COMPARISON AND EVALUATION OF METHODS FOR HANDLING DATA CLUSTERS IN REGRESSION MODELS

Pakistan Institute of Development Economics

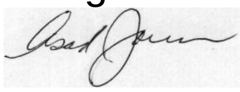
By

Saima Ishaque
PIDE2016FPHDETS07

Supervised by
Dr. Ahsan-ul-Haq

Co-Supervised by
Dr. Asad Zaman

Approved for Sending to External Examiners:



School of Economics
Pakistan Institute of Development Economics, Islamabad

2023

Author's Declaration

I 'SAIMA ISHAQ' hereby state that my PhD thesis titled "comparison and evaluation of methods for handling data clusters in regression models" is my own work and has not been submitted previously by me for taking any degree from Pakistan Institute of Development Economics or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduation the university has the right to withdraw my PhD degree.

Date :____________________

Saima

Signature of Student

Name of Student

Plagiarism Undertaking

I solemnly declare that research work presented in the thesis titled **"Comparison and Evaluation of Methods for Handling DATA Clusters in Regression Models"** is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the **HEC** and **Pakistan Institute of Development Economics, Islamabad** towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD degree, the University reserves the rights to withdraw/revoke my PhD degree and that HEC and the University has the right to publish my name on the HEC/University Website on which names of students are placed who submitted plagiarized thesis.

Students/Author Signature:________________

Name: Saima Ishaque

PIDE2016FPHDETS07

Certificate of Approval

This is to certify that the research work presented in this thesis, entitled: **"Comparison and Evaluation of Methods for Handling DATA Clusters in Regression Models"** was conducted by Ms. Saima Ishaque under the supervision of Dr. Ahsan ul Haq & Dr. Asad Zaman. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Econometrics from **Pakistan Institute of Development Economics, Islamabad.**

Signature: ____________

Student Name: Ms. Saima Ishaque
PIDE2016FPHDETS07

Examination Committee:

a) **External Examiner: Dr. Mudassar Rashid Khawaja** Signature: ____________
Associate Professor
Department of Economics
COMSATS, Islamabad

b) **Internal Examiner: Dr. Iftikhar Hussain Adil** Signature: ____________
Assistant Professor
NUST, Islamabad

Signature: ____________

Supervisor: Dr. Ahsan ul Haq
Assistant Professor
PIDE, Islamabad

Signature: ____________

Co-Supervisor: Dr. Asad Zaman
Ex-Vice Chancellor
PIDE, Islamabad

Signature: ____________

Dr. Shujaat Farooq
Head, Department of Economics & Econometrics
PIDE, Islamabad

ACKNOWLEDGEMENTS

Doing PhD Econometrics for me is not just degree acquisition rather it will stay will me as a life learning experience. The most awaited day of declaring acknowledgements makes me hearten thankful to my “ALMIGHTY ALLAH’’ for granting me the opportunity to accomplish this task. I am debited to all who have supported me on this path and to those who left me alone for making myself ready to work as independent researcher. I owe many thanks to all of my teachers including Dr. Amina, Dr. Saud, Dr. Attique, Dr. Hafsa and Dr. Zahid for being there to guide and support in many situations when I lacked the courage.

Lastly, I mention my gratitude to my supervisors Dr. Asad Zaman and Dr. Ahsan-ul-Haq for being encouraging, supportive and bearing with my impatience. It was impossible for me to accomplish my work if they would not be there to guide.

ABSTRACT

Mixture models and their variants are widely used in various disciplines. In this thesis we have considered four families of mixture models to address some modeling considerations through empirical applications. We employed a step-by-step model selection process to address the potential usefulness of the taken approaches as well as we incorporated auxiliary variables for prediction purpose. We have incorporated various notorious issues relevant to mixture modeling in the specific empirical setup including dependencies of observations, dependence between covariates and indictors, sparse data and model selection. We applied step 1 and step 3 approaches for separate discussion of distal outcomes and covariates inclusion in modeling setup of latent class cluster model, regression mixtures, growth mixtures and Markov models. Uni and multivariate mixed mode data are employed. Unconditional and conditional models are estimated and compared through a model evaluation kit consisting of absolute, relative, and bootstrapping based criteria. In latent class cluster models job quality typology is searched and compared for basic unconditional, for direct effects case and for continuous factor versions of latent class models. The explored job quality describes four clusters in terms of job quality variations for considered sample of American workers. For latent class regression case we find non-presence of differential effects of job satisfaction predictors. For Growth mixture variants in empirical application of employment status growth patterns we have found three clusters of active, inactive and mediocre active participants over the age span of 16years.In Markov modeling setup variants further improve the model fit compared to growth models since autocorrelation, heterogeneity of data and measurement error is simultaneously addressed in this case. The transitions and switching probabilities for three clusters of employed ,unemployed and inactive are calculated and compared.

Key words: Mixed-mode data, mixture modeling, auxiliary variables, direct effects, Step-3 analysis, unobserved heterogeneity, typology building, differential effects, growth differences, transition status.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1.Introduction

Cluster analysis is a subject of interest in various domains including applied statistics, data mining to machine learning. Clustering of data is done for partitioning the sample space into sub-samples or subgroups for finding the joint distribution of few or several variables. These variables or features can be further combined for formation of new features based on sub-groups. The common features can guide further for treatment of groups conditional to group differences(Tuma & Decker, 2013). The common features or response patterns help to build typology or label the classes. This labeling is not done for sake of naming the classes rather it may serve for further empirical analysis or focused group studies. A very simple example can be taken from business studies, where consumer preference surveys are done commonly for targeting the potential customers, the group of potential customers can also be found by exploring consumer data, by grouping data patterns with the help of cluster analysis. A step forward approach in clustering leads researchers far beyond grouping of data, where they may address prediction and causal links for different issues through clustering.

