessary to create custom computer programs to obtain the estimates, since the model formulated above can be readily estimated in standard software for latent class analysis such as Mplus (Muthén and Muthén, 2012), ℓEM (Vermunt, 1997), or Latent GOLD (Vermunt and Magidson, 2013). The present analysis uses Latent GOLD 5.0; program code and data is provided in the appendix.

However, freely estimating the class sizes will often lead to class size estimates that differ from the official proportions of citizens living in each neighborhood, say, π_s. If this is not desired, a possible solution is to fix the class sizes to these known official proportions. However, this entails the complication that the meaning of the classes in terms of the observed indicators must also be restricted in the estimation procedure, to ensure, for instance, that a class chosen to represent the Zuid neighborhood has the correct fixed proportion. Moreover, due to the presence of covariates this restriction needs to be made in terms of the marginal $\sum_x \hat{P}(S|x)$. In short, while possible, imposing such restrictions can be challenging in practice.

A simpler solution to fix the marginal priors to known neighborhood proportions is to poststratify the posterior in Equation 4. After running the EM algorithm to convergence, the maximum-likelihood estimates of the (mis)classification rates $P(y_j|S = s)$ are simply the average estimated posterior probability of observing $S = s$ within each category of y_j,

$$\hat{P}(y_j|S = s) = \frac{1}{n\hat{P}(S = s)}\sum_{y_{j'\neq j}}\sum_{x}\hat{P}(S = s|x, y_1, y_2, y_3). \tag{6}$$

We then poststratify this estimate by replacing Equation 6 with

$$\hat{P}(y_j|S=s) = \frac{w_s}{\sum_s w_s} \sum_{y_{j'\neq j}} \sum_x \hat{P}(S=s|x, y_1, y_2, y_3), \tag{7}$$

where the poststratification weights are chosen to reflect the known proportions,

$$w_s = \pi_s / \hat{P}(S=s).$$

From Equation 4 it can be verified that this poststratification corresponds to one cycle of the EM algorithm, in which the marginal priors $P(S = s)$ are fixed to π_s, and the M-step is performed separately for each variable.

4 Results

4.1 Model fit

The previous section made certain choices for the specification of the model. A first step in analysis is to see whether these choices were necessary to fit the data, or whether a simpler model could have been fit. Conversely, some of the choices may have been too restrictive and it is of interest to see whether these restrictions fit the data. Table 3 investigates this by comparing different specifications using as fit measures the BIC, AIC, AIC3 and CAIC (see Vermunt and Magidson, 2013, p. 60). All of these information criteria seek a balance between model-data fit and model complexity, but do so in slightly different ways, leading to different preferred models shown in bold face in the table (lower values are better). The total "bivariate residual" (Vermunt and Magidson, 2013; Oberski et al., 2013b) is also given, in the last column (BVR): a sum of chi-squares measuring the distance between observed and expected frequencies in the bivariate cross-tables.

The rows in Table 3 correspond to models that result from combining the choices of survey mode effect with the number of latent method classes. The first model fitted (row 1) was the seven-class local independence model in Equation 1. None of the fit measures prefer this model, clearly indicating residual dependence. Model 1 as well as models 2–4 had one method class, which simply means there was no method factor (as can be verified from Equation 2). Models 8–10, on the other hand, were fit to investigate whether the previous Section's choice of two method classes should be extended to three classes. Comparing these three groups it is clear that all fit measures prefer a model with two method classes. Moreover, the total BVR is most reduced when going from no method factor at all to a method factor with two classes. There is indeed a dependence between the two survey measures that is adequately accounted for by allowing for two method factor classes.

Among the models that had two method classes, models 5–7 differ on how the survey mode effect was specified. Model 5 only related the survey mode to the true neighborhood S, but not directly to the survey measures (in Figure 3, no direct arrows from x to y_2 or y_3). Model 7 is the model shown in Figure 3, allowing for both direct and indirect effects. It is

Table 3: Model fit measures for different specifications of the latent class model. The best fit according to each measure has been indicated in **bold face**.

