Psychologie
Faculteit der Sociale Wetenschappen

Constructing latent classes to predict dropout in interventions for multiproblem young adults

Master thesis Methodology and Statistics

August 2016

Sandra van Zoelen

Student number: S1316494

Internal supervisor: Dr Zsuzsa Bakk

External supervisor: Dr Carmen Paalman, Academische werkplaats bij De Nieuwe Kans

Acknowledgements

I would like to thank the people who have helped me the past six months to finish the last project of the master program 'Methodology and Statistics'. First of all I would like thank my supervisor Dr. Zsuzsa Bakk for her guidance, many helpful ideas and all of her advices on this thesis. Secondly, I would like to thank my advisor Laura van Duin at the 'Academische Werkplaats' for her guidance during my internship and for providing me the dataset for this thesis project. Her enthusiasm for this field of research has been catching and inspired me to write this thesis. Last of all I would like the wonderful people around me, especially my boyfriend Mitchell, for all the support and encouragement during this process.

Abstract

Multiproblem young adults form a major problem for the community and intervention programs are often not successful. One major problem is the large amount of dropouts and little is known about the cause of this. This study tries to cluster participants based on indicators that are predictive for dropout. The outcome could be used to alternate the available program to the needs of each sub group. Because there are many possible indicators it may be better to narrow down the number of included indicators before performing latent class analysis. Normally this would be done using latent class variable selection, but because we deal with small sample size and we prefer classes predictive of drop out it may be better to use another method. The option tested in this study is using the supervised learning methods lasso regression and random forest to select variables that are predictive for drop out and to use these variables as indicators for LCA. A simulation study is used to compare the classification error rates from these two models with a LCA model using al the indicators.

The outcome was that when there were five or ten important indicators and the predictors were strong, the method using lasso regression performed better than the other two methods. For the other situations it did not matter which method was used considering prediction accuracy. Taking a look at the young adults only the method using random forest could find a well-fitting model. However, none of the models formed a good prediction model for drop out. We can thus conclude that the variables used in this study are not predictive for drop out. Larger sample studies must confirm this. We have learned from this study that variable selection can be a good thing to do in some situations, further studies must test how this procedure works with different sample sizes and a larger amount of indicators.

Samenvatting

Multiprobleem jong volwassenen vormen een groot probleem voor de samenleving en interventies zijn vaak niet succesvol. Een groot probleem hierbij is het grote aantal uitvallers en het is onduidelijk wat hier de oorzaak van is. Deze studie probeert de deelnemers middels latente klassen analyse te clusteren op basis van variabelen die uitval voorspellen. De uitkomst kan gebruikt worden om bestaande interventies aan te passen en beter aan te laten sluiten op de behoeften van de verschillend subgroepen. Er zijn veel mogelijke indicatoren beschikbaar, daarom is het wellicht beter om dit aantal te beperken voor het uitvoeren van de LCA. Normaal wordt hiervoor latente klas variabelen selectie voor gebruikt, maar omdat we hier te maken hebben met een kleine steekproef is het wellicht beter om een andere methode te kiezen. De opties die in deze studie gebruikt worden zijn de ‘statistical learning’ technieken lasso regressie en ‘random forest’ om variabelen te selecteren die uitval het best kunnen voorspellen. Met behulp van een simulatie studie worden de classificatie fouten van deze methodes vergeleken met die van LCA met alle indicatoren.

De uitkomst van deze studie was dat bij vijf of tien belangrijke indicatoren en wanneer uitkomst sterk samenhangt met deze, de lasso regressie methode beter werkt dan de andere twee. Voor de andere situaties maakten het niet uit welke methode gebruikt werd, kijkend naar de voorspelling’s precisies. Wat betreft de jong volwassenen data kunnen we concluderen dat alleen de methode die ‘random forest’ gebruikt een goed passend model kon vinden. Geen van de modellen hier kon echter een goede voorspelling maken voor uitval. Op basis van deze studie kunnen we dus zeggen dat de hier gebruikte variabelen niet voorspellend zijn voor uitval. Onderzoeken met grotere steekproeven moeten uitwijzen of dit daadwerkelijk zo is. We kunnen uit dit onderzoek opmaken dat het onder sommige omstandigheden belonend is om het aantal indicatoren te beperken, volgende studies moeten testen hoe deze methode werkt bij een grotere steekproef en een groter aantal indicatoren.

Table of contents

Introduction

1.1 Multiproblem young adults

Multiproblem young adults comprise a vulnerable group who have had problems making the transition from adolescence to adulthood and who encounter difficulties in different aspects in life, such as income, low or no education, housing, delinquency, social relationships and health (Berniz, 2010; Osgood, Foster & Courtney, 2010). Most of them have had a history with the mental health, childcare and juvenile justice system because of an unstable family environment and delinquent behaviour before the age of eighteen. Family problems and early criminal behaviour can have disadvantageous influences on many aspects later in life (Van Domburgh, Vermeiren, Blokland & Dorreleijers, 2008; Loebe & Farrington, 2000). Early offending is associated with low school motivation, poor social skills, unemployment, depression, suicide and early substance abuse (Grant & Dawson, 1998; Lewis, Shanok, Grant, & Ritvo, 1983; Loebe & Farrington, 2000; Spies & Davelaar, 2015). Delinquent behaviour often occurs in the presence of risk factors such as low intelligence, parental neglect, physical punishment, parental abuse, single parenthood and low school motivation, which are just a few of many risk factors for child and later serious and violent offending (Stouthamer-Loeber, Wei, Loeber & Masten, 2004; Noom, van der Veldt, Houdt & Slot, 2009). The more risk factors are present, the higher the chance of occurrence of problem behaviour is. Psychopathology is related to development and continuity of delinquent behaviour and mild intellectual disabilities increase susceptibility for the above mentioned risk factors (Noom, van der Veldt, Houdt & Slot, 2009). Risk factors for the continuity of delinquent behaviour in (young) adulthood are hard drug use, gang membership and serious delinquent behaviour during adolescence (Stouthamer-Loeber, Wei, Loeber & Masten, 2004).

1.2 Profiles in youth delinquency

Three types of youth offenders were distinguished by theory by Moffitt (1993) and Aguilar, Sroufe, Egeland & Carlson (2000). The first one is the childhood limited type, in which delinquency stops during puberty or adolescence, the two other types show more persistent delinquent behaviour. The second type is the adolescent onset type, these youth start during adolescence and persist during adulthood. They start with minor offences, which often lead to more serious offences in the future (Loebe & Farrington, 2000; Snyder, 1998). They are mostly influenced by delinquent friends and are quit susceptible for interventions

(Noom, van der Veldt, Houdt & Slot, 2009). The last type is the lifetime persistent type, these youth commit many and severe offences. Problem behaviour starts often already at lower school. It mostly starts with externalizing and aggressive behaviour, which predicts delinquent behaviour later in life. Youth from the lifetime persistent type often show neurological abnormalities before birth and have dealt with multiple risk factors early in life (Noom, van der Veldt, Houdt & Slot, 2009). Committing minor offences before the age of thirteen increases the risk of committing violent and chronic offences later (Lahey and Waldman, 2005; Loeber et al., 1993; Loeber & Farrington, 2000; Moffitt, 1993; Patterson et al. 1998). The adolescent onset and the lifetime persistent type can both be split up in aggressive and non-aggressive delinquency.