Having said, in the following we brief about major techniques of data clustering with the underlying assumptions followed by the relevant concise critique of such techniques. Later we will address the chosen clustering techniques in this thesis. In its basic version, clustering of data can be done through graphical displays to explore uni or multi-model nature of the data. To further compare the graphical displays, we can apply proximity measures to quantify the distance amongst observations. There are huge number of statistical formulae in literature for handling data clusters these ways but optimal proximity measures are all times unresolved and controversial issue in applied studies. The major drawback against such techniques is the subjective choice of proximity measures leading the researchers to different clustering solutions even if applied on the same dataset. Another distinguished approach to handle data clusters is Hierarchical clustering technique; the main feature of such techniques is to perform clustering in multiple steps, where data is partitioned in agglomerative or divisive sense based on some arbitrarily chosen proximity measures. For small and inherent hierarchical data these techniques are most suitable, but these techniques are subjective and irreversible, implying the limitation of no-way back for correction or updating of groups if once done wrongly(Farrelly et al., 2017).When a particular partition is providing no underlying hierarchy, Optimization clustering techniques are an alternative clustering choice to partition the objects into a set number of clusters. These techniques utilize various numerical criterions to measure the distances between units of measurement. The concepts of homogeneity and separation are key concepts for such clustering indices. The numerical criterions for optimization are huge in number, conflicting and subjective choice of criterion is again the main issue with this family of clustering methods. Because of immense range of optimization algorithms, we find controversial and subjective choice to prefer one criterion over another. Using different optimization strategy for the data we may end-up in this case with conflicting solutions given the same data and given the same technique(Delgado, Gómez-Skarmeta, & Martín, 1997) .The above briefly noted mainstream clustering techniques are criticized mainly for being heuristic. The criticism is raised for non-consideration for data structures in the above said approaches. For the critical discussion of graphical, hierarchical, proximity measures, and optimization techniques see into(Duran & Odell, 2013).

To address such limitations an alternative family of techniques to handle data clusters is model based. These methods are based on mixture densities ,and applied in many scientific disciplines under various titles such as model based clustering methods ,latent class cluster analysis ,mixture models (Hagenaars & McCutcheon, 2002).In these methods a formal statistical model is hypothesized for the population from which we draw the sample for analysis. Depending upon the nature of data, we can take uni-or multivariate probability density functions for the subpopulations. For finding clusters in data parameters can be used further to calculate the posterior probabilities of cluster association for separating of alike items into meaningful classes whilst employing mixture densities Additionally composition as well proportion of class mix can be evaluated in such techniques (B. S. Everitt, S. Landau, M. Leese, & D. C. a. Stahl, 2011b). Also in mixtures based cluster models for enumeration of clusters and for finding most workable statistical models, objective statistical criteria do exist and support the analysis(Brian S Everitt et al., 2011a).These clustering techniques based on finite mixture densities are generic and adapted according to the nature of data in many contexts (Biernacki, Celeux, & Govaert, 2000) . As in any statistical model we can even impose restrictions on the parameters to obtain parsimonious models and there exist several tests to check the validity and performance of such exploratory models (Banfield & Raftery, 1993).

Additionally in these methods scaling or normalization of variables is not required contrary to standard cluster methods. In these models, it is comparatively easy to deal with variables of mixed mode data (different scale types) and there are more formal criteria to make decisions about the number of clusters and other model features. Being a probabilistic clustering approach, though each object is assumed to belong to one class or cluster, but the objects class membership is never certain in these models like hierarchical clustering models. Belonging to certain clusters is decided by individual's posterior class-membership probabilities which are computed from the estimated model parameters and his observed scores. This makes it possible to classify other objects belonging to the population from which the sample is taken, which is not possible with standard cluster techniques. The cluster modeling approach described above is quite general: It deals with mixed-mode data and (correlated) error structure. An important extension of this model is the inclusion of covariates to predict class membership. Conceptually, these models distinguish endogenous variables that serve as indicators of the latent variable from exogenous variables that are used to predict to which cluster an object belongs (B. Muthén & Asparouhov, 2002b).

Another edge of mixture models over other exploratory frameworks lies in extended application apart from clustering. These extensions include regression mixtures for measuring differential effects across subpopulations (Di Mari, Oberski, & Vermunt, 2016) and latent class growth analysis for measuring developments of the construct under study overtime(Daniel S Nagin & Tremblay, 2005). Another variant includes Markov/latent transition analysis for measuring transitions of belonging from one subgroup to another subgroup over time (Magidson, Vermunt, & Tran, 2009b).

The choice of above said mixture model versions are conditional to data nature. The modeling environment specific to cross sectional analysis and longitudinal data analysis poses different issues and choices to handle data for meaningful analysis based on mixture models. For employing the above said 4 main variants of mixture models firstly we are bound to align data structure in accordance with the modeling choice to be made. Since some require longitudinal data frame and some are suitable for cross sectional data for finding core objective of class differences (clusters).Secondly for utilizing these techniques conditional to data nature one equally important consideration has to be made in advance and that is theoretical validity of the model in exploratory empirical settings.This implies the model selected for cross sectional or longitudinal empirical analysis should be backed by the cross sectional or longitudinal nature of the query in theoretical sense also. Though the briefly introduced mixture model-based variants for clustering solutions are unique in various perspectives (these will be explained in further chapters).

These mixture models for cross sectional and longitudinal analysis share some common issues to be addressed like problem of class enumeration (class/cluster size), decision of adding auxiliary variables /covariates before clustering or after clustering(step 1 vs step 3 analysis discussed in next chapter).In this thesis we evaluated models with a rigorous model selection procedure for class enumeration issue as well as compared alternative approaches to include auxiliary variables in each setting of four broad mixtures models (latent class cluster ,latent class regression, growth mixtures and Markov models).Along with these methodological concerns of mixtures we have addressed here problems associated with cross sectional modeling setup for latent class models in face of failure of local independence assumption of indicators with standard latent class cluster analysis and failure of homogenous sampled population assumption in real data settings using longitudinal mixture analysis.

We claim here that utilization of the mixture models for confirmatory or exploratory purposes is far from its potential in economics as cited in (Varian, 2014)and (Athey & Imbens, 2019).In domain of economics specific to quality of employment we find two applied studies utilizing only latent class cluster analysis . First case is by Benach, Vanroelen, Vives, and De Witte (2013) who followed three steps to build the clusters for job quality using European macro data by latent class cluster analysis.They utilized the resulting typology by studying relationship of job quality to relevant auxiliary variables. They have focused on giving an overview of the influence of employment quality within the European salaried workforce on the selected outcomes. They discussed relationships between job typology and employment related outcomes and health related outcomes for different groups. The narrative of study is pure economics followed by policy implications. And the technical issues related to model selection are not taken under consideration. Second case is of Vermeylen, Wilkens, Biletta, and Fromm (2017) who addressed self-employed heterogeneity of European Labor force with the help of latent class clustering. They advocated for the standard labor status division as imprecise to explain the diversity of working conditions experienced by the self-employed. They highlighted the negligence to consider heterogeneity of self-employed leads to imprecise basis for the formation of policies aimed at improving conditions for the specific class. They further emphasized for applying cluster analysis for understanding the heterogeneity hidden in status of people performing a job. An empirical estimation based on 2015 data from the sixth European Working Conditions Survey is used to identify five clusters of job quality.