	Mode effect	Method classes	# par.	L^2	BIC	AIC	AIC3	CAIC	Total BVR
1.	No effect	1	153	1198	-10101	-1514	-2870	-11457	15.9
2.	Only $x \rightarrow S$	1	165	505	-10694	-2183	-3527	-12038	15.9
3.	$x \rightarrow S; x \rightarrow y_3$	1	181	443	-10624	-2213	-3541	-11952	13.5
4.	$x \rightarrow S; x \rightarrow (y_2, y_3)$	1	195	162	-10788	-2466	-3780	-12102	0.99
5.	Only $x \rightarrow S$	2	183	205	**-10844**	-2447	-3773	**-12170**	3.08
6.	$x \rightarrow S; x \rightarrow y_3$	2	199	133	-10783	**-2487**	**-3797**	-12093	0.39
7.	$x \rightarrow S; x \rightarrow (y_2, y_3)$	2	213	105	-10695	**-2487**	-3783	-11991	0.21
8.	Only $x \rightarrow S$	3	201	151	-10749	-2465	-3773	-12057	0.71
9.	$x \rightarrow S; x \rightarrow y_3$	3	217	100	-10666	-2484	-3776	-11958	0.38
10.	$x \rightarrow S; x \rightarrow (y_2, y_3)$	3	231	83	-10566	-2473	-3751	-11844	0.15

preferred by AIC. While Model 5 is preferred by the BIC and CAIC, the total BVR is still relatively large; the bootstrapped p-value for this measure of misfit is p = 0.010, indicating that this model may be too parsimonious. Since the misfit relative to model 7 can only result from the omission of the two direct arrows, the source of the misfit can be discovered by inspecting the individual BVR's for these cross-tables (between x and each of y_2 and y_3). Theses BVR's show that most of the misfit was caused by the direct question, y_3. A model only allowing a direct effect on y_3, model 6, was therefore also fitted to the data as a compromise between Models 5 and 7. This model is preferred by AIC3. The total BVR after freeing only this one direct effect rather than both has a bootstrapped p-value of 0.120. Model 6 therefore appears to fit well in both relative and absolute terms, and provides a reasonable balance between complexity and fit. This model was therefore selected.

4.2 Error rate estimates

From the final model we calculate the estimated overall (mis)classification rates $\hat{P}(y_j|S)$ by marginalizing over the mode and method. These (mis)classification rates, as calculated using Equation 7, are shown for the three measures of neighborhood of residence in Table 4. Given each "true neighborhood" (columns), the conditional probability of observing a particular category of the three measures (rows) is shown. For example, when answering the direct question, respondents who belong to the latent class labeled "Nieuw-West" (i.e., truly living in Nieuw-West) have an estimated probability of 0.1287 of answering mistakenly that they live in West.

Overall, as one, might expect, Table 4 shows the administrative register is estimated to be of the highest measurement quality overall. This can be seen by comparing the diagonal estimated correct classification probabilities between the three measures. The average misclassification weighted by class size, $\sum_s P(S = s)[1 - P(y_j = s|S = s)]$, was 0.5% for the administrative register, 7% for the postcode-derived survey measure, and 4.8% for

Table 4: Classification rates for the three observed measures (Model 6). Shown are the estimated conditional probabilities of observing each category given the latent class. The (**boldface**) diagonal in each table indicates the estimated proportion of correct classifications. Note that missings are considered incorrect classifications here.