1.3 Multiproblem young adults in Rotterdam

Rotterdam has many multi-problem young adults, fourteen percent of the adolescents in Rotterdam between the age of fifteen and twenty-seven do not follow education and are unemployed and two third of them do not have a starters qualification (Spierings, Tudjman, Meeuwisse & Onstenk, 2015, Spies & Davelaar, 2015). This may not seem to be a large group, but this group is responsible for a large amount of the criminality in this city. These young adults who encounter problems in the fields of for example education, employment and housing can come to 'het Jongeren Loket' where they receive help finding a job or education or where they are possibly directed to a daytime intervention program. Multiproblem behaviour is assessed with the Self Sufficiency Matrix (SSM), a scale to measure the skills in daily functioning (Hammink & Schrijvers, 2013). Eighty percent of the 5700 young adults who have an intake at 'het Jongeren Loket' falls within the target group of multiproblem young adults. 4300 of them go into an obligatory searching period, the month after intake in which the young adults try to find a job or education by themselves before they are eligible for a daytime program or an allowance. Most interventions consist of a daytime program which prepares the participants to go back to school or work. 33 percent of the group that goes into searching period does not come back after the obligatory month. Apart from the young adults who do not come back after the searching period, another 1800 drop out during the intervention and only 840 (fifteen percent of the initial intake number) have a positive outflow to work or education. This shows that not all of the intervention programs for these young adults are successful and there is need to take a closer look on why this is the case. The focus of this study is aimed at the young adults who go into searching period and/or register

for the intervention program but who do not enter the program, this will be denoted with drop out in this paper. Our main substantive interest is to understand what is it that prevents them from coming back?

No research has yet been done concerning dropouts in interventions within this population, it would however be useful to investigate which characteristics are related to drop out. It is imaginable that experiences from their youth and their contacts with the social services in the past are related to susceptibility for treatment now, since it is proven that this also relates to the severity of problems later in life. The young adults may have adverse memories from facilities which makes them suspicious for interventions. It would therefore be possible that the earlier mentioned risk factors for problems later in life are also related to drop out. Some subpopulations who have specific patterns of risk factors may be less susceptible than others and may need a different approach. As mentioned before it is proven that having dealt with multiple risk factors leads to a higher chance of negative outcomes later in life. This could also be the case for the outcome of dropout.

1.4 Gaining an insight into the population of multiproblem young adults and dropout

To get a better understanding about why the young adults drop out from the track it would first be useful to investigate what the group of multiproblem young adults looks like and which characteristics are related to drop out. Mulder, Vermunt, Brand, Bullens & van Marle (2012) found for example four different groups using (LCA) in juvenile offenders with different risk factors who reacted differently to intervention. LCA is an unsupervised learning technique that classifies cases based on categorical indicators and has been widely used in psychology to reveal underlying structures (Nylund, Asparouhov & Muthén, 2007). LCA is a technique that compares response patterns and classifies response patterns that are alike. As mentioned above, we expect that certain patterns of risk factors relate to higher chances of drop out and with LCA it would be possible to capture these response patterns and to analyse its relationship to dropout. Research on this topic is still missing in literature, most research about delinquency and multiproblem behaviour has been done about adolescents and adults aged from eighteen to sixty, not specifically about young adults which is found to be a different subpopulation (Osgood, Foster & Courtney, 2010). Since young adults need to be studied as a separate population they may also differ in risk factors and LC's. Thereby has dropout never been taken into account in studies, even while this is one of the main issues in interventions.

Because drop out is an aspect that is worth including in the analysis, it would be possible to build a LCA model that aims at explaining differences in dropout. It is imaginable that there is a particular subgroup that has dealt with certain risk factors in their youth and that this subgroup is less prone to entering a intervention program than another subgroup where this is less the case. Performing LCA with the goal of predicting drop out could show for who the available approaches and programs work and for who alternatives should be found. Differences in item probabilities among classes could be used to direct what extra needs should be taken in consideration in adopting the available programs.

A simpler option may be to just analyse the effects of the risk factors directly on the outcome variable. It is however, as early described, found that the co-occurrence of multiple risk factors increases the chance of more severe problems and thus perhaps also affects drop out. In LCA it is possible to find certain subgroups that have dealt with a certain combination of similar risk factors and this co-occurrence could be related to drop out. With other predictive analyses this should be done with many interaction terms and this would lead to unreliable results in a limited sample size.

Traditionally a LCA model would first be fitted to the data, after which the corrected class membership would be used as an independent variable in analysing the relationship with the outcome. The problem here is that there is a large amount of possible indicators which requires a large sample size to get a reliable result in LCA, which is not always achievable. The performance of LCA depends among others on the amount and quality of indicators (Dziak, Lanza & Tan, 2014; Wurpts & Geiser, 2014). With many indicators and small sample size the risk of data sparseness and item probabilities of one or zero increases. The latter risk is unwanted because it gives the idea of perfect indicators with no standard errors which is impossible in reality. Minimizing the number of (low quality) indicators in LCA is expected to improve the precision of the parameter estimates and the performance of the classification because noise variables are removed from the model (Dean & Raftery, 2010). Dean & Raftery (2010) have been the only ones investigating these possible problems in LCA and found that more indicators did not necessarily lead to unreliable results and emphasized that more research is needed in what influences the reliability of LCA models. Little research has been done after the required sample size in LCA and there is still some uncertainty about the reliability of the estimates in LCA with small sample size. Another problem with many indicators is that it becomes more difficult to get a clear overview of the model when there are many noise variables present.

The mainstream procedure to decrease the number of indicators is Latent Class Variable Selection, this model can iteratively find the most important variables for LCA (Dean & Raftery, 2010; Ghosh, Herring & Siega-Riz, 2011). The LCA model that would be fitted in this technique does only involve indicators and no covariates and the relation to the outcome variable would be analysed separately in the next step. The first step is to fit the LCA model with the maximum number of LC's that is possible to be identified with all the available indicators included. The variables are then ranked based on the summed variability between classes. High variability between classes indicates that the variable distinguishes good between LCs. Then the model with the fewest possible variables that are needed for an identifiable two class model is fitted, which uses the variables with the highest rank. In the next steps variables are in- and excluded by comparing the BIC values of the model with and without the variable considered in this step. Two cut-offs are used to in or exclude variables: If the change in BIC is below the first cut-off, the variable is completely deleted from the selection procedure. If the change in BIC is between the two cut-offs, the variable is not included and will be put on the end of the list of possibly included variables. The exclusion step works in a similar way. This procedure is repeated till the list of possibly included variables is empty or till the list of included variables does not change after multiple inclusion and exclusion steps. In the latter situation the algorithm will stop because of convergence.