With reference to other mixture models based on latent class regression analysis/ regression mixture, growth mixtures or Markov models we have detailed the research gap in next chapter for each modeling scheme since there is much applied work in other study disciplines using these methods. Briefly the research gap stems from two perspectives, first from empirical applied case studies void in discipline of economics, second from the class or cluster formation based on more rigorous approach.

Though empirical applications of mixture models for exploratory objectives are rare in economics still the applied studies have not offered any model building process of these techniques with related and competing model extensions. The chosen four family of methods has found a growing number of important applications in many other fields of social sciences(J. D. Vermunt & Vermetten, 2004; Von Davier & Yamamoto, 2004). We could not find a single empirical application in economics exploring data by applying these major techniques of mixture models for cross sectional and longitudinal data. The application of clustering methods we will employ will help us to know the static picture of various groups as well as it will reveal dynamic aspect in terms of growth and transitions in parameters of interest. We attempted to utilize these methods for meaningful empirical findings in novel contexts. The empirical applications are carefully chosen to fil the void of microdata-based case studies for job quality and related issues, and for employment status changes in terms of growth and transitions (status change).

Further, none of the studies have applied the four modeling versions of these mixture models addressing the strategy of model evaluation in case of violation of local independence assumption with presence of heterogeneity in sampled data. In this thesis we utilized mixed mode/sparse/real-life data with extended versions of standard models of latent class cluster analysis, regression mixtures and extensions including growth mixtures and Markov models. We have employed various approaches to incorporate auxiliary information (covariates or distal outcomes) and finalized model by bootstrapping and cross validation in most data settings.

1.2.Research Questions

The research questions addressed in this thesis include:

1. Are there qualitatively distinct subgroups within data who reveal precise patterns of responses?

2. What if the indicators are not conditionally independent, how to incorporate local dependence of indicators across cross sections and across time?

3. How to choose and decide between the homogenous or heterogeneous sampled data frame?

4. Is there change between latent statuses across time? If so, how this change can be characterized? How to find best model to express an individual's particular status at a given time and the dynamic status over the time?

5. How to validate differential effects of predictors across groups for meaningful class differences?

6.Is there significant different clusters based on growth trajectories of development for the chosen longitudinal data?

7. How inclusion of auxiliary variables affects class enumeration and class prediction for cross sectional and longitudinal data?

1.3. Objectives of the Study

The objectives of the thesis are:

i. To apply various exploratory methods to find hidden structure of the chosen data, the methods are to be applied within the cross-sectional mixed mode and longitudinal categorical data settings.

ii. To compare and evaluate the performance of nested and non-nested models in each data application by a comprehensive mix of various model selection tools.

iii. To find best model by a step-by-step model building process.

iv. To explore the economic construct under study for typology building and for pointing focused policy.

v. To identify movement between the subgroups over time by applying various Markov modeling procedures.

vi. To observe change and growth in selected variables over time for various clusters.

1.4. Significance of the study

Fonseca (2013);Janssen, Walther, and Lüdeke (2012);Janssen et al. (2012) suggested that in absence of proven general theories in scientific disciplines clustering of data can help discover regularities in data which is crucial for empirically based theory building. We like to emphasize the usefulness of learning from data itself. From the methodological perspective, our analysis may differ from most of the studies in economics for being non-parametric. In this thesis the data-driven approaches will directly learn from the data as suggested by (Catalina Rubianes & Annoni, 2016).

Model based clustering, regression mixtures, growth mixtures, Markov models are four distinctive applied statistical tools in various scientific domains and for finding more about focused group analysis. Their collective application and detailed model evaluation is not practiced on synthetic data as well empirical application for real data in economics is missing. Data in economics contain more outliers and can be messy. Class-membership is never known a priori in many real data settings. Distributional assumptions of standard normal also not hold for these kinds of data naturally. The above said techniques guided us to make model selection by comparing the results in terms of convergence properties, fit and residuals, parsimony, and interpretability.

We could not find a single empirical application in economics exploring data by applying major techniques of mixture models for cross sectional and longitudinal data. The application of clustering methods we will employ will help us to know the static picture of various groups as well as it will reveal dynamic aspect in terms of growth and transitions in parameters of interest. The vagueness in said claim will be complimented by the introduction of methods and economic issue in next chapters (chapter 2 and chapter 3).

1.5. Outline of the thesis

In the following chapters we have presented firstly literature review specific to four kind of modeling schemes focusing applied side. Then methods literature review is followed by studies on approaches of operationalizing the theoretical construct of quality of employment (job quality). Next chapter discussed the formalization of methods involved in analysis. Then the succeeding chapters in sequence present applications of above said four types of mixture models. We have ended this thesis with conclusion and policy implications in last chapter.

CHAPTER 2

LITERATURE REVIEW OF METHODS

2.1. Introduction

In each study mentioned in this section we will highlight the technical aspects related to the applied studies utilizing mixture modeling variants. In first section, we have presented studies with no covariates followed in next section by studies employing techniques for incorporating covariates in the standard models. Standard models of the four modeling schemes are discussed primarily for categorical variables. Testing of violation of standard model assumptions and the methods for handling of auxiliary variables included in these models are elaborated in section of extension studies. The studies in this review are naturally divided for application in its simplest versions for cross sectional data followed by sections which expand for complex modeling families for exploratory longitudinal data analysis.

B. Muthén and B. Muthén (2000) has classified statistical approaches into variable oriented approaches and person-centered approaches. Person centered approaches contrary to its competing approach model the relationships between individuals rather than variables. This big family of models has various variants for a variety of variables and for variety of data mixtures. Here we considered mixture models to highlight the limitations and contributions made by many researchers.

