Statistical Modelling in Surveys without Neglecting *The Undecided*: Multinomial Logistic Regression Models and Imprecise Classification Trees under Ontic Data Imprecision

Julia Plass
Department of Statistics, LMU Munich
julia.plass@stat.uni-muenchen.de

Paul Fink
Department of Statistics, LMU Munich
paul.fink@stat.uni-muenchen.de

Norbert Schöning
Geschwister Scholl Institute of
Political Science, LMU Munich
norbert.schoening@gsi.uni-muenchen.de

Thomas Augustin
Department of Statistics, LMU Munich
augustin@stat.uni-muenchen.de

Abstract

In surveys, and most notably in election polls, undecided participants frequently constitute subgroups of their own with specific individual characteristics. While traditional survey methods and corresponding statistical models are inherently damned to neglect this valuable information, an ontic random set view provides us with the full power of the whole statistical modelling framework. We elaborate this idea for a multinomial logistic regression model (which can be derived as a discrete choice model for voting behaviour) and an imprecise classification tree, and apply them as a prototypic illustration to the German Longitudinal Election Study 2013. Our results corroborate the importance of a sophisticated, random set-based modelling. Furthermore, by reinterpreting the undecided respondents' answers as disjunctive random sets, general forecasts based on interval-valued point estimators are calculated.

Keywords. Ontic data imprecision, survey methodology, election polls, multinomial logistic models, discrete choice models, imprecise classification trees, conjunctive random sets, disjunctive random sets, epistemic prediction, German Longitudinal Election Study 2013 (GLES 2013)

1 Introduction

Although pondering between several options is characteristic for human beings, indecisiveness of respondents is not reflected in most surveys. Instead it is common to force a precise answer, and at best to provide an additional category "Don't know" for those that are not decided. Frequently, in the framework of the analysis respondents reporting this "Don't know" category are no longer taken into consideration as those answers are understood as unusable. In many cases indecisive respondents are able to definitely exclude some options, which is not expressed by category "Don't know", and additionally characteristics of indecisive and decisive respondents may systematically differ. Consequently, the common proceeding leads to a substantial loss of information in data collection and biased results in the analysis of data.

In order to deal with this problem, it is necessary that questionnaire designers allow for multiple answers as "option A or option B" or at least provide ways to construct them. Hence, the preferences of the indecisive respondents are reflected in the most informative way and we are able to distinguish between different types of indecisive respondents. In this sense, we explicitly account for the heterogeneity within the group of indecisive respondents.

In order to embed this idea into a proper statistical modelling framework, we mainly will rely on the notion of *ontic sets* in the sense of Dubois and Prade ([15, 16]) as well as Dubois and Couso ([11]). They stressed the importance of differentiating between two views of a set, one representing precise collections of elements (*ontic view*) and the other reflecting incomplete knowledge about a particular precise value (*epistemic view*) ([12]). As answers of indecisive respondents are interpreted as ontic sets, we will call data that are coarse induced by indecision like "A or B" *data under ontic imprecision*.

Our paper is structured as follows. In Section 2 we will recapitulate some notions mainly based on random set theory ([19]) that have already been investigated in the framework of ontic sets ([11, 12]). In this context, we will emphasize the applicability of ontic sets to the general analysis in the presence of answers of indecisive respondents, where the focus will be on incorporating the idea of the ontic view into multinomial logistic regression analysis and classification trees in order to

model heterogeneity of respondents by their covariates. By briefly digressing into the epistemic view, in Section 3 interval-valued forecasts will be constructed. The aforementioned techniques are used in an illustrative analysis based on the German Longitudinal Election Study that is briefly presented in Section 4. Corresponding results are shown and compared to those obtained from classical statistical analyses in Section 5.

For sake of simplicity, we focus on categorical data of nominal scale, yet adaptation to ordinal scale for other applications may be derived only with little additional effort. Moreover, an extension to coarse categorical covariates under ontic data imprecision may be achieved with similar arguments.

2 Data under Ontic Imprecision: Basic Idea and Extending some Statistical Approaches

As argued in the introduction, it is crucial to distinguish between the ontic and epistemic view and thus between *random conjunctive sets* and *ill-known random variables* ([11, 12]). In this section we focus on *random conjunctive sets*, underlying the ontic view.

2.1 General Analysis

As we regard the case of categorical data with a finite state space, it is sufficient to focus on the definition of *finite random sets*, which can be considered as a simplification of the more general definition of random closed sets. A finite random set is a mapping $Z^* : \Omega \to \mathcal{P}(S)$ such that for any $A \subseteq S$ holds: $Z^{*-1}(\{A\}) = \{\omega \in \Omega : Z^*(\omega) = A\} \in \mathcal{A}$, where S denotes the state space, $\mathcal{P}$ the power set and $(\Omega, \mathcal{A})$ the underlying measurable space, equipped later with a probability measure P (e.g. [20]). In other words, a finite random set is characterized by a measurable mapping on the power set. Couso and Dubois call this notion *random conjunctive set* or *(ontic) set* ([11, 12]).