	Latent class ("true neighborhood")						
	Centrum	West	Nw-West	Zuid	Oost	Noord	Zdoost
Class size	0.1063	0.0902	0.0972	0.4100	0.0990	0.1096	0.0876
Official stats.	0.1061	0.1738	0.1775	0.1720	0.1555	0.1108	0.1036
Register: postcode							
Centrum	**1.0000**	0.0000	0.0000	0.0000	0.0000	0.0022	0.0000
West	0.0000	**0.9947**	0.0000	0.0000	0.0050	0.0000	0.0028
Nieuw-West	0.0000	0.0000	**0.9921**	0.0000	0.0000	0.0044	0.0000
Zuid	0.0000	0.0000	0.0029	**0.9994**	0.0000	0.0000	0.0058
Oost	0.0000	0.0053	0.0025	0.0000	**0.9950**	0.0022	0.0000
Noord	0.0000	0.0000	0.0025	0.0006	0.0000	**0.9912**	0.0000
Zuidoost	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.9914**
Survey: postcode							
Missing	0.0656	0.0531	0.0722	0.0972	0.0584	0.0658	0.0634
Centrum	**0.9321**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
West	0.0023	**0.9416**	0.0000	0.0012	0.0000	0.0000	0.0000
Nieuw-West	0.0000	0.0000	**0.9278**	0.0000	0.0000	0.0000	0.0000
Zuid	0.0000	0.0053	0.0000	**0.9016**	0.0000	0.0000	0.0000
Oost	0.0000	0.0000	0.0000	0.0000	**0.9416**	0.0000	0.0000
Noord	0.0000	0.0000	0.0000	0.0000	0.0000	**0.9342**	0.0000
Zuidoost	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.9366**
Survey: direct question							
DK	0.0023	0.0051	0.0198	0.0065	0.0050	0.0241	0.0110
Centrum	**0.9774**	0.0027	0.0000	0.0047	0.0049	0.0000	0.0000
West	0.0113	**0.9600**	0.1287	0.0029	0.0049	0.0000	0.0000
Nieuw-West	0.0023	0.0295	**0.8490**	0.0011	0.0000	0.0000	0.0000
Zuid	0.0000	0.0027	0.0025	**0.9796**	0.0024	0.0000	0.0000
Oost	0.0068	0.0000	0.0000	0.0012	**0.9804**	0.0022	0.0000
Noord	0.0000	0.0000	0.0000	0.0006	0.0024	**0.9737**	0.0082
Zuidoost	0.0000	0.0000	0.0000	0.0035	0.0000	0.0000	**0.9808**

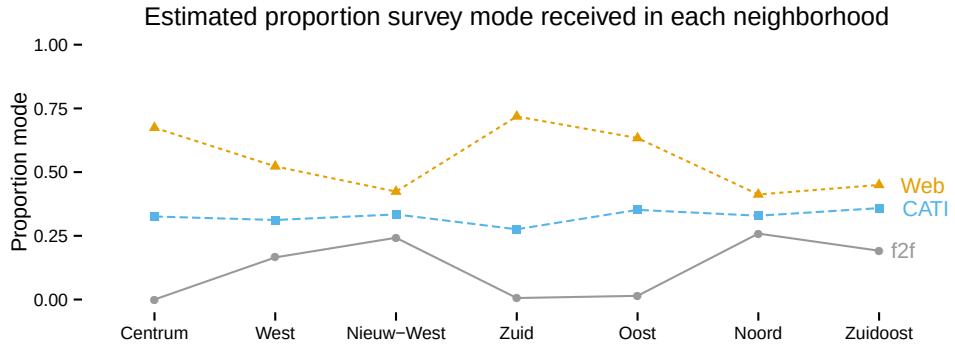

Figure 4: Model-based estimates of the proportion residents within each (true) neighborhood who received one of the survey modes (relationship between x and S).

the direct survey question. However, these misclassification rates include the "missing" category as "misclassification". If missingness were to be ignored, the misclassification rates for the postcode-derived and direct survey measure would be 0.15% and 3.8%, respectively. That is, ignoring the missing data, the postcode-derived measure is estimated to contain less measurement error than the administrative register. However, missingness in the postcode-derived survey measure does appear to be larger for the Zuid neighborhood (deviance 18.5 on 6 df, $p = 0.005$), so that it may not be ignorable for certain purposes.

Based on the model, it is possible to estimate the relationship between the survey mode x and the true neighborhood S. This relationship is shown as the estimated proportion of people within each true neighborhood who were interviewed in a particular mode in Figure 4. The "Web" mode, which includes a small percentage of completed paper-and-pencil questionnaires, is the most prominent in all neighborhoods, followed by CATI and face-to-face. Of course, this distribution reflects the sequential mixed-mode design employed by the municipality. Figure 4 primarily shows that this mix of modes differs strongly over neighborhoods. For example, in Centrum zero respondents were interviewed face-to-face, while in Nieuw-West and Noord about 25% of respondents were. Since measurement quality may differ over modes, the neighborhood misclassification differences shown in Table 4 may potentially be partially explained by mode mix differences over neighborhoods. However, the model allows us to disentangle these factors by separating the misclassification rates by mode.