This literature review is based on models for longitudinal and cross-sectional data. For cross sectional data the models cited are mainly latent class cluster analysis (model based clustering is another label for this class), latent class regression models ,and its improved variants for categorical data analysis under mixture frame work by (Agresti, 2003)and (Skrondal & Rabe-Hesketh, 2004).For longitudinal data we have presented studies which have utilized methods for finding development trajectories in the variable of interest over time or age, and which have addressed transitions from initial status of belonging one class to another class over time .Broadly the categories of methods for handling data clusters in such longitudinal developmental course are growth mixture models, Markov models and regression mixtures. One common factor amongst all these distinctive methods in both time frames is that these methods are based on at least one discrete (categorical)latent variable, which divides the population into a finite number of mutually exclusive and exhaustive latent categories (classes).

2.2. Model based Clustering Applications

In case of multivariate or mixed mode data the task to identify and define groups through going-over data again and again becomes quite cumbersome. Therefore beyond visual inspections and descriptive statistics multivariate grouping techniques such as model based cluster analysis are suggested in (Magidson & Vermunt, 2002).Model based clustering techniques with continuous latent variable or latent class cluster analysis (LCCA) with discrete latent variables groups data based on their common response patterns (Brian S Everitt et al., 2011a).The basic assumptions of model based clustering analysis based on discrete latent variables is(Sinha, Calfee, & Delucchi, 2021).

According to (Lanza, Flaherty, & Collins, 2003)**,**the role of categorical latent variable is crucial to find meaningful class divisions within response patterns. And for finding qualitative and quantitative differences in these models categorical latent variable is more justified when a phenomenon under study is inherently categorical rather than continuous. Brian S Everitt et al. (2011b) introduce the basics of this technique, they provide evidence for considering it parallel to conventional cluster analysis for being exploratory though this method is model based and allows also for the computation of fit indices for finding most parsimonious model amongst competing models. In this way, the problem becomes the choice of best model fit for the data contrary to standard optimization rule-based clustering algorithms. Another crucial advantage of latent class cluster analysis pointed by Jeroen K Vermunt, Tran, and Magidson (2008) is calculation of measurement error since the latent variable is not perfectly measured with the indicators taken to describe the latent construct. In LCCA, the inherent measurement error present in indicators is considered at allocation level of individuals to various clusters based on class membership probabilities. Variable nature is crucial for deciding about the distribution structure of the latent class models. From the generalized family of models different nested models can be derived for discrete and exponential distributions. The nature of variables lead to choice these models, for details on various possible latent class models see into (Magidson & Vermunt, 2004).For distinction and choice between continuous and discrete latent variables discussion see into (Lanza et al., 2003)and (B. Muthén & B. Muthén, 2000). Mindrila (2020b) also justifies in her article the remarkable utilization of latent class clustering techniques and the reason to pick these techniques over other competing methods.

When researchers are faced with interdependence of two or more than variables occurring in data analysis (Hagenaars & McCutcheon, 2002) explains how latent class analysis helps the researchers primarily in two ways. They explain its simplest exposition through cross-classification of suspected variables and diagnostic for detection of interdependence. The analysis presented is specific to handle the source of possible interdependence from two sources. The first cause of interdependence is hidden or latent explanatory variable causing the codependence of two or more variables and the second cause described is non-independence of variables in account of missing variables (or information). The article beautifully guides about how the preliminary diagnostics can help for understanding any concept under study. The key model building steps in model-based clustering analysis suggested are to always start with independence model which should be base line and then to improve the model by adding further classes. For comparing the nested models the improved fit in absolute and relative terms is suggested as the evaluation criteria in this article. Hagenaars and McCutcheon (2002) have further explained theoretical underpinning of such models for finding of alike individuals within individuals and elaborated its usefulness in addressing data heterogeneities. Mindrila (2020a)has discussed steps of finite mixture clustering within the context of American adolescents. She has utilized data based on stratified multistage clustering technique collected from education department. She has utilized Fourteen binary survey items for measuring bullying and cyber-bullying. Seven of these variables measured traditional bullying victimization, whilst the other seven variables measured cyber-bullying victimization. In this study, naturally traditional and non-traditional victimized classes were expected in young people considered, most parsimonious and more distinctive solution suggested 4 clusters. The no of classes was selected based on relative fit indices and entropy. The study explains the non-victimized class as being highest, more than 90 %, while between the left-over marginalized class, further variations of bullying experiences are reported. The study although does not consider the possibility of non-independence of bullying behaviors, and the most victimized class is not highlighted in terms of its background or other social orientations. Therefore, badly treated remained unexplained that could be further explained with linking to external available variables. In summary non handling of dependence between observations and role of covariates seems missing in this paper. Similarly Lanza et al. (2003) has provided various applications for health and psychological data. The analysis examples in whole book are based on local independence assumption of chosen indicators. Mainly three studies are revised in various sections of the book to explain interesting features and labeling of classes based on posterior membership probabilities. The writers highlight the significance of such methods for exploring quantitative as well as qualitative differences in any data. How and why labeling of clusters should be based on the posterior member ship probabilities, and relevant technical measures of clusters enumeration and over all fit including well separated clusters are discussed there. The book is a comprehensive guide on model building in case of cross sectional and longitudinal data. Applications in this book are specific to nominal data so nixed mode data handling is not discussed .In labor market context Clark (2005) conducted a study to explore the existence of any typologies of social contract and psychological balance, the labor market data was taken from 2 European countries whose labor conditions were aligned in terms of employee obligations and contracts. Basic latent class cluster analysis was employed to validate three social contract theory hypotheses about employee and employer psychological contract. The article mainly addressed specific subject theory in detail, whereas main limitations in addition to pointed out by researchers were non-inclusion of relevant dependencies or direct effect of residuals plus the role of covariates was also limited to step 1 analysis (see section 2.4). Hennig and Liao (2013) provide a thoughtful demonstration of the application of cluster analysis to socio-economic stratification with the help of latent class cluster analysis. Much of their paper is concerned with incorporating nominal and ordinal variables in a cluster analysis in a manner that is based on theoretical background of the specific subject under any study based on subject matter know how. Their objective was to transform chosen indicators to be optimally used in dissimilarity matrix of conventional distance-based clustering technique. Nonetheless, the main critique against their clustering philosophy is the detection of clusters even in the absence of clear clustering patterns which is explained as discrete socio-economic strata in their application. The findings of the study are sensitive to data choices about preprocessing and standardization of variable**.** McParland et al. (2014)also worked on mixed data for clustering in a socio-economic application. They applied model-based clustering on categorical observations as representative of a latent continuous variable. Then by using a mixture of factor analyzers model data was clustered in their application. They have utilized Bayesian priors for estimations of parameters and allocation of units into different clusters , instead of routine BIC suggested in literature for mixture models they have opted for mixture variant of Bayesian information criterion proposed by (Fraley & Raftery, 2007). Compared to Hennig and Liao (2013) who have used parametric bootstrap test for investigating about clustering structure in their data but they did not reported model fit for such results. In this regard, this study utilized a cross-validation approach to validate the results.