The important characteristic of an ontic set is that it represents a precise collection of elements in the sense that there is no true element of S underlying, but the set itself constitutes an entity of its own ([11]). Answers like "A or B" may be regarded as an ontic set $\{A,\ B\}$ as there is no unique preference. Therefore, the nature of coarse data under ontic imprecision is well represented by the ontic view. Consequently, this leads to a power set based view, meaning an extension of the classical precise state space S to $S^* = \mathcal{P}(S) \setminus \emptyset$, with the asterisk stressing ontic imprecision. Thus, basing the analysis on S^*, and therefore regarding coarse categories as own entities, provides the main idea of dealing with ontic imprecision. The one and only difference compared to the classical case is the adapted state space S^*.

Hence, by reinterpreting the random conjunctive set as precise random variable, classical probability theory and all statistical methods based on it are applicable. In other words, the idea of the adapted state space is independent of the statistical method and exploiting this idea further for formulating regression models and classification trees in the next sections should be regarded as an example.

A short example shall be given already here. It consists of calculating the probability of respondents, who are at least indecisive between particular options C_0, by the probability of the family of corresponding supersets $\mathcal{C} = \{T \subseteq S : C_0 \subseteq T\}$ to

$$P_{Z^*}(\mathcal{C}) = \sum_{C \in \mathcal{C}} P_{Z^*}(C)\ , \tag{1}$$

which is essentially a summation over singletons of the space S^* (cf. [11, p. 8]).

2.2 Regression Analysis

Generally, the main goal of regression analysis consists of modelling the relation between several covariates X and a dependent variable Y, without claiming to describe necessarily the causal impact of variables. In our case the dependent variable is assumed to be coarse under ontic imprecision, whereas we address precise covariates. As we restrict ourselves to a coarse categorical variable of nominal scale, a multinomial logit model is an appropriate statistical model.

2.2.1 Multinomial Logit Model

In this section it is mainly referred to [17, pp. 329-331]. A more thorough treatment of discrete choice models can be found for instance in [29]. We denote by $Y_i \in S = \{1, \ldots, c\}$ the random variable describing the response of individual $i = 1, \ldots, n$. Assuming a multinomial logit model, the probability of occurrence of category $s \in \{1, \ldots, c-1\}$ for i with given covariate values $\mathbf{x}_i$ is set to be

$$P(Y_i = s \,|\, \mathbf{x}_i) \ = \ \pi_{is} \ = \ \frac{\exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s)}{1 + \sum_{r=1}^{c-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r)}\ , \tag{2}$$

with $\tilde{\mathbf{x}}_i^T = (1, \mathbf{x}_i^T)$ and category specific regression coefficients $\boldsymbol{\beta}_s = (\beta_{s0}, \beta_{s1}, \ldots, \beta_{sp})^T$ referring to p covariates. Because of the redundancy resulting from the fact that all probabilities add up to one, the corresponding probability for the so-called reference category c can

be determined by

$$P(Y_i = c \,|\, \mathbf{x}_i) \;=\; \pi_{ic} \;=\; 1 - \pi_{i1} - \ldots - \pi_{ic-1} \\ =\; \left(1 + \textstyle\sum_{r=1}^{c-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r)\right)^{-1} .$$

This corresponds to the side constraint that the regression coefficients of category c are set to zero.[1]

Expressing Equation (2) in terms of the linear predictor $\eta_{is} = \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s$, one obtains the logarithmised chances and the relative risks of category $s \in \{1, \ldots, c-1\}$ and reference category c by

$$\log\left(\frac{\pi_{is}}{\pi_{ic}}\right) = \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s \quad \text{and} \quad \frac{\pi_{is}}{\pi_{ic}} = \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s) \; . \quad (3)$$

Accordingly, the exponential of β_{sj} $(j = 1, \ldots, p)$ expresses how the chance for category s compared to the reference category c changes if the value of a certain covariate x_j is increased by one unit in the case of metric covariates or if x_j is taken instead of reference category x_J in the case of categorical covariates.

2.2.2 A Multinomial Logit Model Based Approach under Ontic Imprecision

The redefinition of the original precise state space $S = \{1, \ldots, c\}$ of Y to the state space $S^* = \mathcal{P}(S) \setminus \emptyset$ of Y^* is crucial for adapting the multinomial logit model to account for ontic imprecision, treating answers of indecisive respondents as own categories, as already pointed out in Section 2.1.

Consequently, the number of categories of the dependent variable Y^* amounts to the cardinality of the new state space S^* $(m = |S^*| = |\mathcal{P}(S) \setminus \emptyset| = 2^{|S|} - 1)$. It formalizes the idea that no longer for each $Y \in \{1, \ldots, c\}$ but for each $Y_i^* \subseteq \{1, \ldots, c\}$ probabilities $\pi_{i1}^*, \ldots, \pi_{im}^*$ are modeled and coefficients $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_{m-1}^*$ are estimated. Hence, the probability of occurrence of category $s \in \{1, \ldots, m-1\}$ for i with given covariate values $\mathbf{x}_i$ is determined by

$$P^*(Y_i^* = s \,|\, \mathbf{x}_i) \;=\; \pi_{is}^* \;=\; \frac{\exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_s^*)}{1 + \sum_{r=1}^{m-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r^*)}$$

and for reference category m by

$$P^*(Y_i^* = m \,|\, \mathbf{x}_i) \;=\; \pi_{im}^* \;=\; 1 - \pi_{i1}^* - \ldots - \pi_{im-1}^* \\ =\; \left(1 + \textstyle\sum_{r=1}^{m-1} \exp(\tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_r^*)\right)^{-1} .$$

[1] In order to ensure identifiability it is important to include a side constraint for the regression coefficients into the basic model. Alternatively, any other category may be chosen as reference category or a symmetric type of constraint like $\sum_{r=1}^{c} \beta_r^T = (0, \ldots, 0)^T$ can be applied (e.g. [30]).