The drawback of this procedure is that in the first step a model is fitted with a large number of LCs when you have many possible indicators. When the ratio between the number of possible indicators and sample size is small this may lead to a unreliable model because of data sparseness. The ranking of the variables is based on this model so it would be possible that variables that are placed in front are not definitely the most important but are kept in the model. Small sample size could thus pose problems in this procedure. Other drawbacks are that this model was developed for continuous indicators and that the procedure takes a lot of computation time because many LCA models need to be fit because of the many including and excluding steps. Another comment that could be made for this particular research question would be that we would like to create classes that are related to drop out, but this procedure may leave out indicators that are average indicators for LCA but are perhaps important in predicting dropout. Because small sample size is the main problem in this study with a sample size of 219, we deal with categorical indicators and it is preferable to construct LC's that are related to dropout, it may be affordable to find an alternative approach to decrease the number of indicators.

Because latent class variable selection is the only available method in literature to reduce the number of indicators I will introduce a new method that may work better in certain situations. This method would be selecting variables for LCA based on their relationship to the outcome variable, which is advantageous because the LC's will be more predictive for the outcome variable than in latent class variable selection. A disadvantage is that the variables that are important in predicting the outcome are possibly not the best variables to conduct latent class analysis. If the LC's are weak it may not be possible to draw conclusions about the differences between LC's and there would be no basis to adjust the intervention programs. This method has not yet been used but may be better in estimating dropout from LC than latent variable selection, or traditional LCA including all indicators when sample size is small. To select important variables for predicting drop out a supervised learning method could be used, such methods can be modelled to increase prediction accuracy.

Random forest for example is a supervised learning technique that is based on decision trees but has greatly improved prediction accuracy at the cost of interpretability (Sut & Simsek, 2011). This procedure combines bagging and random selection of features at each node (Breiman, 1996; Breiman, 2001). This decreases variance by averaging and decorrelating the fitted trees. In random forest the most important variables can be determined by using the variable importance which uses the mean decrease in Gini index. The Gini index can be seen as a measure of node purity, lower values indicate more identical values for the classes in a node and thus better model fit. Variables that have a high mean decrease Gini index thus contribute to increased node purity. The formula for the Gini index is shown in equation 1, where $\widehat{P}_{mk}$ is here the proportions of class K in split m

$$G = \sum_{k=1} \hat{P}_{mk} \left(1 - \hat{P}_{mk}\right). \quad (1)$$

Random forest is often used in ecology studies because it performs well with many variables and few cases, it would therefore also be a suitable procedure in this study because we deal with small sample size (Cutler, Edwards, Beard, Cutler, Hess, Gibson & Lawler, 2007).

Another method that performs well in selecting important features is Lasso penalized regression, which performs subset selection and linear regression at once (Osborne, Presnell, Turlach, 2000). With the lasso penalized regression we try to find the lasso coefficient $\hat{\beta}_{\lambda}^{L}$, the coefficient that shrinks the regression parameters that minimize equation 2. The equation is calculated based on 1,2,3…n participants and 1,2,3…p variables where *xij* is the observation

of participant *i* on variable *j*. RSS is here used to denote the residual sum of squares, λ the tuning penalty, y the depended variable and β the slopes.

$$\sum_{i=1}^{n}\left(Y_i-\beta_0-\sum_{j=1}^{P}\beta_j xij\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|=RSS+\lambda\sum_{j=1}^{p}\left|\beta_j\right|. \qquad (2)$$

The shrinkage parameter can be determined by using cross validation, which decreases uninfluential estimates (Tibshirani, 1996; Wu & Lange, 2008). Prediction accuracy and interpretation are improved because some uninfluential variable's parameters are set to zero, which reduces the variance on the cost of increasing the bias slightly. Lasso penalized regression performs well even when the number of parameters exceeds the number of cases (Sancetta,2016; Wu & Lange, 2008). With Lasso penalized regression mostly a few variables are found to be greatly influential and these could be used for LCA, in this way the problem of many indicators is overcome and it is known in advance that the LC's are related to dropout.

1.5 Research questions

This study investigates a substantive and a methodological research question. The methodological questions is: Which approach performs best in constructing LC's that can be used to predict the outcome variable with small sample size and many indicators: LCA with all the variables included or LCA with the risk factors that are found to be important by a supervised learning method? These methods are used because evidence is still needed that decreasing the number of indicators can improve the LCA method. The latent class variable selection method is not considered here because we aim here at explaining an outcome variable and we deal with small sample size and categorical indicators. The sub question would be: which supervised learning procedures would perform best in combination with LCA and should be used to decrease the number of indicators? LCA with all the indicators and LCA in combination with a supervised learning technique will be compared in a simulation study on the basis of prediction accuracy of the outcome variable.

In the simulation study multiple datasets will be created with different numbers of important indicators, different strengths of predictors of outcome and overlapping or not overlapping important indicators and predictors. The substantive research question will be

used to demonstrate the above described procedures and is: can the participants be divided in LC's that are related to drop out? Both models will be fit to the data to see which model creates LC's that are most related to dropout: Latent class variable selection or LCA using the variables that were found to be highly influential in the supervised learning technique.

For the methodological research question I expect that the LCA with all the indictors will perform better when the number of important variables is large in comparison to the number of noise variables. With fewer important indicators it would be expected that the LCA in combination with the supervised learning techniques performs better because the model with all indicators will then include many variables that do not contribute to predicting the outcome. I expect that the results for the supervised methods will be better when the important indicators overlap with the important predictors because they select variables based on the relation with the outcome and would then also select good clustering variables. The supervised learning method is expected to perform better because we are dealing with small sample size and many indicators, when all indicators are included this could lead to data sparseness. Comparing the two supervised methods I would expect that the result will be approximately the same. Both methods are proven to perform well under the circumstance of small sample size.

For the substantive research question I expect that the variables that are important for clustering will be more or less the same as the variables that are important for predicting drop out. I expect that the classes to be found will include the same variables that were found by Moffit (1993) because age of onset of problem behaviour and presence of aggressive behaviour are documented in this study and are expected to be related to susceptibility to intervention. Not all variables overlap with the ones used in their study so the result may differ somewhat.

For answering the substantive research question the data of a dossier study among multiproblem young adults aged eighteen to twenty-seven in Rotterdam will be used. The data is part of a larger project in which a lot of information about the participants is gathered from the beginning of an intervention program till one year after the start of the program. Most of the multi problem young adults have a history with social services and had a crime record before the age of eighteen. They often come from unstable families that could not offer them the needed support. Because they are likely to have had problems already before the age of eighteen most of them have a dossier at the 'Raad van de Kinderbescherming', in which information is provided about the history of delinquent behaviour, family circumstances and

provided social services. Fifteen categorical variables will be created from this file and will be used as the risk factors in the prediction and as the grouping variables in the LCA. 219 dossiers are available for this study.