The top rows of Table 4 also show that there is a large discrepancy between the estimated class sizes and known population proportions from official statistics. In particular, the affluent Zuid neighborhood is overrepresented, while the lower social-economic status-neighborhoods West, Nieuw-West, Oost, and Zuidoost are underrepresented. The most likely explanation for this phenomenon is differential nonresponse. Particularly, noncontacts are often higher in areas with low socio-economic status (Stoop, 2005), reducing the effectiveness of sequential mixed mode as a means of canceling out differential nonresponse. As discussed in the previous section we used poststratification to correct the model estimates in table 4, weighing the posteriors so that class sizes correspond to these known proportions.

Figure 5 shows how the quality of the two survey measures differs over the three survey

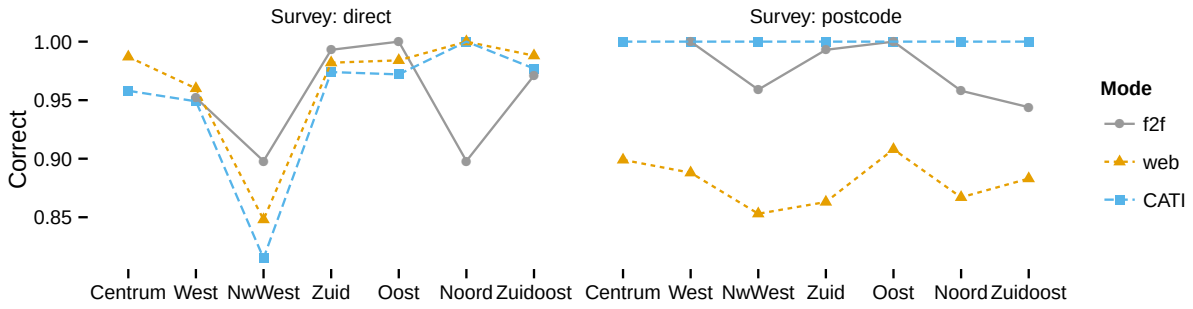

Figure 5: Estimated correct classification probabilities (excluding missing) by survey mode for the two survey measures.

modes. Plotted is the probability of observing the correct category for each of the two survey measures, among respondents who answered in one of three survey modes. For the direct question, Figure 5 demonstrates some differences between the modes; in particular, with the exception of one neighborhood, correct classifications are somewhat more prevalent in the face-to-face mode (triangles) than in the other two modes. The largest quality differences between the modes are observed in the postcode-derived measure, however. When interviewed over telephone using CATI (dots in the Figure), the measurement quality of the postcode-derived survey measure is estimated to be perfect. The reason for this is that over the phone, respondents were read their administrative register value for the postcode and asked whether this was correct or not; if not, the value was corrected in the register. This procedure was only officially followed over the telephone. However, an interviewer visiting a respondent's home in a different neighborhood than indicated by their postcode may also have provided some correction to the survey values, leading to a higher quality using face-to-face data collection (triangles). The web mode (squares) was estimated to yield the lowest quality here, with correct classification rates between 0.85 and 0.90. Overall, the conclusion of Table 4 that Nieuw-West yields the highest classification error remains even after conditioning on mode.

It should be noted at this point that the differences in measurement quality over survey modes need not be fully due to the mode of data collection only, since the choice of mode was not randomized, but determined by response status. These differences may reflect differences in mode, but also may be related to the type of respondent who is offered the corresponding mode.

The effect of the method classes is shown in Figure 6 as a marginal profile plot. Membership of class 2 appears to be most strongly related to providing a missing value on both survey measures. Thus, the method factor here predominantly represents a tendency to refuse to report one's neighborhood in a survey. While this tendency is strong within class 1 with a probability of 0.858 of omitting the postcode, only about 9% of respondents are estimated to belong to this class.