In this way, one obtains own regression coefficients for each coarse category, which exactly reflects the underlying idea that different types of indecisive respondents are regarded as own group.

In summary, one can account for ontic imprecision within categorical variable Y of nominal scale by incorporating coarse answers as own categories into a multinomial logit model. Apart from the up to exponential increase in the number of categories nothing changes: All statistical methods refining and extending the classical multinomial logit model, like penalization approaches, flexible covariate modelling or random effects under repeated measurements (e.g. [30]), and their fundamental statistical properties, like consistency and asymptotic normality of estimators, can be transferred. In this way, the here addressed adaptation of the multinomial logit model serves as an example for incorporating the power set based idea into categorical regression models.

2.3 Classification Trees

Whereas in regression we are mainly interested in the estimation of the regression coefficients, which provide a structural interpretation of the data, in the framework of classification trees one major goal is to predict the value(s) of a dependent variable (called *class variable* Y later on) of a future observation, based on values of some independent, so-called feature, variables. Learning a classification tree involves recursively partitioning the full data space as it is available in the beginning, into disjoint subspaces by splitting with respect to some (in-)homogeneity criterion. A most favourable property of a single classification tree from a statistical modelling point of view is that it still allows a structural interpretation, while such is lacking in the even more prediction orientated ensemble of trees, so-called bags or forests.

In the framework of classification trees there are numerous algorithms available that are able to deal both with nominal and numerical variables, some even account for missingness at random, for instance Quinlan's ID3 [23] and Breiman's CART [9] and their successors. They share the concept of selecting splitting feature variables performing the partitioning by a similarity measure, in our context the entropy. For sake of simplicity we confine ourselves to class and feature variables of nominal scale.

In order to calculate the entropy and decide on a splitting feature variable, it is required to estimate the class' probabilities, classically achieved by the corresponding relative frequency. Abellan and Moral [4] introduced *imprecise classification trees* by changing the estimation to involve imprecise probability mod-

els. As a split criterion they favoured a maximum entropy approach and presented in [4] an adaptation of Quinlan's ID3 algorithm, both of which for sake of simplicity we employ.

Yet there are more general approaches, where for instance the full entropy range is taken into account, as in [18] or [13], the latter naturally growing a forest. Further improvements of the initial imprecise algorithm also include the concept of bagging [2, 3].

In our analyses in Section 5.3 we grow classification trees accordingly to [4] but relying on a Nonparametric Predictive Inference (NPI) model for estimation of the class probability distribution within a node instead, yet an Imprecise Dirichlet Model would have been also applicable; see [10] for a more detailed introduction to NPI for categorical data and [4] or [18] for a description on how an imprecise classification tree based on it is actually constructed. Yet, we briefly recall the estimation with NPI within a tree's node.

Each node of the tree consists of a collection of observations. They are assigned to nodes in such a way that they form the aforementioned disjoint subspaces in an optimal way with respect to the splitting criterion. In the context of an entropy based splitting criterion the probability distribution of the class variable is required. In [4] the assumption of a precise probability distribution is relaxed to a credal set leading to a maximum entropy split criterion approach. According to NPI the predictive probability that for a virtual next observation the class variable attains a value y_i of its state space is within the following interval

$$P(Y = y_i) \in \left[\max\left(0, \frac{n_i - 1}{n}\right), \min\left(\frac{n_i + 1}{n}, 1\right) \right], \quad (4)$$

with n_i the number of observations having a class value of y_i and n the overall number of observations, both with respect to the node under consideration.

In the situation where the class variable is only observable under ontic imprecision, we embed ontic sets into the framework of classification trees properly by a redefinition of the class variable as a finite random set, thus basing the analysis on the power set of the class variable space, similarly to the regression analysis. This is a direct implementation of the crucial idea, allowing us to reinterpret the ontic sets as a new precise class variable, i.e. an answer "A or B" is interpreted now as the precise class "AB". Therefore, any classification tree technique might be applied that is able to deal with a precise classification variable, regardless of the underlying probability model(s). This power set based technique is frequently applied in the framework of multi-label classification (e.g. MODEL–n in [8]). Due to the increased number of classes the concept of entropy correction ([27]) becomes more important, besides substituting Y by Y^* in (4).

Furthermore, basically any classification technique may be applied, after the state space of the variables under ontic uncertainty is substituted by its power set. The classification trees serve as a feasible example.

3 Interval-valued Forecast

We consider the same data situation, but change our perspective and the aim of our analyses. Instead of modelling the underlying structure of voting (in)decisions, we now turn to forecasts based on an epistemic reinterpretation of our data.

Let's assume that our main interest lies now in forecasting certain events by enforcing a final decision expressed by a variable Y_{final}. In the context of voting behaviour such a situation arises when a forecast on the election result is required. Under the assumption that the final decision is precise and consistent with the data collected now, this means a precise true value is underlying the set-valued response.