Methods

2.1 Design

The data for the substantive research question comes from a larger cohort study conducted by 'Academische Werkplaats bij De Nieuwe Kans'. The study consist of four measurement moments (T0, T1, T2 and T3), neurological research and a dossier study. It is also documented whether and when participants drop out.

2.2 Procedure

For this study the dossiers will be used for clustering and drop out will be used as outcome variable. Participants that have come to 'het Jongerenloket' but who have not returned after the searching period or who do come back but do not actually start the program are considered as drop outs in this case. The participants were recruited at 'het Jongerenloket' and 'De Nieuwe Kans' or another intervention program. There was an experimental and control group, which are taken together in this study because we want to investigate the group of multiproblem young adults as a whole and not the difference between programs. The participants who are approached at the intervention have already been through the searching period and cannot become drop outs but they will be included in the sample. This will be done because their information is still informative about why they did pas this period and leaving them out would even further decrease the sample size.

2.3 Material

The dossiers at the 'Raad van Kinderbescherming' are used to derive the risk factors. Examples of variables are: how much research has been done by the child protection services, how many delicts are committed, was there a presence of child abuse, truancy, violence offenses and protective research. Most of the variables are categorical, but there are also some interval variables as age of first crime and age of first investigation. The interval variables will be made categorical by creating variables first crime/investigation took place before the age of fourteen.

2.4 Statistical procedures

Analyses will be performed using R Statistical Software (R Development Core Team, 2013). The random forest procedure will be performed using the package randomForest. The lasso penalized regression will be performed using glmnet. The LCA model in the simulation study will be fit using the function *e1071::lca* from the *e1071* package. For the real data the *poLCA* package will be used. Two different packages are used because the *e1071* does not work with data frames, the algorithms and thus the results are however the same for the two packages. The logistic regression will be performed using the *glm* and *predict* function from the *glm* package.

For choosing the best performing supervised learning method for the LCA procedure the classification error of predicting dropout from LCs will be compared. The optimal values for lambda will be found by using tenfold cross validation and will be used to find the variables with non zero estimates. The variables that are included with this optimal lambda value will be used for performing LCA. To be able to fit an LCA model with at least two classes there need to be at least three variables selected. With lasso analysis it is not possible to set a minimal number of variables that should be selected, so only LCA models are fit in the situations where at least three variables are selected. In the cases were not enough variables are found the LCA analyses are stopped, instead logistic regression will be performed. The results of the logistic regression will be separately reported.

In Random Forrest the subset of possible splitting variables will be set to the roots of the total number of predictors. In this procedure variable importance is used to select the most important variables. The drawback here is that there is no absolute cut off for the amount of predictors to be selected. The five variables with the highest mean decrease Gini index are selected in each sample. The five most important variables in this procedure will be used in the LCA model. The LCA model to which the two supervised learning methods will be compared is the model with all fifteen indicators. A simple LCA model without any covariates will be fit. For each method the optimal number of classes in the LCA model will be determined, this will be based on the Bayesian Information Criteria (BIC) value, which is found to be the best indicator in model comparison (Nylund, Asparouhov & Muthén, (2007). The formula for the BIC is presented in equation 3, where n is the number of parameters:

$$\text{BIC} = 2 \times \log(\text{maximized likelihood}) - n \times \log(n) \qquad (3)$$

The BIC will be calculated for 1, 2, 3 and 4 class models. Except for the lasso method, for which the maximum number of classes depends on the number of selected variables. When three variables are selected the maximum number of classes is two and with four variables the maximum number of classes is four. Because it is likely that different variables are chosen by the different methods, the number of optimal classes can also vary over the methods. The model with the optimal number of classes will be the final model and from this model the corrected class memberships will be used to predict the outcome variable. The independent variable, class membership, is categorical and the outcome variable, dropout, is also categorical and thus logistic regression will be used to analyse the relationship between LC and outcome. The predicted odds ratio from this analysis will be used to predict the outcome and will be compared to the actual outcome. This will be done for all LCA models and then the predictive accuracy (number of respondents correctly predicted as drop out or none drop out) of predicting dropout form LC will be calculated and averaged per condition and method. The model with the highest predictive accuracy in a specific condition will be designated as the best performing model for that condition. Because it is possible that in the same sample the number of optimal classes differ, the number of classes should also be taken into account. The fact that more or less classes provide a lower BIC does not mean that these classes are interpretable and meaningful, this should be taken into account in choosing the best model. The number of selected classes will be documented and compared with the actual number of classes. It is preferable that the number of classes found is identical to the actual number of classes, which will be three is the simulation study. All of the above procedures will be performed on the simulated data sets and on the real data example.

2.5 Simulation setup

2.5.1 Conditions simulation

The conditions of the simulation study will be varied on three aspects. The first aspect will be the number of important indicators for LCA, because it is not known in advance how many important indicators will occur in the real data. If the methods cannot find all the indicators this would have consequences for the model and the prediction of drop out. The number of indicators will be varied on three levels, one with no important indicators, one with five important indicators and one with ten important indicators. The strength of the LCA

model will increase as there are more important indicators. The number of important predictors for the outcome variable is set to five.

The second aspect is a weak relation between predictors and outcome and a strong relation between predictors and outcome. The strength of the unimportant predictors will be the same in all situations but the strength of the important predictors will have two levels. In the next section will be explained how the relation between predictors and the outcome variables is modelled.

The last aspect that will be varied is whether the important indicators overlap with the important predictors or not. In the situation with only unimportant indicators this is of course not possible. For the situation with five important indicators the important indicators will be selected in a way that they do not overlap, partly overlap and completely overlap. In the complete overlap situation the five variables that are strong indicators for LC will also be the five variables that are predictive for the outcome variable. In the partial overlap situation there will be two variables that are strong indicators for LC and these will also be predictive for the outcome variable. The other three variables that are strong indicators for LCA will not be predictive for the outcome variable. In the no overlap situation the five variables that are strong indicators for LC will not be predictive for the outcome variable. For the situation with ten indicators it would not be reasonable to have no overlap between important indicators and predictors so here we will only have partly and complete overlap. In the situation with complete overlap five out of the ten variables that are important indicators for LC will also be predictive for the outcome variable. In the partial overlap condition two out of ten indicators that are important indicators for LC will also be predictive for the outcome variable. If the important indicators are not the same ones as the important predictors, latent class variable selection would find meaningful classes that are not related to the outcome and LCA with supervised learning methods would find non meaningful classes that are related to the outcome. In section 2.2 it will be explained how the value for the outcome variable is determined.