To further interpret the method classes, the probabilities in Figure 6 were converted to log-odds ratios by converting each line in the figure to odds, taking the ratio of the odds for Method class 1 to those for Method class 2, and then taking the log of that ratio. This yields a type of logistic regression coefficient that shows to what extent each category is overclaimed

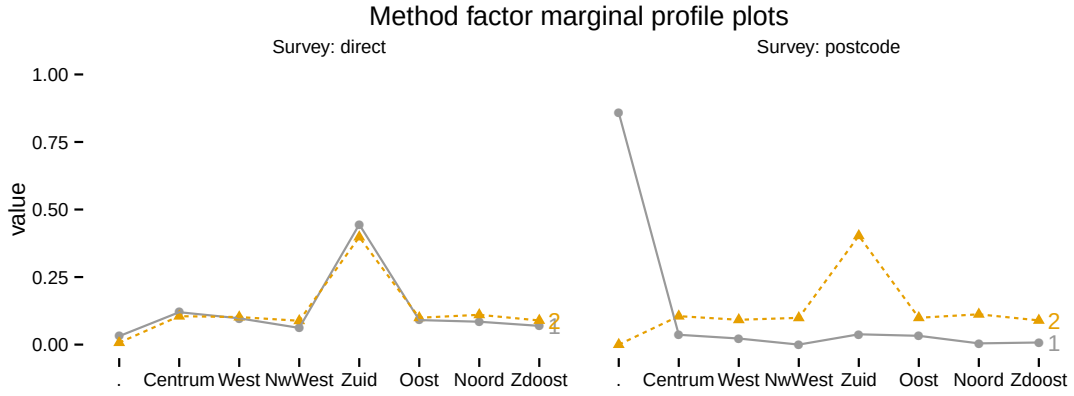

Figure 6: Estimated conditional probabilities of choosing each category within method classes (profile plot). Method class 1 is estimated to contain 9% and Method class 2 is estimated to contain 91% of respondents.

(positive log-odds) or underclaimed (negative log-odds) relative to true neighborhood of residence. Whenever the line for Method = 1 is above that for Method = 2 in Figure 6, there is overclaiming of this neighborhood and the log-odds will be positive. The resulting log-odds for the direct measure have been plotted against the neighborhoods' social-economic status (SES) scores[5] in Figure 7. It can be seen that those neighborhoods with a high SES score tend to be overclaimed, while those with a low SES score are underclaimed within Method class 1. Thus, the 9% of respondents who are estimated to belong to this class appear to show socially desirable tendencies in stating where they live.

5 Discussion and conclusion

This chapter demonstrated how measurement error rates can be estimated in survey and administrative measures of a categorical variable without the need for a golden standard. Using linked survey-register data from the municipality of Amsterdam on the neighborhood of residence, a multiple latent variable model estimated error rates, both in the administrative register and in the survey for different data collection modes. The register was indeed estimated to be of excellent quality with an average 0.5% error rate. In this case, therefore, often-made assumptions about this administrative register's quality do not appear to be far from the truth.

The average error rate estimated here (0.5%) was lower than that estimated from expensive audits performed by the local government, which had put the estimated average error rate at about 1.7%. On the one hand, thus, latent variable modeling may have somewhat overestimated the quality of the administrative register. On the other hand, the costs of full-scale audits compared with latent variable modeling are gigantic, meaning that a small inaccuracy still leaves latent variable modeling an attractive alternative. More studies comparing the results of such audits with those using the latent variable approach are needed, however. Par-

[5]These were calculated using the descriptives in Table 2 as $(\text{Income} + (100 - \text{Unemployment}) + \text{Education})/3$.

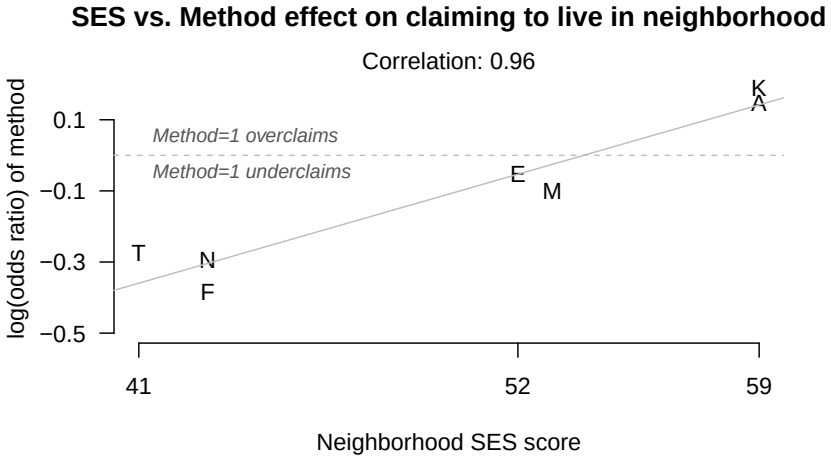

Figure 7: Scatterplot of a constructed neighborhood SES score (see Table 2) and the log-odds ratio (logistic regression coefficient) of choosing each neighborhood when the method class equals 1. A: Centrum, E: West, F: Nieuw-West, K: Zuid, M: Oost, N: Noord, T: Zuidoost.