In this way, set-valued elements A^* of S^* are no longer interpreted as own entities, but are regarded as incomplete knowledge, which for every event B from the space $(S, \mathcal{P}(S))$ is given by (cf. [7, p.185])

$$P(Y_{\text{final}} \in B \,|\, Y^* = A^*) \in \begin{cases} \{0\}, & \text{if } B \cap A^* = \emptyset \\ \{1\}, & \text{if } B \supseteq A^* \} \\ [0,1], & \text{otherwise} \end{cases},$$

postulating that the final answer is compatible with the initial information from the ontic view.

This corresponds to an epistemic view of modelling[2]. However, models should be cautiously interpreted as the data were originally obtained under ontic imprecision, yet it may be justified for modelling purpose.

In the context of the epistemic view Couso and Dubois ([11]) consider *ill-known random variables* Y_{epist} with precise, but incomplete realizations y_{epist}. An *ill-known random variable* Y_{epist} is a multiple-valued mapping $Y_{\text{epist}} : \Omega \to \mathcal{P}(S)$ described by the disjunctive set of mappings

$$\left\{ Y_{\text{precise}} : \; Y_{\text{precise}}(\omega) \in Y_{\text{epist}}(\omega) \;, \forall\, \omega \in \Omega \right\},$$

where $Y_{\text{precise}} : \Omega \to S$ is a precise random variable. Thus, Y_{epist} is interpreted as the collection of several precise models that can be deduced from incomplete knowledge.

[2] First steps towards statistical modelling under epistemic data imprecision can be found in ([21]).

Taking the reinterpretation as disjunctive sets seriously, the range covering the true probability of a certain event of interest E can be expressed by Dempster's lower and upper probabilities ([14]) that are

$$\underline{P}_{Y_{\text{epist}}}(E) = \sum_{Y_{\text{epist}}(\omega)\subseteq E} p(\omega) ,$$
$$\overline{P}_{Y_{\text{epist}}}(E) = \sum_{Y_{\text{epist}}(\omega)\cap E\neq\emptyset} p(\omega) ,$$

where p is the probability mass function of P ([11]).

Thus, the proportion of an option E can be forecasted by the sample counterparts $\widehat{I}(E)$ of the interval

$$I(E) = \left[\underline{P}_{Y_{\text{epist}}}(E) , \ \overline{P}_{Y_{\text{epist}}}(E) \right] . \qquad (5)$$

As the difference between the values of the lower and the upper probability represents the lack of knowledge induced by indecisive answers, it is apparent that the length of this interval can be interpreted as the extent of the underlying epistemic imprecision.

In order to account additionally for statistical uncertainty due to finite sampling, confidence intervals for $I(E)$ may be calculated. This leads to so-called uncertainty regions aiming to cover both: imprecision due to incompleteness and statistical uncertainty ([31]).

4 Data

Until now the German Longitudinal Election Study (GLES) ([25]) is the most elaborated German electoral poll and currently focuses on three federal elections (2009, 2013, 2017). The sampling method of the initial data set of the *GLES* 2013 is a (3-step) random sample, which is treated here in our illustrative analysis as a simple random sample. As voting intentions before the election day are of main interest, we consider the preliminary study of *GLES 2013*, which is a face-to-face interview two months prior to the election day.

To our present knowledge there is not any pre-election study allowing indecisive respondents to express their voting intention by multiple answers. The main advantage of *GLES 2013* is that respondents are also explicitly required to report their voting intention's certainty ("certainty")[3] along with the assessments of several parties (*q21a-q21h*[4]). Those and the respondent's current voting intention[5], collected in a precise answer, allow us the construction of a variable "ontic", reflecting the respondent's indecision by multiple answers. The procedure for our construction of the variable "ontic" is as follows: While for all "very certain" respondents the reported party of the variable "vote" is taken, the party or parties with maximal assessment are chosen for the respondents that are "fairly certain" explicitly allowing by construction indecision between the corresponding parties. For the respondents that decide for "neither/nor" or "not certain at all" parties with maximal and second highest assessments are taken.The chosen way of construction of the variable "ontic" is to some extent arbitrary, but at least it accounts reasonably for ontic imprecision. In the following we focus on the second vote, as similar steps and explanations hold for the first vote as well.

	case 13	case 126	case 1515
certainty	very certain	fairly certain	neither/ nor
vote	GREEN	SPD	CD
assessCD	-1	-1	+3
assessSPD	+2	+1	+3
assessFDP	-4	0	0
assessLEFT	-4	+1	-5
assessGREEN	+4	-3	+2
	⇓	⇓	⇓
ontic	GREEN	LEFT:SPD	CD:GREEN:SPD

Table 1: Construction of variable "ontic" (example)

The examples in Table 1 illustrate the way of construction by means of three randomly chosen respondents.[6]As our goal consists of demonstrating the difference in results from an analysis including ontic imprecision and a classical analysis, such a constructed variable is required.

Partly due to the construction of variable "ontic" several respondents had to be excluded[7]. All conducted filtering steps (e.g. excluding voters of smaller parties or non-voters) that reduced the sample of initially 2003 to 1196 respondents can be found in [22]. The associated loss of information caused by the reduced

[3] *q13* with categories "very certain", "fairly certain", "neither/nor" and "not certain at all"

[4] Each measured on a scale from "-5" ("a very negative view of this political party") to "+5" ("a very positive view of this political party")

[5] The German election system mixes elements of election by proportionality and by majority. The voters have two votes (*q11ab*: second vote, *q11aa*: first vote). The second vote is generally considered as more important, because the proportion of seats in the German Bundestag mainly is allocated according to the second vote. The first vote determines the direct representative of an election district in the Bundestag.