Hagenaars and McCutcheon (2002) have addressed various model-based clustering variants for continuous and discrete latent variable. They have broadly divided the approaches into traditional and non-traditional cases for finding data clusters. The traditional approach was taking latent class in its original version as it’s been done in a forementioned studies, non-traditional methods were applied to incorporate the residual dependence of observations. This study moves further from standard traditional model-based clustering approaches and offers a variety of solutions in case of violations of standard model assumptions. Two distinctive solutions offered in this article to correct for local dependencies are to include the dependence between indicators when estimating the model, and second was to include continuous factor to handle dependence of observations. They tested their offered remedies with real data settings in clustering, regression and latent factor analysis contexts. Various models were tested under these conditions and were also compared in terms of improvement for relative and absolute fit of the models. The examples justified the improvement in results by incorporating non independence of variables The issue of parsimonious model in context of sparse data was separately evaluated in this article parametric bootstrapping technique based on adjusted chi 2 for sparse data is suggested and explained. The paper offers nice introduction to major model-based clustering scenarios through simple examples. The study by(Magidson & Vermunt, 2004) is another example which incorporates violation of basic assumption of non-independence, though the data taken is of continuous nature. The writers presented two examples of continuous data: one with inclusion of active covariates in the basic model and second for dependencies of residuals between indicators and covariates in terms of their effect on model building and interpretations. The study gives another simple cause and effect case for departure from traditional model-based clustering to improved modeling options. The models though are compared only in terms of absolute fit. For further related applications see into Lanza, Collins, Lemmon, and Schafer (2007) who explained to specify a latent class cluster model for the empirical data in Psychological study. Latent class cluster analysis has been applied to explore about development in child behaviors by(Nylund, Bellmore, Nishina, & Graham, 2007) and in marketing typology building by(Wedel & Kamakura, 2000).

2.3. Mixture Models for Categorical Longitudinal data

2.3.1. Mixtures of Regression

Variants of finite mixture models for categorical and mixed mode longitudinal data are discussed in this section. These methods are exploratory for dealing with unobserved population heterogeneity. And also called random coefficients models in standard statistical books when the nature of latent variable is considered continuous for random coefficients (MacCallum & Austin, 2000); Wedel and DeSarbo (1994).

In generally Mixture of Regressions/ latent class regression (LCR) are among the most widely used approaches for dealing with heterogeneity in regression problems. For applications in closely related field to economics see market segmentation(Tuma & Decker, 2013).Latent class Regressions (LCR) models have been particularly popular in various other fields including psychology, education and criminology; for such application see into (Andrews & Currim, 2003b) and (Bierbrauer, Trück, & Weron, 2004). Technical aspects of regression mixtures are very nicely discussed in (Khalili & Chen, 2007).Wedel and Kamakura (2000) presents a nice introduction to latent class version of mixture models. They highlight the strength of regression mixtures as a special case of finite mixtures to accommodate hypothesis testing within statistical standard theory. They discuss flexibility of such models to accommodate dependent variables of scale types other than nominal or ordinal. They further describe the possibility of estimating conditional and unconditional models in their article as strength of regression mixtures. In Hartzel, Agresti, and Caffo (2001) we find non-technical introduction to mixtures of regressions and their potential for use in socials sciences. The paper presents theoretical claims of standard regression models with a list of applications scope in marketing research. Extensions of the basic structure influenced by background variables are also described with two empirical applications. One is for trade performance show, and another application for conjoint based study is reported in the study. In this paper the writers discussed different classes in terms of their different attributes and provide preference of discrete latent variable for classifying groups over the continuous latent variable in such cases. Extending to Hartzel et al. (2001) for various discrete and categorical variables Jeroen. K. Vermunt (2005) presented mixed-effects applied cases priory extensively discussed in (Skrondal & Rabe-Hesketh, 2004) and Hartzel, Agresti,&Caffo, 2001).The writers presented an extension of mixed effects logistic regression model for cases in which the dependent variable is a discrete latent variable measured with multiple indicators. The writers offered improved maximum- likelihood based solution for the multivariate case and adapted the E step of the EM algorithm making use of the conditional independence assumptions implied by the model. Then model was illustrated with an example from organizational research in which they built a latent task-variety clustering. After controlling for individual-level covariates they found for same data clear indication for between-cluster variation in the latent class distribution of different clusters.

Yamaguchi (2000) has also discussed mixed effect model for Japanese women for measuring gender biases in case of vote and support for female participation in workforce.The data is taken from the national base for measuring attitudes of Japanese people towards social norms in society. The models opted to formalize the relationship between indicators for various possible groups of samples are regression extensions of log-linear latent-class models with group variables. The application focuses on predictors of three latent classes of gender role attitudes among Japanese women. Three class regression solutions emerged in their case as the best solution, and the classes were labeled as “anti-work gender equality supporters” “traditional gender role supporters,” “pro-work gender-equality supporters,” following their probabilistic response patterns differences. Each class had different characteristics, which were explained in terms of distinctive parameters. Kim, Vermunt, Bakk, Jaki, and Van Horn (2016) have guided in their article for the model building process for including a latent class predictor in regression mixture models. For this purpose, they designed low to high separated mixtures. First, they estimated an ‘unconditional model’ which included no latent class predictors. The model was explained in terms of differences between classes in regression weights. Then the writers included latent class predictor into the model and compared whether there were any substantive differences in the model results. Contribution of predictors in class predictors was questioned by comparing different models and the difference was compared in terms of differential effects in class structure to the baseline no predictors case. Exclusion of predictors may lead to bias in class sizes was also addressed in that study. Finally, they used an applied dataset to show the effects of omitting the direct effect from the latent class predictors to an outcome variable.