Sample sizes and number of indicators will be set to the values in de real data, 220 and 15 respectively. This will be done because the aim of the simulation is to show whether the methods give reliable results in this sample and not to show what sample size would be sufficient to conduct such an analysis. In total there will be twelve different situations (table 1) and per situation hundred data sets will be generated.

Condition number	Number of important indicators	Strength of relation between predictors and outcome	Overlap
1	0	Medium	No
2	0	Strong	No
3	5	Medium	Complete
4	5	Strong	Complete
5	5	Medium	No
6	5	Strong	No
7	5	Medium	Partial
8	5	Strong	Partial
9	10	Medium	Partial
10	10	Strong	Partial
11	10	Medium	Complete
12	10	Strong	Complete

Table 1. Overview of the situations in the simulation study

2.5.2 Calculation of the values for the indicator and outcome variables

The indicators and the outcome variables will both be categorical, as is the case in the real data. For every condition hundred datasets are created to get a reliable result. The data is created by first creating a LC variable with three latent classes, this is the number of classes that is mostly found in this population. The indicators are created based on equation 4 and 5:

$$\ln\left(\frac{p}{1-p}\right) = 0.7 + \beta_1 x_j. \quad (4)$$

$$\beta_1 x_j = s_j * \left(yi - \frac{\sum y}{n}\right). \quad (5)$$

$\beta_1 x_j i$ is here the parameter for case i and indicator j which will be used for the logistic regression equation (4). N presents the number of cases , yi is the original class membership of case i, and s the strength of the relationship between indicators j and outcome which is manipulated in this simulation. With the function Rbinom the odds ratio according to logistic

regression equation 4 was used to generate binomial values for the indicators. For the noise indicators a value of 0.5 was used and for the strong indicators a value of 2 was used.

For the outcome variable the same method is used as for the indicators. The outcome variable is based on all fifteen indicators. Equation 5 was per predictor multiplied with the strength of that predictor and these values were summed and used in equation 4. A value of 2 was used for medium predictors and a value of 4 was used for strong predictors.

Results

3.1 Simulation study

In table 1 can be found which are the circumstances for each condition number which will be used in the other tables. For each of the twelve conditions the average correct classification ratio of the outcome variable and number of selected classes was compared. The classification ratios of drop out are shown in table 2 and the number of selected classes in the LCA model are shown in table 4. In table 3 can be found how many times different amount of variables were selected using the lasso in each of the conditions.

First let us look on the last column of table 2 that presents how many times the lasso found enough variables to be able to perform LCA. We can see here that this is often the case in de datasets with a strong strength for the relationship between predictors and outcome. In the other datasets the relationship is so weak that (almost) all coefficients are shrunken to zero and LCA was not performed in this case. In the datasets with a strong relation between predictors and outcome this only happens occasionally. The classification ratio's found in table 2 are calculated based on the dataset for which LCA is performed, these are thus not reliable for the lasso model in the medium strength condition because this statistic is based on too few datasets. For the Lasso procedure it was also documented in table 2 what the classification ratios were for the lasso penalized logistic regression for the samples for which too few variables were found to perform LCA. This statistic is thus only reliable for the samples in which few LCA in combination with lasso are performed.

	Average classification ratio of the outcome				
Condition number	**LCA with all indicators**	**LCA with lasso**	**LCA with random forest**	**Lasso penalized regression**	**Number of LCA lasso**
1	0.64	0.60	0.65	0.66	1
2	0.62	0.63	0.58	0.70	98
3	0.67	0.68	0.68	0.69	49
4	0.72	0.76	0.75	0.84	99
5	0.64	0.65	0.63	0.65	6
6	0.66	0.72	0.61	0.69	97
7	0.65	0.00	0.65	0.65	0
8	0.60	0.65	0.60	0.68	78
9	0.64	0.66	0.64	0.66	6
10	0.61	0.71	0.63	0.72	98
11	0.67	0.65	0.67	0.67	10
12	0.64	0.70	0.68	0.73	92

Table 2. Overview of average classification ratio of the outcome variable per condition

Next, in table 3, we see that in the situation where all fifteen indicators are equally weak related to latent class it is obvious that it makes no difference which method is used. This is the case because, as can be seen in table 3, all methods found that the one class model fitted best. The regression models for the outcome are thus the same for all three methods. In the situation of weak indicators it would thus not make sense to perform LCA with this sample size.

When five out of fifteen indicators are strong, the two class model was mostly found to fit best by all three methods. The LCA model including all variables selects the two class model in each sample under this condition, for the other methods there was more variation in the number of classes that were selected.

With ten important indicators the method including all indicators performed best in choosing the correct number of classes. This LCA model is thus the best fitting model. Comparing the classification ratios however, it shows that the supervised learning methods perform better under some circumstances even though the correct number of classes is not

always chosen. The variables chosen in this model thus do contribute to the prediction of the outcome but not to the model fit.

Taking a look at the results for the logistic regression in table 2 we see that the results for this method are comparable with the other two methods. We are here only looking at the datasets with medium predictors because enough logistic regression analyses are here performed to get a reliable result. In table 3 can be seen that in many datasets here no variables are included and the intercept only model was used. In other cases mostly one or two variable are found. It is thus appealing that logistic regression with only an intercept or with one or two independent variables gives the same result as applying logistic regression with LC as independent variable, which is based on all variables.

	Number of variables selected in lasso procedure				
Condition number	0 variables	1 variable	2 variables	3 variables	4 or more variables
1	48	35	16	1	0
2	0	0	2	16	82
3	1	15	35	38	11
4	0	0	1	2	97
5	23	53	18	4	2
6	0	0	3	3	94
7	68	26	6	0	0
8	0	2	20	32	46
9	35	38	21	5	1
10	0	0	2	7	91
11	31	37	22	9	1
12	0	0	8	29	63

Table 3. Overview of how many times the lasso choose a particular number of variables in each condition

The prediction accuracy rates for the lasso are even slightly higher than for those for the random forest procedure. This is especially the case with none overlapping important indicators and predictors and a strong relation between the predictors and the outcome, the

accuracy rates are 0.10 higher than for the random forest method here. In situation 4, 6 and 10 the accuracy rates are highest for the lasso, compared to the other situations, and this is also where the correct amount of classes, three, is more often found to fit best compared to other situations. The aspect overlap is different in these three situations, in one there is no overlap, in one partial overlap and in one complete overlap. Overlap between important indicators and important predictors thus does not seem to be an important aspect. In the samples with a medium relationship between predictors and outcome the accuracy rates lay closer together. In these samples the actual relationship between latent class and outcome is also less strong. In all situations the differences between the methods are small and with the highest prediction accuracy equalling 0.76 we can say that there is not a very strong prediction model found.