[6] Translations of German abbreviations of political parties are used here. Considered parties are: *Christlich Demokratische Union Deutschlands* (CDU) and *Christlich-Soziale Union in Bayern* (CSU) representing throughout Germany one option only (here denoted by CD), *Sozialdemokratische Partei Deutschlands* (SPD), *Die Linke* (LEFT), *Bündnis 90/Die Grünen* (GREEN), *Freie Demokratische Partei* (FDP).

[7] In voting studies sample loss is rather common. Usually empirical analyses are reduced to those parties, who entered the German Bundestag finally (e.g. [28]).

CD	SPD	GREEN
495	271	125
LEFT	FDP	GREEN:SPD
106	39	36
CD:SPD	CD:FDP	GREEN:LEFT
35	18	15
LEFT:SPD	CD:GREEN:SPD	GREEN:LEFT:SPD
14	17	13
CD:FDP:SPD		
12		

Table 2: Absolute frequencies of constructed variable "ontic" (second vote)

sample size is undesirable, but unavoidable for an ontic analysis illustrated by this data set. Because of the underrepresentation of indecisive persons induced by the current design of the questionnaire, which implicitly excludes indecisive respondents by the preceding filtering of the "certainty" item (cf. [22]), we expect less marked differences between an ontic and a classical analysis, described in the following sections.

The resulting illustrative data set containing variable "ontic", whose absolute frequencies are given in Table 2, forms the basis of the following analysis.[8]

5 Data Analysis

The principal goal consists of comparing the results obtained by an analysis using the constructed variable "ontic" (cf. Section 4 and [22]) to a classical analysis excluding all uncertain respondents. This issue will be considered in this section with regard to the findings from Section 2. Hereby, we focus on the second vote, only where mentioned explicitly the first vote is considered. All analyses are based on complete cases, dependent on the variables effectively under consideration. We performed our analyses with the open-source statistical software R [24]. The code is available on request from the authors.

5.1 General Analysis

The analysis incorporating ontic imprecision is based on $S^* = \mathcal{P}(S) \setminus \emptyset$, where

$$S = \{\text{CD}, \text{SPD}, \text{GREEN}, \text{LEFT}, \text{FDP}\}$$

is the state space. Since only 13 elements of S^* are attained in the addressed data set, we adapted S^* to cover those values of variable "ontic" only (see Table 2).

If for instance the probability of respondents is of interest that are (at least) indecisive between party "SPD" and "GREEN", according to Equation (1) all probabilities referring to respondents that are (at least) indecisive between both parties have to be summed up, which can be estimated by associated relative frequencies to

$$\begin{aligned}
&\widehat{P}_{Z^*}\big(Z^* \supseteq \{\text{GREEN, SPD}\}\big) \\
&= \widehat{P}\Big(\big\{\omega :\, Z^*(\omega) = \{\text{GREEN, SPD}\}\big\}\Big) \\
&\quad + \widehat{P}\Big(\big\{\omega :\, Z^*(\omega) = \{\text{CD, GREEN, SPD}\}\big\}\Big) \\
&\quad + \widehat{P}\Big(\big\{\omega :\, Z^*(\omega) = \{\text{GREEN, LEFT, SPD}\}\big\}\Big) \\
&= \frac{36}{1196} + \frac{17}{1196} + \frac{13}{1196} \;\approx\; 0.06\,.
\end{aligned}$$

The estimated proportion of indecisive respondents is 0.13, calculated analogously. Consequently, if just decisive respondents are considered an amount of 13% of respondents are not taken into account. As respondents are excluded because of the value of the variable of interest itself, we are concerned with a *not missing at random* situation and thus ignoring the indecisive respondents may lead to biased results. This is particularly fatal for a theoretical understanding of voting decisions as well as from a practical campaigners' view, because this percentage covers those respondents that are of particular interest.

5.2 Regression Analysis

In order to analyse the heterogeneity within the coarse dependent variable Y under ontic data imprecision, the models presented in Section 2.2 are applied. The multinomial logit model has a longstanding tradition in the context of modelling voting behaviour[9].

In our analysis the variable "ontic" represents the coarse dependent variable, where "SPD" is chosen as reference category. Generally, it is important to choose all reference categories in such a way that interpretations enable answering the question of interest. For our illustrative purpose we use a very simple voting model with only two covariates[10], namely socio-demographical variable "religious denomination" (*q228*) as well as variable "most important source of information" (*q97*). In both variables certain categories were aggregated. Thus, variable "religious denomination" here only takes values "Christian" and "non-Christian", where the categories of "most important

[8] Absolute frequencies of singletons differ from those of variable "vote" due to the construction of variable "ontic".

[9] Actually, the multinomial logit model is the simplest model of the discrete choice family. Although it has several disadvantages for the modelling of voting behaviour as discussed by [6], for the sake of our illustrative application yet the multinomial logit model is appropriate, because it shows the basic concept in handling data under ontic imprecision, which can be extended analogously to more tailored models.