Sánchez and Puente (2015a) presented the differential effects of mismatch inherent in educational and skills of workers for job satisfaction outcome. The related job quality indicators included were salary, promotion chances, number of working hours and kind of tasks performed. Using Survey of Quality of Working Life, the writers found that highly educated individuals show higher levels of dissatisfaction than those with low qualification. Separate consideration of educational and skill mismatches was emphasized in conclusion. (Sánchez & Puente, 2015b) further endorsed for Education and skill mismatches for Spanish labor market by employing regression mixtures and explained significant differential effects of relevant labor market indicators on job satisfaction.

Growth mixtures are a special version of regression mixtures where we find differential effects of time or age for bringing change in growth patterns of the sampled population in hand. Since these are specific to time or age predictor and serve additional purpose of finding change in parameters differences for different groups over time therefore this family of methods is distinctive from regression mixture. For discussion of growth mixture variants see into (Agresti, Booth, Hobert, & Caffo, 2000). Here we present some relevant case studies:

2.3.2. Growth Models

When the aim is to explore trajectories of growth amongst heterogeneous population over the time or age, we come across mainly three categories to address such issues with varying modeling assumptions. These modeling categories for finding development or change in construct under study constitute average or simple growth models, mixture growth and restricted mixture growth models i.e., latent class growth models. In present discussion we have differentiated studies accordingly .Andruff, Carraro, Thompson, Gaudreau, and Louvet (2009) and (Tony Jung & Kandauda AS Wickrama, 2008)

For confirmatory modeling perspective of growth models amongst many Skrondal and Rabe-Hesketh (2004)discussed generalized linear models as a parametric tool for quantifying differences in random samples under various modeling options of linear and non-linear time trends in longitudinal data. They discussed importance and choice of the link functions for choosing conditional distribution of data and in case of (ordered or unordered categorical data) they voted for multinomial distributions in these books. Either for discrete or continuous data they treated in the book, the techniques were standard parametric based on iid assumption for the distribution. Another confirmatory clustering technique mentioned in Bryk and Raudenbush (1992) offers robust sandwich estimators for confirmatory hypotheses testing in case of possible violation of conditionally independent observations in categorical repeated measurements over time. The methods which utilize the inherent nature of categorical variables keeping time domain inclusive and explicitly addressing individual difference in confirmatory settings are discussed by (Bryk & Raudenbush, 1992) and

(Skrondal & Rabe-Hesketh, 2004). These methods are called random-effects models or standard growth models by (Jeroen K Vermunt, 2017); (D. Nagin, 2009).In a special setting Jeroen K Vermunt and Liesbet Van Dijk (2001)have measured individual differences in the outcome variable when time equals zero and change in the effect over time are modeled by allowing the intercept and slope coefficients to vary across entities. These orthodox growth modeling approaches support for an average growth estimate and a single estimate of variance of the growth parameters, the effect of covariates on the variance and growth parameters is also assumed same across parameters in the study. For various applications of these methods look into (Agresti et al., 2000);(B. Muthén & Asparouhov, 2002b).

According to D. Hedeker and R. J. Mermelstein (1998)the standard growth models are extremely valued tools for handling with longitudinal data. Though unobserved heterogeneity is considered in these methods, but the modeling framework is limited on the untestable conjecture of a multivariate normal distribution for random coefficients. In a study by Jeroen K Vermunt (2017) we find a supportive review of cases lacking normal distributional assumption leading to biased results under standard growth models framework.

As an alternative to standard growth models D. Nagin (2009) proposed homogeneous individual growth trajectories within a class for latent class growth models, in growth mixture modeling approach by B. Muthén and L. Muthén (2000) we find general growth mixture modeling framework where no restriction on nature of growth trajectory is suggested for each subgroup .In their paper the writers have introduced us to the comparisons of standard growth curve modeling which is multilevel and random effects based. They provide a comprehensive for the infeasibility of homogenous individual growth approach for more real situations occurring in social as well as health sciences. They further present the alternative methodologies of Latent class growth and Growth mixture models. The addition of issues faced in such models and their possible solutions are also contributed to the same article**.**

According to J. Reinecke and D. Seddig (2011) the Growth mixture model appear to be the most adjustable approach for incorporating inter individual differences in intra-individual change considering unobserved heterogeneity within a larger population. Muthèn (2004) has conducted an extensive study for explaining theoretical differences amongst various growth modeling options given real life data applications. The writers have drawn a clear distinction between conventional growth models and growth mixtures, they have discussed the extensions with the hybrid models of discrete and continuous latent variables following (Jeroen K Vermunt, 2017). The incorporation of auxiliary information of distal outcomes in educational setting is also provided in the paper. The part which interests more to be discussed with reference to present work is the applied example of latent class growth mixture modeling as a restricted variant of more general framework growth mixtures for categorical data proposed by (B. O. Muthén, 2002). The data taken for elaboration of comparative performance of both models is used also by Daniel S Nagin (1999)to study delinquency behavior of individuals over age 10 to 32. Both models were tested on the same data, and 3 class mixture solution was decided on the base of relative information criteria.

Daniel S Nagin (1999) proposed using the mixture growth model for the analysis of developmental trajectories with the purpose to identify distinctive groups of individual trajectories by means of continuous latent variables. The writer also explained to work discrete latent variables for classifying distinct growth trajectories. This case becomes a special variant of growth models referred as latent-class growth model where purpose of the exploratory technique is to identify subgroups or clusters showing different developmental patterns or trajectories over the time or age. They also explained the background theory and introduced two conceptualizations of latent class growth/trajectory models in case of approximations of a continuous but unknown distribution of population heterogeneity, and for concrete trajectories that can be treated as substantively important entities. Jeroen K Vermunt (2017) has used both techniques growth mixtures and latent class growth mixtures in the same model for chosen case study of schizophrenia incidence measured at consecutive weeks. The writer included continuous random effects within discrete latent class growth model latent class to incorporate unobserved heterogeneity within the given sample. Another example by Verbeke, Fieuws, Molenberghs, and Davidian (2014) is Latent class growth model in multilevel framework. The defining assumption of this data setting is the distinctive parameters related to nested observations with reference to latent classes and for within classes different growth patterns . These classes are allowed to differ with respect to the growth class distribution of their members. The article explains about the class-specific growth curves based on mixture components or number of classes.