	LCA with all indicators			**LCA with lasso**			**LCA with random forest**		
Condition number	1 class	2 class	3 class	1 class	2 class	3 class	1 class	2 class	3 class
1	100	0	0	1	0	0	100	0	0
2	100	0	0	87	11	0	100	0	0
3	0	100	0	0	49	0	22	78	0
4	0	100	0	0	95	4	0	95	5
5	0	100	0	2	4	0	69	31	0
6	0	100	0	6	72	19	51	49	0
7	0	100	0	0	0	0	92	8	0
8	0	100	0	62	16	0	78	22	0
9	0	22	78	1	5	0	28	72	0
10	0	18	82	3	83	12	28	72	0
11	0	20	80	0	10	0	26	74	0
12	0	20	80	1	97	2	1	97	2

Table 4. Overview of the number of classes that are selected per condition

3.2 Multiproblem young adults

The sample consisted of 218 male participants after deleting two participants with missing values. The frequencies of the variables are presented in table 4. As can be seen here, most variables are not equally distributed. This often influences the strength of the LCA

model negatively. 62 percent of the sample has started the intervention, so this would be the expected classification ratios of drop out by chance. The classification ratios of drop out from the regression model using latent class as independent variable should thus be higher than 0.62 to improve prediction of drop out.

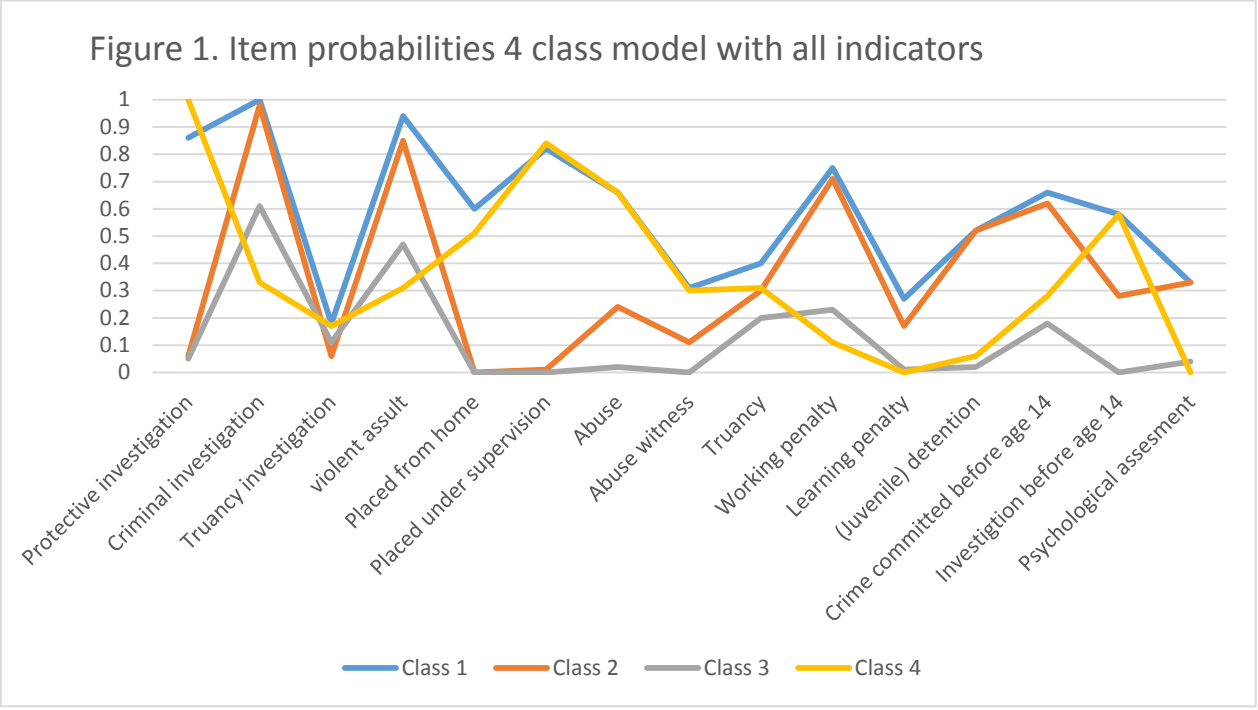
Variable	No	Yes
Started intervention	82	136
Protective investigation	126	92
Criminal investigation	46	172
Truancy investigation	191	27
Violent assault	67	151
Placed under supervision	140	76
Abuse	140	78
Abuse witness	182	35
Truancy	152	66
Working penalty	109	109
Learning penalty	189	29
(Juvenile) detention	150	68
Crime committed before age 14	118	100
Investigation before age 14	147	71
Psychological assessment	174	44
Been removed from home	163	54

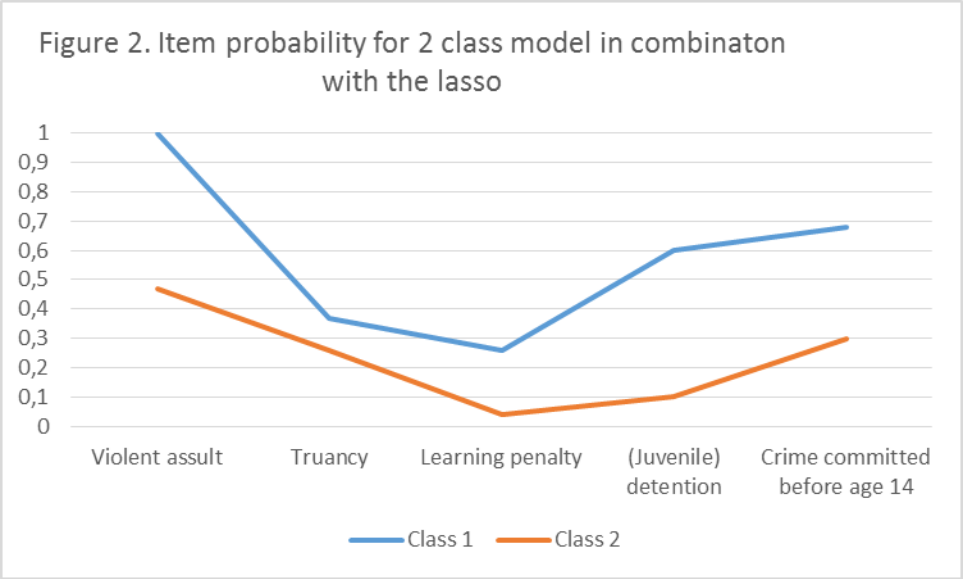
Table 4. Frequencies for the indicators and outcome

The first model that was fit was the LCA that included all indicators. The model with four classes had the lowest BIC value here. The model did not fit well (X^2 (155, 218) = 17958,17 = $p<0.5$). The classes were approximately equally distributed with predicted class memberships of 0.279, 0.291, 0.291 and 0.138. In figure 1 the item probabilities are charted for the four class model. What we see here is that the item probabilities for most items lay around the middle and there are not clear differences between classes visible. We can say that one group with relatively high rates of item probabilities for most items and one group with relatively low item probabilities for most items can be distinguished. There is one group that scores especially high on variables related to criminal behaviour and one that scores especially high on variables related to family circumstances. However, the difference between classes are hard to distinguish which is partly caused by the large amount of indicators. The

regression model with the class membership as independent variable predicted all outcomes to be positive (all participants started the intervention). The model did thus not perform better than by chance.