[10] Recent models of voting behaviour use policy distance, party identification and socio-demographical variables and yield a remarkable fit and prognostic validity (cf. [5])

Coefficient	ontic		classical
	CD	G:S	CD
intercept	0.37	−1.47***	0.13
rel.christ	0.32*	−0.05	0.49***
info.tv	0.01	−0.29	−0.01
info.np	−0.05	−1.67**	−0.01

Table 3: Comparison of results (second vote).

source of information" are translated to "television", "newspaper" and "other source", the latter also covering "radio", "internet" and "talking to other people". Every reclassification is subject to avoid categories with only few observations in order to decrease statistical uncertainty. By including "most important source of information" as a covariate into the model, we assume that the way how voters inform themselves of the federal election influences their voting intention. Nevertheless, one cannot exclude an opposite (causal) direction as respondents who vote for particular parties potentially avoid or prefer certain information sources because of the way this party is represented in it. This needs to be kept in mind when interpreting the model's results.

For reasons of conciseness estimated regression coefficients are shown just for category "CD" and "GREEN:SPD" (G:S) here.[11] With $n_{\text{CD}} = 508$ and $n_{\text{G:S}} = 36$ they form the largest groups of decisive and indecisive respondents, respectively, such that the interpretation of corresponding regression coefficients is comparably trustworthy. Especially in the context of estimators for indecisive groups, we remark that some of the regression coefficients' calculations are based on few observations, and thus corresponding interpretations have to be treated cautiously.

Furthermore, in context of interpretation one should check by taking the statistical significance[12] into account whether the regression coefficients vary just randomly. The small sample size within several groups of variable "ontic" may be responsible for non-significant estimators. Thus, from an increase in sample size statistical uncertainty is reduced and potentially significant results can be obtained.

Considering the results of the second vote analysis presented in Table 3 (ontic)[13], for Christian respondents the probability of electing "CD" instead of "SPD" is increased by the multiplicative factor $\exp(0.32) = 1.38$ compared to non-Christian respondents under the ceteris paribus assumption of unchanged other covariates.[14] Furthermore, regression coefficients closely to zero indicate that no influence of covariate "most important information source" on the probability of electing "CD" in comparison to the reference category "SPD" may be verified.

The crucial property of the multinomial regression under ontic imprecision consists of estimating own coefficients for the different indecisive groups. For instance, for respondents reporting "newspaper" as their most important information source in comparison to those naming another information source the probability of being indecisive between the two parties "GREEN" and "SPD" instead of voting for "SPD" is decreased by the factor $\exp(-1.67) = 0.19$ on the ceteris paribus premise. Likewise investigations are important for election campaigners to adjust their strategies adequately, as they show how potential voters differ from the core voters of a party (as here "SPD") in the choice of their favourable information source.

Results from a classical analysis that chooses variable "vote" as response variable and takes only those respondents into consideration that are "very certain" or "certain" may be found in Table 3 as well, again just displaying coefficients for "CD".

Comparing results from both analyses, estimators of similar magnitude are obtained throughout. In this way, the classical and the generalized approach reflecting ontic imprecision do not contradict each other.

The importance of our ontic set based modelling is corroborated even stronger when we consider the first vote instead. Now the analyses reveal remarkable differences partly associated with a change in sign. Thus, some covariates have an amplifying effect on the dependent variable in one analysis, while in the other analysis a weakening effect is underlying (cf. Table 4), yet those are not statistically significant.

Although the complete case analysis and the carried out filtering steps mainly induced by the questionnaire design led to a further decrease in the number of indecisive respondents, this illustrative analysis already shows striking differences between both analyses. Because of the here provided proof of concept for an ontic analysis, it is strongly suggested to include the option of reporting multiple answers such that those can be

[11] Estimated regression coefficients for the other categories may be found in [22]

[12] "***", "**" and "*" denotes statistical significance of level $\alpha = 0.01$, $\alpha = 0.05$ or $\alpha = 0.1$, respectively.

[13] Covariates "religious denomination" and "most important information source" are dummy coded with "non-Christian" and "other source" as reference category, respectively. The estimates quantify the difference between the group under consideration and the reference category (rel.christ: "religious denomination" is "Christian"; info.tv, info.np: "most important information source" is television, newspaper, respectively).

[14] Despite the name "CD" and the above results indicating a strong Christian relation, nowadays the "CD" parties understand themselves as a general conservative party with members and supporters regardless their religious affiliation.

Coefficient	ontic		classical
	CD	G:S	CD
intercept	0.33	−1.41 **	−0.12
rel.christ	0.37 **	−0.25	0.52 ***
info.tv	−0.02	−0.32	0.25
info.np	−0.12	−1.69 **	0.13

Table 4: Comparison of results (first vote).

	ontic	vote	classical
Scenario 1	0.407 (0.040)	0.425 (0.050)	0.446 (0.041)
Scenario 2	0.704 (0.026)	0.796 (0.031)	0.817 (0.042)

Table 5: Correct classification rate (standard deviation) for second vote based on 10-fold cross-validation

included into the analysis in an appropriate way. In cases of large data sets with numerous indecisive respondents, we even expect increased differences in the estimation of regression coefficients.

5.3 Classification Trees Analysis

In a first scenario the settings are the same as we explored in the regression analysis, thus considering "ontic" coarse class variable and "religious denomination" and "most important source of information" as split feature variables, in the same scaling as previously in section 5.2 (Scenario 1). We are considering this setting to retain direct comparability with the regression analysis, yet we are aware that a classification tree's ability lies in reducing the sample space by discovering few favourable independent variables out of a potentially huge number of candidates. Therefore, we are not expecting an outstanding performance in this scenario. As discussed above we decided in favour of a Nonparametric Predictive Inference model as underlying (imprecise) model of the classification tree. We choose the most frequent class as prediction rule in the leaves, thus enforcing a precise result. Furthermore, we grew imprecise classification trees on the data set neglecting the undecided, but in this case we chose "vote" as the dependent variable as a counter part to the classical regression analysis. In order to assess the predictive ability of the trees a 10-fold cross-validation each was performed.