ticularly relevant in the Dutch context is that one estimate is below the officially mandated 1% limit and the other above it.

The two survey measures, one direct and the other derived from a question asking the respondents' postcode, clearly showed a lower quality. When asked over the web, for instance, postcode-derived neighborhood of residence was estimated to have an error rate of between 10% and 15%. Such an error rate may be acceptably small for certain applications, while for others it may lead to rather large distortions in the final analyses of interest. For example, when the neighborhood of residence is used to link survey answers to background variables on the neighborhood level, measurement error in these linked background variables will result.

Some limitations of the present study remain. First, the administrative register in question is of course collected all over the Netherlands, not just in one municipality. This chapter's results can therefore not be generalized to the GBA register as a whole. Second, the survey mode was not randomized so that it is difficult to attribute quality differences over survey modes to either mode or selection effects. Third, the assumption that measurement error in the administrative register is conditionally independent of the survey answers may be problematic in some cases – especially when there are strong personal incentives to provide incorrect values that are consistent with the register, for example when renting illegally. This possibly explains the higher estimate of quality obtained here. Fourth and finally, because only one register measure was available it was not possible here to incorporate a method factor for the register as well as the survey measures. Future research may also include additional covariates beyond survey mode.

In summary, latent variable modeling of linked survey-register data can be viewed as an attractive method of investigating the assumption that administrative registers are perfect.

Moreover, it is the only way of estimating measurement error in "objective" survey questions without requiring a gold standard. Since objective variables such as the neighborhood of residence play a central part in many analyses, this approach appears to be a fruitful way of preventing distortions in such analyses.

References

Bakker, B. F. (2009). *Trek alle registers open! [Open up the registers!]*. VU University, Amsterdam.

Bakker, B. F. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, 66(1):8–17.

Biemer, P. (2011). *Latent Class Analysis of Survey Error*. Wiley, New York.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement error in nonlinear models: a modern perspective*, volume 105. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Boca Raton, FL.

Entwisle, B. and Elias, P. (2013). *New Data for Understanding the Human Condition: International Perspectives*. OECD, Paris, France.

Fuller, W. (1987). *Measurement error models*. John Wiley & Sons, New York.

Gomez, S. L. and Glaser, S. L. (2006). Misclassification of race/ethnicity in a population-based cancer registry (united states). *Cancer Causes & Control*, 17(6):771–781.

Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics (JOS)*, 28(2).

Hagenaars, J. A. P. and McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge University Press, Cambridge, UK.

Kreuter, F., Müller, G., and Trappmann, M. (2010). Nonresponse and measurement error in employment research nonresponse and measurement error in employment research: Making use of administrative data. *Public Opinion Quarterly*, 75(5):pp. 880–906.

Lazarsfeld, P. (1950). The logical and mathematical foundation of latent structure analysis (ch. 10) and the interpretation and mathematical foundation of latent structure analysis (ch. 11). In Stouffer, S., Guttman, L., Suchman, E., Lazarsfeld, P., Star, S., and Clausen, J., editors, *Measurement and Prediction*, pages 362–472. Princeton University Press.

Maudsley, G. and Williams, E. (1996). Inaccuracy in death certification–where are we now? *Journal of Public Health*, 18(1):59–66.

McCutcheon, A. (1987). *Latent class analysis*. Sage Publications, Inc.

Muthén, L. K. and Muthén, B. (2012). *Mplus User's Guide, Seventh Edition.* Muthén & Muthén, Los Angeles, CA.

Oberski, D. (2013). The latent class MTMM model. *Psychological Methods*.

Oberski, D., Kirchner, A., Eckman, S., and Kreuter, F. (2013a). Evaluating the quality of survey and administrative data with multitrait multimethod models. In *Registerdata meeting*. Tilbrug University.