Daniel S Nagin and Tremblay (2001) explained how Latent version of growth analysis differs from standard growth analysis or random coefficients approach. They described that contrary to standard latent growth modeling techniques in which individual differences in both the slope and intercept are estimated as average estimates by employing random coefficients. The proposed Latent class growth model imposes equality across individuals within a class by estimating class slope and class parameter of change. The edge of imposing this restriction is remarked for segregating individual changes from multiple trajectories (changes). Comparisons of the distribution of individual differences in change within the data by a finite set of polynomial functions relative to a single distribution of change are discussed also in (Daniel S Nagin & Tremblay, 2005).Extension of such models for multivariate data is provided in(Daniel S Nagin, Jones, Passos, & Tremblay, 2018). The article guides about how to utilize the information within multivariate longitudinal data for finding the path of change in outcome of interest. In this application the writers demonstrated for latent clusters of individuals following similar change patterns across many symptoms of chronic disease measured by their hemoglobin, blood and CO2 levels. The unique likelihood function of the multi-level model was demonstrated with few related examples.The study also cites important studies to differentiate this class and its specifications from standard growth models. Van De Schoot, Sijbrandij, Winter, Depaoli, and Vermunt (2017) present a comprehensive guide for those who are conducting latent growth or related analysis, the paper highlights the importance of various steps to make sure of validity of analysis. Helbling and Kanji (2018)estimated the growth models in two steps. They estimated an unconditional growth curve model without explanatory variables to assess the average trajectory of life satisfaction of young workers in Germany between 2010 and 2014. In this example the latent intercept represented average initial life satisfaction at base year, and the latent slope parameter represented average rates of change in life satisfaction over the total span of four years period. The variances of the latent growth parameters indicate whether there is inter-individual variation in initial life satisfaction and in the rate of change of life satisfaction across young workers related to contractual insecurity defined in subjective and objective terms. Latent growth curve modeling in this study is used to distinguish individual wellbeing trajectories in relation to job insecurity over a period of five years. Further for studies on class enumeration issue in growth mixtures see into (Tony Jung & Kandauda AS Wickrama, 2008; Daniel S Nagin & Tremblay, 2001; Karen L Nylund, Tihomir Asparouhov, & Bengt O Muthén, 2007; K. L. Nylund, T. Asparouhov, & B.O. Muthén, 2007).

In the following we present literature on those specific methods where individuals can be followed for their initial class membership to shifted membership. These are special variant of mixture models, where dynamic aspects of data clustering are measured and tested.

2.3.3. Markov Models

Lanza et al. (2003)have discussed latent Markov models for adolescents who had participated in the Add Health study. It was a 2-wave study for same individuals in high school start years and college start year, so time gap was 1 to 2 years varying between individuals. Explicitly, students were explored in terms of their engagement to normative and non-normative behaviors (like lying, acting loud or stealing) over the taken time period. Latent transition model was applied to answer additional questions followed to simple latent class cluster analysis. These additional questions were set in terms of development of youngsters' behaviors for negative attitudes over the time. Though time span was not taken to be long but still writers expected possible presence of some individuals who have changed their belonging to different latent groups over the time. 5 class solution was found to be most parsimonious based on relative info criteria such as AIC, BIC. Following the same lines, a somewhat different study was conducted by Collins and Lanza (2009) for investigating about the differences in emergence and sustainability of depression symptoms in young individuals over the time, the construct was built in similar fashion but for different sample of data in the last 2 high school years. Further extensions with inclusion of covariates, and other related issues of latent transition models are explained in (Auerbach & Collins, 2006a).These studies are specific and more applied in psychological or physical health transitions contexts(Bartolucci, Farcomeni, & Pennoni, 2019), Lanza, Bray, and Collins (2013)

The Markov models are applied in mainly domains of psychology , criminology , political science and other life sciences (Langeheine, Stern, & Van de Pol, 1994), (K. Nylund et al., 2007), (Dayton, Counseling, & Development, 1991).We can find big list of similar applications in (Lanza et al., 2003),(Collins & Lanza, 2009). One of The limitations of all above mentioned studies is that these studies either have not included auxiliary information in predicting class memberships or movement between class membership of individuals, and if have included then the relative closely related alternatives of step 1 approach is not taken into account. Also, these studies are based on the strong 1st order Markov assumption implying data contains no memory which implies second or higher order Markov chains possibility is not even tested or reported. For the said case if we agree with the imposed assumption of this most restrictive case still we need to justify the independence of observations over time through some common measures of dependence structures like longitudinal bivariate residuals proposed by (Jeroen K Vermunt & Magidson, 2013).As explained by Jeroen K Vermunt and Magidson (2013) if the model of 1st order Markov process holds on any data then the dependence structure of observations should show values less than 3 when measured in terms of bivariate residuals. Leonard J Paas, Jeroen K Vermunt, and Tammo HA Bijmolt (2007b) applied solely latent Markov models (with no competing alternative models) justified the validity of financial multivariate financial data by using longitudinal bivariate residual diagnostics. They tested the 1st order Markov assumption and justified theoretically the order of Markov chains. The study is one of its natures and one of very few applications of latent Markov models in finding possible classes of households in terms of various financial portfolios. Another work is the application in measurement/ prediction of future employment status by Magidson, Vermunt, and Tran (2009a) using CPS data over the years. The writers have offered this modeling as a cost-effective strategy for case of measuring economic variables where measurement error is always expected to be present for various technical and non-technical reasons. The models used are general version of mixed Markov with its special variants of latent Markov and mover stayers. For the specific data, mover stayer version turns to be the best where one class of people who change their group membership over time whilst others remain stayers and do no change over time their employment status. The covariates or predictors of class transitions are not discussed and three employment status categories of employed, unemployed and not in labor force are considered. The study claims the potential of Markov models for finding about hidden economic constructs.

F. Van de Pol and R. Langeheine (1990)contribute to earlier reviews on modern day Markov chains models. The writer has discussed various extensions of Markov models including manifest and latent Markov models. They also briefly treated restricted cases such as the single Markov chain, the mover-stayer model, with other extensions for multivariate latent class models. The inclusion of multiple group factor and invariance of parameters is also tested and the basic theoretical set up for applied restriction variants of such models is presented with real life data application. From methodological perspective the paper is an introduction to basic frame works of Markov as well as latent class Markov models.Though these kind of models are not the typical ones to be used for clustering of data but when the transition framework or the change from initial stage to latter stage brings the membership probabilities of the entities over the time then we may get a comparative pic of groups or clusters over time. The nested models with various assumption of being homogenous, stayer, and random class were simulated and were compared being nested models in terms of chi square based likelihood ratio statistic.