The results are to be found in the first row of Table 5, with respect to the second vote. For a fair comparison we measure the accuracy for both data situations by the correct classification rate (columns *ontic* and *classical*), and furthermore in case of the ontic data sets we checked the prediction result of "ontic" against "vote" (column *vote*). Any value of "vote" which was contained in the predicted coarse category was considered correctly classified. Furthermore the standard deviation is reported.

As it is clearly visible the predictive ability of the imprecise trees is unsurprisingly poor, and an inspection of the underlying trees reveals the culprits. The selection of the independent variables only allows growing of 13 different trees, which only in case of a strong dependency between the independent and depend variables leads to reasonable accuracy results. Furthermore when looking at the relative class frequencies in the root nodes, the category of "CD" is with over 40% by far the most observed one. While the construction of most trees involved at least one split, category "CD" is still predicted in a vast majority of the tree's leaves, in few cases even in all.

In further analyses, we incorporated more independent variables, allowing a higher variation in potential trees (Scenario 2). Further splitting candidate variables were the party identification (*q119*), the person's social stratum (*q192*), the sex (*q1*), general political interest (*q3*) and the personal economic situation (*q17*). With those and the previous variables the same analysing steps were repeated, but now with the accuracy nearly doubling in either scenario as the second row of Table 5 indicates. Especially the party identification has a high influence.

Similar prediction results as above are obtained when considering the first vote, instead of the second, displayed in [22]. Quite interestingly, the correct classification rate is lower when we are predicting the "ontic" variable than in the case when predicting "vote". In the second scenario there is a notable gap of around 10%, which is mainly caused by an ontic coarse class prediction, whereas vote is (naturally) precise.

In both scenarios the classical procedure of omitting the undecided persons leads to better results, when just considering the predictive ability, yet with the help of our ontic view we are able to identify hard to classify respondents.

A major reason for the small differences between the classical and ontic analyses is the comparably little percentage of undecided persons (less than 10% within the data under consideration). As mentioned in the discussion in the regression analyses, this is partly due to the conducted complete case analysis and the construction of variable "ontic", but more gravely imposed by the design of the questionnaire. When allowing for multiple answers directly in variable "vote", we expect an increase in the accuracy of the ontic prediction, as the number of hard to precisely classify, indecisive persons raises.

5.4 Interval-valued Forecast

In Section 3 the epistemic view has been used in order to calculate interval-valued forecast $I(E)$, which will be illustrated in this section.

For instance, if one is interested in the forecasted proportion of respondents electing "CD", by referring to the absolute frequencies of variable "ontic" in Table 2 and to Equation (5), the interval-valued forecast

$$\widehat{I}\big(\{\text{CD}\}\big) \;=\; \left[\, \frac{495}{1196} \,,\; \frac{495+35+18+17+12}{1196} \,\right]$$

is obtained. All fractions that are included in the lower bound refer to respondents who vote for the "CD" party for sure while all fractions that are used within the calculation of the upper bound concern respondents who generally could imagine to vote for it. Political studies gradually proceed to calculate the fraction of "potential voters" which corresponds to the upper bound of interval $\widehat{I}(E)$ (cf. [1]).

Nevertheless, forecasts are commonly based on respondents that are characterized by a high degree of certainty concerning their voting intention only. In our data example there are $n = 1096$ respondents that are "very certain" or "fairly certain" according to their voting intention, where 490 of those intend to vote for "CD" and thus the naive estimated forecasting probability results in

$$\widehat{P}_{\text{naive}}\big(\{\text{CD}\}\big) \;=\; \frac{490}{1096}\,.$$

As indecisive voters may systematically differ from respondents that are sure of their voting intention, the proportion in terms of interval $\widehat{I}(E)$ contains valuable information that is not expressed by $\widehat{P}_{\text{naive}}(E)$. Because of the difference between these groups it is important to treat results ignoring indecisive respondents with caution.

In practice forecasting the proportion of a set containing more than one element is of considerable relevance: Frequently, for instance in Germany, the main interest is the voters' percentage not just for a particular single party, but for a coalition. In this context the interval-valued forecast $\widehat{I}(E)$ becomes of particular interest, as respondents that are indecisive between the parties contained in the coalition of interest E are incorporated for sure. Thus, these coarse observations constitute a precise vote for the coalition (e.g. [22]).

6 Concluding Remarks

While currently data under ontic imprecision are still neglected in statistical analysis, they could prove a valuable source of information. Especially in context of election studies incorporating the different types of "The Undecided" into statistical analyses becomes increasingly important as more and more voters decide shortly before the election day (cf., e.g. [26]). Once the practitioner changes the state space, the statistical methods remain the same, as we could demonstrate. Even as the group was comparably small and we were forced to assess indecisiveness indirectly by constructing an ontic variable, we corroborated in our data example that including the undecided respondents did make a difference. Therefore, as now appropriate statistical methodology has been proven to be available, we strongly recommend allowing for multiple answers directly within questionnaires.