The second model that was fit was the LCA in combination with the random forest procedure. The random forest model was fit and the mean decrease Gini index was compared for the fifteen variables. The five variables that scored highest were selected for the LCA model, these were: Having committed an violent assault, presence of truancy, been imposed a learning penalty, having been in a (juvenile) detention institution and having committed a crime before age 14. The LCA model with these variables found that the LCA model with two classes fitted best. This model did fit well (X^2 (20,218) = 19.49 = p<0.9). The first class was smaller than the second class with predicted class membership proportions of respectively 0.372 and 0.628. The item probabilities for this model are charted in table 2. Class 2 scores lower on all variables, having almost zero score on being imposed a learning penalties and having been in a (juvenile) detention. When comparing these item probabilities with those from the all indicators included model we see that they are quit the same for both models when we take the classes that are alike in the four class model together. Even though they are much alike the fit of the model in combination with the random forest procedure is far better. Using the predicted class membership of this model again all outcomes were predicted as positive. Also this model does not perform better than by chance. Comparing the item probabilities from figure 1 and 2 we see here that the probabilities are approximately the same in both models, the model with five variables however, has better fit. Reducing the number of variables has not changed the item probabilities but has benefit model fit.





In the last model the lasso regression was used to select the variables for LCA. Only one variables had a non zero coefficient after the lasso penalty. This variable was whether protective investigation had taken place. Because only one variable was selected no LCA model could be fit. Logistic regression was performed on dropout with the variable found with the lasso to see whether this performed better than the LCA methods. Also with this analysis all outcomes were predicted to be positive. None of the models seemed to be good for predicting drop out.

Discussion

4.1 Conclusion

In this study it was investigated whether reducing the number of indicators in LCA improved the performance of the LCA model aimed at predicting an outcome variable. It was expected that this is the case using a supervised learning method when many noise variables are present and when there is overlap between important indicators and important predictors. It was also expected that the supervised learning methods performed better when the relation between predictors and outcome was strong.

In the simulation study it was found that that LCA performs better in combination with a supervised learning method in some, but certainly not all, circumstances. Under the condition of equally weakly related indicators all methods found that the one class model

fitted best, which resulted in equal prediction accuracy rates for the three methods. The strength of the relationship between the predictors and the outcome variable did not make a difference here. This is in line with what was expected from previous studies, when sample size is small you need stronger indicators to build a well-fitting model.

Overall, the number of classes was underestimated by all methods. The correct number of classes was only found in the case of ten strong indicators using the model including all indicators. The model with ten strong indictors was also expected to be the strongest LCA model. Comparing the two supervised learning methods it was found that the main disadvantage of the lasso method was that there were not always enough variables selected to perform LCA with more than two classes when the relationship between predictors and the outcome variable was not strong enough. However, the Lasso method outperformed the other two methods in the case of five and ten partly overlapping strong predictors. When the predictors are strong enough the lasso method seems to perform better than the random forest method, in the case of medium predictors no difference between the three methods could be discovered. Even the logistic regression using mostly only up to two variables performed equally well. This suggest that LCA is not needed in these cases. That the models perform equally weak can also be caused by the fact that the relation between LC's and outcome is less strong and thus prediction accuracies are expected to be lower in this situation. It can be concluded that in some situations it can be advantageous to reduce the number of indicators and that ,overall, including all indicators does only lead to better selection of the number of classes and not so much the prediction of the outcome. To build only a strong LCA model may would thus be better to use all indicators, but not when the aim is explaining an outcome variable. To successfully apply the supervised methods strong predictors must be available and they must at least partly overlap with the important indicators.

The substantive research question was whether LC's could be constructed in order to predict drop out. This was definitely not the case in this sample. None of the methods could give a good prediction of drop out. The circumstances in this sample seem to be in line with situation 1 from the simulation study: weak indicators and weak predictors. The predictors that were found to be important for predicting dropout using the random forest procedure are partly in line with those found in studies that investigated risk factors for later problem behaviour, which are: having committed a violent assault and having committed a crime before the age of fourteen. It was also found that penalties that were given during youth

(learning penalty and (juvenile) detention) influenced drop out. The relation with drop out was not strong however, which was proven by prediction of drop out which were not better than by chance.

The lasso regression could only find one variable that had a none zero coefficient and the LC's that were found by the other two methods were not predictive for the outcome. This is not in line with the expectations from previous studies. It could be possible that drop out is not as strongly related with risk factors as the severity of problem behaviour and criminality. It is thus still unclear what influences drop out and whether it is even possible to predict it. There are also some remarks that can be made on this study which can influence the result, besides of course the small sample size. A part of the sample of approached when being already in an intervention program and these could not become a drop out anymore, this could have resulted in underestimation of the dropout rates. When there would be more variability in dropout rates, its prediction may be easier. It would have been better to have recruited participants before going into searching period instead of already being in an intervention.

Comparing the three methods in this data set, some of the disadvantages of them have surfaced. The one that is most clear is that the lasso regression may restrict the number of indicators too much, only one variable was selected by this method so no LCA model could be fit. To use this method for LCA it would be preferable to be able to select at least the three most important variables. However, the LCA model that includes all variable did not perform well either. The model did not fit well which resulted in classes that were not very clearly distinguishable. The number of classes found was four, this was also not to be expected looking at previous research. The random forest procedure worked best in this dataset, the fit of the model was better than for the on including all variables, even though the item probabilities were quit similar. This could be a proof that noise variables can influence the performance in a negative way. Even though the model in combination with random forest could not find a good predictive model for drop out, it could make a good clustering model. It distinguished a class that had more problems on most items than the other group. Especially criminal activity seemed to be a larger problem in this group. Considering these variables the difference in risk factors between the clusters seems to be the amount of problems that were present.

4.2 Directions for further research

There are some components that are not taken into account in this study but that can possibly have an influence on the performance of the methods and should therefore be further investigated. Sample size for example was fixed in this simulation because it was not important for the substantive research question. It would however be interesting to see how these techniques work in larger samples. Can it also be affordable to reduce the number of indicators even when there is a sufficiently large sample? Evidence is also still missing on the question what actually is a sufficient large sample. Comparing the performance of LCA models with different sample sizes can contribute to this.

Another point for further studies would be the number of available indicators. In this study fifteen indicators were available with a sample size of 220, this is a relatively large amount of indicators in psychology but there are other fields of research in which they even have far more indicators with even smaller samples. The advantages of these techniques can therefore possible be greater for research field like biometrics and genetics.

Considering the substantive research question many questions are still unanswered. A well fitting two class could be found that showed an overall high and an overall low response pattern. But this response pattern did not seem to be predictive of drop out. More possible risk factors should be investigated to find the reason why dropping out poses such a large problem. Possibly more recent events are more influential and this is definitely worth studying.