Oberski, D., Van Kollenburg, G., and Vermunt, J. (2013b). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, 7(3).

Pavlopoulos, D. and Vermunt, J. (2013). Measuring temporary employment. do survey or register data tell the truth? *Survey Methodology*.

Scholtus, S. and Bakker, B. F. (2013). *Estimating the validity of administrative and survey variables through structural equation modeling: a simulation study on robustness*. CBS Discussion Papers. Statistics Netherlands, The Hague.

Stoop, I. A. (2005). *The hunt for the last respondent: Nonresponse in sample surveys*. Sociaal en Cultureel Planbureau, Voorburg.

Vermunt, J. K. (1997). Lem 1.0: A general program for the analysis of categorical data. *Tilburg: Tilburg University*.

Vermunt, J. K. and Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Statistical Innovations Inc., Belmont, MA.

Wallgren, A. and Wallgren, B. (2007). *Register-based statistics: administrative data for statistical purposes*. Wiley, New York.

West, B. T. and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74(5):1004–1026.

A Program input and data

Listing 1: Latent GOLD syntax for Model 6.

```
options
  algorithm tolerance=1e-007 emtolerance=0.01 emiterations=550
     nriterations=70;
  startvalues  seed=0 sets=50 tolerance=1e-005 iterations=70;
  bayes   categorical=0 variances=0 latent=0 poisson=0;
  missing includeall;
  output
    parameters=effect standarderrors profile probmeans=posterior
    bivariateresiduals estimatedvalues=model;

variables
  dependent stadsdeelDirectSurvey, stadsdeelPCSurvey, stadsdeelRegister;
  independent surveyModeJoined nominal;
  latent   Cluster nominal 7;

equations
  Cluster <- 1|surveyModeJoined ;
  stadsdeelDirectSurvey <- 1|surveyModeJoined + Cluster;
  stadsdeelPCSurvey <- 1 + Cluster;
  stadsdeelRegister <- 1 + Cluster;
```

Complete data (without missing)

#	su_di	su_pc	reg	mode	freq
1	A	A	A	cati	138
2	A	A	A	web	265
3	A	E	A	web	1
4	A	E	E	web	1
5	A	K	K	cati	3
6	A	K	K	web	3
7	A	M	M	web	1
8	E	A	A	cati	4
9	E	A	A	web	1
10	E	E	E	cati	109
11	E	E	E	f2f	59
12	E	E	E	web	168
13	E	E	M	cati	2
14	E	F	F	cati	24
15	E	F	F	f2f	1
16	E	F	F	web	19
17	E	F	M	f2f	1
18	E	K	E	web	2
19	E	K	K	cati	3
20	E	K	K	web	1
21	E	M	M	cati	2
22	F	A	A	cati	1
23	F	E	E	cati	5
24	F	E	E	f2f	2
25	F	E	E	web	4
26	F	F	F	cati	108
27	F	F	F	f2f	84
28	F	F	F	web	127
29	F	F	K	cati	1
30	F	F	N	cati	1
31	F	K	K	cati	1
32	K	E	E	cati	1
33	K	E	K	web	2
34	K	F	F	f2f	1
35	K	K	K	cati	458
36	K	K	K	f2f	10
37	K	K	K	web	1044
38	K	K	N	web	1
39	K	M	M	cati	1
40	M	A	A	web	2
41	M	K	K	cati	1
42	M	K	K	web	1
43	M	M	E	web	2
44	M	M	M	cati	141
45	M	M	M	f2f	6
46	M	M	M	web	234
47	M	N	N	f2f	1
48	N	K	K	cati	1
49	N	M	M	cati	1
50	N	N	F	f2f	1
51	N	N	F	web	1
52	N	N	M	web	1
53	N	N	N	cati	150
54	N	N	N	f2f	101
55	N	N	N	web	161
56	N	T	T	cati	1
57	N	T	T	f2f	1
58	T	K	K	web	4
59	T	T	E	f2f	1
60	T	T	K	f2f	1
61	T	T	T	cati	128
62	T	T	T	f2f	62
63	T	T	T	web	145
64	W	E	E	f2f	1
65	W	E	E	web	1