As the underlying idea is somewhat generic, the in here presented analyses by a multinomial regression model and imprecise classification trees are just the tip of the iceberg. One may think of more complex methods to study the data set, mutatis mutandis. For simplicity we restricted ourselves to the case of a nominal scale of the variable under ontic imprecision, yet the adaptation to an ordinal scale is achievable with little additional effort as well. In further studies it is worth considering not only the dependent variable under ontic imprecision but also the covariates. In principle, this is achievable by involving the power-set based idea again.

Acknowledgements

We are grateful to two of three anonymous reviewers for their very helpful remarks, also stimulating further research.

References

[1] Großteil der Wähler würde sich noch umstimmen lassen. *Süddeutsche Zeitung*, 16 August 2013. Accessed 24 January 2015, `http://www.sueddeutsche.de/politik/umfrage-zur-bundestagswahl-die-meisten-waehler-wuerden-sich-noch-umstimmen-lassen-1.1747539`.

[2] J. Abellán and A. Masegosa. Bagging decision trees on data sets with classification noise. In S. Link and H. Prade, editors, *Foundations of Information and Knowledge Systems*, pages 248–265. Springer Berlin Heidelberg, 2010.

[3] J. Abellán and A. Masegosa. An ensemble method of using credal decision trees. *European Journal of Operations Research*, 205(1):218–226, 2010.

[4] J. Abellán and S. Moral. Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, 18(12):1215–1225, 2003.

[5] J. Adams, S. Merrill, and B. Grofman. *A Unified Theory of Party Competition: A Cross-National Analysis Integrating Spatial and Behavioral Factors*. Cambridge University Press, Cambridge, 2005.

[6] R. Alvarez and J. Nagler. When politics and models collide: Estimating models of multiparty elections. *American Journal of Political Science*, 42(1):55–96, 1998.

[7] T. Augustin, G. Walter, and F. Coolen. Statistical inference. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 135–189. Wiley, 2014.

[8] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth Books, Monterey, CA, 1984.

[10] F. Coolen and T. Augustin. A nonparametric predictive alternative to the Imprecise Dirichlet Model: The case of a known number of categories. *International Journal of Approximate Reasoning*, 50(2):217–230, 2009.

[11] I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55(7):1502–1518, 2014.

[12] I. Couso, D. Dubois, and L. Sánchez. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables*. Springer, Cham, 2014.

[13] R. Crossman, J. Abellán, T. Augustin, and F. Coolen. Building imprecise classification trees with entropy ranges. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger, editors, *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 129–138, Innsbruck, 2011. SIPTA.

[14] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[15] D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, New York, 1988.

[16] D. Dubois and H. Prade. Formal representations of uncertainty. In D. Bouyssou, D. Dubois, M. Pirlot, and H. Prade, editors, *Decision-Making Process: Concepts and Methods*, pages 85–156. ISTE & Wiley, London, 2009.

[17] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: Models, Methods and Applications*. Springer, Berlin, 2013.

[18] P. Fink and R. Crossman. Entropy based classification trees. In F. Cozman, T. Denœux, S. Destercke, and T. Seidenfeld, editors, *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*, pages 139–147, Compiègne, 2013. SIPTA.

[19] G. Matheron. *Random Sets and Integral Geometry*. Wiley, New York, 1975.

[20] H. Nguyen. *An Introduction to Random Sets*. CRC Press, Boca Raton, Florida, 2006.

[21] J. Plass, T. Augustin, M. Cattaneo, and G. Schollmeyer. Towards statistical modelling under epistemic data imprecision: Some results on estimating multinomial distributions and logistic regression under coarse categorical data. Under revision for ISIPTA '15, preprint temporary available at `http://www.statistik.lmu.de/~jplass/forschung.html` (20.03.2015).

[22] J. Plass, P. Fink, N. Schöning, and T. Augustin. Statistical Modelling in Surveys without Neglecting "The Undecided": Multinomial Logistic Regression Models and Imprecise Classification Trees under Ontic Data Imprecision - extended version. Technical Report 179, University of Munich, Department of Statistics, 2015. `http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-23816-6`.

[23] J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[25] H. Rattinger, S. Roßteutscher, R. Schmitt-Beck, B. Weßels, and C. Wolf. Vorwahl-Querschnitt (GLES 2013), 2014. GESIS Datenarchiv, Köln. ZA5700 Datenfile Version 2.0.0, Accessible from `https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5700&tab=3&ll=10&notabs=&af=&nf=1&search=gles&search2=&db=e`.

[26] O. Schirg. Wahlforscher: Jeder Dritte ist noch unentschlossen. *Die Welt*, 10 August 2001. Accessed 22 January 2015, `http://www.welt.de/print-welt/article467015/Wahlforscher-Jeder-Dritte-ist-noch-unentschlossen.html`.

[27] C. Strobl. Variable selection in classification trees based on imprecise probabilities. In F. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, pages 339–348, Carnegie Mellon University, Pittsburgh, 2005. SIPTA.

[28] P. Thurner. The empirical application of the spatial theory of voting in multiparty systems with random utility models. *Electoral Studies*, 19(4):493–517, 2000.

[29] K. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

[30] G. Tutz. *Regression for Categorical Data*. Cambridge University Press, 2011.

[31] S. Vansteelandt, E. Goetghebeur, M. Kenward, and G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16(3):953–979, 2006.