Literature

Aguilar, B., Sroufe, L. A., Egeland, B., & Carlson, E. (2000). Distinguishing the early-onset/persistent and adolescence-onset antisocial behaviour types: from birth to 16 years. *Developmental psychopathology, 12*(2), 109-132.

Berzin, S. C. (2010). Vulnerability in the transition to adulthood: Defining risk based on youth profiles. *Children and Youth Services Review, 32,* 487-495. doi:10.1016/j.childyouth.2009.11.001.

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24,* 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32.

Capaldi, D. M., & Patterson, G. R. (1996). Can violent offenders be distinguished from frequent offenders: Prediction from childhood to adolescence. *Journal of Research in Crime and Delinquency, 33*(2), 206-231. doi: 10.1177/0022427896033002003.

Cutler, D. R., Edwards, T. C., Beard, K., H., Cutler, A., Hess, K. T., Gibson, J, & Lawler, J. J. (2007). Random Forests for Classification in Ecology. *Ecology, 88*(11), 2783-2792.

Dean, N., & Raftery, A. E. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics, 62,* 11-35. Doi: 10.1007/s10463-009-0258-9.

Van Domburgh, L., Vermeiren, R., Blokland, A. A. J., & Doreleijers, T. A. H. (2009). Delinquent Development in Dutch Childhood Arrestees: Developmental Trajectories, Risk Factors and Co-morbidity with Adverse Outcomes during adolescence. *Journal of Abnormal Child Psychology, 37*, 93-105. doi: 10.1007/s10802-008-9260-6.

Dziak, J. J., Lanza, S. T., & Tan, X. (2014). Effect Size, Statistical Power, and Sample Size Requirements for the Bootstrapling Likelihood Ratio Test in Latent Class Analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 534-552. doi: 10.1080/10705511.2014.919819.

Gottfredson, M. R., & Hirschi, T. (1990). A General Theory of Crime. Stanford: Stanford University Press.

Grant, B. F., & Dawson, D. A. (1998). Age of onset of drug use and its association with DSM-IV drug abuse and dependence: Results from the National Longitudinal Alcohol

Epidemilogic Survey. *Jounal of Substance Abuse, 10*(2), 163-173. doi: 10.1016/S0899-3289(99)80131-X.

Hammink, A, & Schrijvers, C. (2013). *Klantprofielen van kwetsbare jongeren die zich melden bij het Jongerenloket.* Retrieved from http://www.ivo.nl/UserFiles/File/Publicaties/2013-01%20Klantprofielen%20kwetsbaren%20jongeren%20Jongerenloket.pdf.

Van der Laan, A. M., Blom, M. (2005). *WODC-Monitor Zelfgerapporteerde Jeugd-criminaliteit – meting 2005.* Retrieved from https://www.**wodc**.nl/images/me2006-4-volledige-tekst_tcm44-59658.pd.

Lewis, D. O., Shanok, S. S., Grant, M., & Ritvo, E. (1983). Homocidally aggressive young children: Neuropsychiatric and experimental correlates. *The American Journal of Psychiatry, 140*, 148-153.

Loeber, R., & Farrington, D. P., (2000). Young children who commit crime: Epidemiology, developmental origins, risk factors, early interventions, and policy implications. *Developmental and Psychopathology, 12*, 737-762.

Loeber, R., Wung, P., Keenan, K., Stouthamer-Loeber, M., Kammen, W. B., & Maughan, B. (1993). Developmental pathways in diruptive child behavior. *Developmental and Psychopathology, 5*, 101-132.

Moffitt, T. E. (1993). Adolescence-limited and life-cycle-persistent antisocial behaviour: A developmental taxonomy. *Psychology Review, 100*, 674-701.

Mulder, E., Vermunt, J., Brand, E., Bullens, R. & van Marle, H. (2012). Recidivism in subgroups of serious juvenile offenders: different profiles, different risks? *Criminal Behavior and Mental Health, 22*(2), 122-135. doi: 10.1002/cbm.1819. Epub.

Noom, M. J., van der Veldt, M. C. A. E., van Houdt, M. A. T., & Slot, N. W (2009*). Profielen van delinquente jongeren en bijpassende interventies: Een onderzoek naar een betere afstemming tussen delinquente jongeren en interventies in Amsterdam.* Retrieved from https://www.piresearch.nl/files/700/DMO-OOV+Profielenonderzoek+Rapport+2009-04-02.pdf.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Moodeling: A Monte Carlo Simulation Study.

Structural Equation Modeling: A Multidisciplinary Journal, 14(4), 535-569. doi: 10.1080/10705510701575396.

Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *Jounal of Numerical Analysis, 20,* 389-404.

Osgood, D. W., Foster, E. M., & Courtney, M. E. (2010). Vulnerable Populations and the Transition to Adulthood. *The Future of Children, 20*(1), 209-229. doi: 10.1353/foc.0.0047.

Patterson, G. R., Forgatch, M. S., Yoerger, K. L., & Stoolmiller, M. (1998). Variables that initiate and maintain an early-onset trajectory for juvenile offending. Development and *Psychopathology, 10*(3), 531-547. doi: 10.1017/S0954579498001734.

Sancetta, A. (2016). Greedy algorithms for prediction. *Bernoulli, 22*(2), 1227-1277. doi: 10.3150/14-BEJ691.

Spierings, F., Tudjman, T., Meeuwisse, M., & Onstenk, J. (2015). *Verkenning Riscojongeren: Onderwijs, Arbeid, Zorg en Veiligheid.* Retrieved from http://www.kenniswerkplaats-rotterdamstalent.nl/wp-content/uploads/2011/07/Literatuurstudie-Risicojongeren.pdf..

Spies, H., & Davelaar, M. (2015). *Jongeren, stad en voorzieningen.* http://www.plusconfidence.nl/sites/default/files/Jongeren%20stad%20en%20voorzieningen.docx.

Stouthamer-Loeber, M., Wei, E., Loeber, R., & Masten, A. S. (2005). Desistance from persistent serious delinquency in the transition to adulthood. *Development and Psychopathology, 16,* 897-918. doi: 10.10170S095457940404006.

Tibshirani, R., (1996). Regression and Selection via the Lasso. *Journal of the Royal Statistical Society, 58*(1), 267-288.

Wu., T. T., & Lange, K. (2008). Coordinate Descent Algorithms for Lasso Penalized Regression. *The Annals of Applied Statistics, 2*(1), 224-244.

Wurpts, I. C., & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. *Frontiers in Psychology, 5,* 920. doi: 10.3389/fpsyg.2014.00920.

Ziegler, A., & König, I. R (2014). Mining data with random forests: current options for real-world applications. *WIREs Data Mining Knowledge Discovery, 4*, 55-63. doi: 10.1002/widm.1114.