Acconcia, Carannante, Misuraca, and Scepi (2020) applied one variant of Markov models i.e., latent transition model in context of some economic research question. The objective of the case study is to validate the theory about interlink of poverty to vulnerability of households. Households are nested within, and the data is adjusted for rotating panel analysis over rolling years to include same individuals in the analysis. First order Markov assumption in this specific context may not be justified, since household standard of living is built upon long interdependence of past Markov chains. The reason might be data limitation, still with given limited 3 years data, other variants of Markov models which incorporate possibilities of randomness, stayer, and heterogeneous moves could be fruitful. In the specific scenario homogenous transitions could be well justified to confirm the performance with unconditional latent transition models. The analysis was conducted for multiple groups by cross comparison of gender and social class biases. The models were selected by likelihood ratio test, the cross comparison of latent and Markov models revealed significant changed pattern of transitions from the initial classes to the final class.

Kaplan (2008) presents an overview of quantitative methodologies for the study of stage-sequential development based on extensions of Markov chain modeling. Four methods are presented that exemplify the flexibility of this approach: the manifest Markov model, the latent Markov model, latent transition analysis, and the mixture latent Markov model. A special case of the mixture latent Markov model, the so-called mover–stayer model, is used in this study. Unconditional and conditional models are estimated for the manifest Markov model and the latent Markov model, where the conditional models include a measure of poverty status. Issues of model specification, estimation, and testing are briefly discussed. Jeroen K Vermunt, Langeheine, and Bockenholt (1999) has discussed three distinct sub-modules of Markov/transition models for longitudinal data. Mixture latent Markov model and its further extensions are highlighted for differences and situations regarding model choices in various scenarios. The paper is nontechnical and describes an application of the transition models using national youth measurement survey.The data studied is of unequal measurement occasions with the age gaps of 11 to 19 between individuals. The writer has addressed the most likely issue present in longitudinal data. The issue of heterogeneity, measurement error, and autocorrelation are explained effectively with the most general framework of mixture latent Markov models. The other paper by Crayen, Eid, Lischetzke, Courvoisier, and Vermunt (2012) also confirm mixture latent Markov models superiority over other competing models for the data utilized. The nested models were employed to address auto-correlation and heterogeneity separately. The nested models and mixture latent performed relatively poor for the data considered in this case. The writers justified to employ the most general mixture version since it can incorporate each of the issues present in longitudinal data .For mixture latent Markov models we find another evidence of support in totally different study design by Crayen et al. (2012).In this study the writers have applied the model to study mood regulation of individuals on daily basis. The extent of measurement error and presence of heterogeneity is tested with mixture latent Markov model with few categorical items. The study has identified two latent groups who strive to maintain their mood, and cope with fluctuations on daily basis data. Amongst the two identified clusters, the larger group showed the tendency of remaining somehow successful with their mood fluctuations. The individuals if experienced bad mood during the day, then they managed to transition from bad to good mood state moderately, compared to other class who were more persistent to move and stay in pleasant mood. The models described guide well for measuring the existence of inter-individual differences by Markov variants. Jeroen K Vermunt et al. (2008) compares relative performance of most general Markov model “mixture latent Markov models”, and its nested models. The writer presents methodological differences between the most general techniques to its sub variants, the differences are highlighted mainly in the capacity of these Markov/transition or growth models for handling longitudinal data issues. The addressed issues in this study are measurement error, heterogeneity and autocorrelation with an application of empirical data. The outputs of all models also differ according to the structure and assumptions of the models, whereas the most general case provides maximum information consisting of initial-state probabilities, transition-probabilities, measurement error probabilities, and mixture proportions. Magidson et al. (2009b)also searched for the most parsimonious model amongst competing Markov models to address transition and measurement error in CPS rotational panel. The economic variable labor market status of being employed, unemployed or not in labor force over time is measured and forecasted for the individuals over time. The simulations were performed to confirm the hypothesis of assumed homogenous transition between groups and its aftereffect on the extent of measurement error present in the model. The results provided positive evidence of negative relationship between these two inherited data issues, Also For CPS data mixture latent Markov model with two chains of Markov process, standard latent transition model (based on 1 chain autocorrelation) and a special variant “mover-stayer model” were tested. Ryoo, Wang, Swearer, Hull, and Shi (2018) studied student-centered concerns for bullying by offering model building process employing latent transition analysis in this regard. Specifically, verbal bullying and social exclusion was found as the major concerns by the major profiled group. There was no profile detected that endorsed physical bullying concerns without endorsing other types of bullying. The study does not reports any framework for dealing on an unconditional model when adding covariates and/or distal variables introduced for Markov models by (Asparouhov & Muthén, 2014).Lastly, Lo-Mendell-Rubin (LMR) test by Jeroen K Vermunt et al. (1999) or the bootstrap likelihood ratio test (BLRT) by Nylund-Gibson, Grimm, and Masyn (2019a) for model evaluations are not discussed to make final model choices. Similarly another study by Jeroen K Vermunt et al. (2008) applied bottom up modeling strategy through the presentation of most general modeling structure for locating group structures and dynamics within these constructs. They addressed comprehensively solutions to issues emerging in longitudinal data such as codependence, measurement error and spurious change. They proposed mixed version of latent Markov models as an expanded version with covariates of the mixed Markov latent class model by(F. Van de Pol & R. Langeheine, 1990).For further applications in various contexts see in(Bartolucci, Farcomeni, & Pennoni, 2014; Bartolucci et al., 2019);(Leonard J Paas, Jeroen K Vermunt, & Tammo H Bijmolt, 2007a) (Van De Schoot et al., 2017) For State selection (Bacci, Pandolfi, Pennoni, & Classification, 2014).For various applications see (Auerbach & Collins, 2006b)who applied LTA to data on alcohol use during the transition to adulthood.

2.4. Step 3Extensions

In the above review of methods for cross-sectional and longitudinal data we presented studies on subfamilies of finite mixture modeling techniques including latent class cluster analysis, mixture regressions analysis, growth analysis and Markov/transition analysis. In the following we present some studies on extensions of these models for one important issue of incorporating auxiliary variables.

Role of concomitant or auxiliary variables is well discussed in various academic model based clustering applications (Brian S Everitt et al., 2011a; Verbeke et al., 2014)**.**The auxiliary variables also said independent, covariates or external expand the basic classified model for other purposes such as prediction or structural modeling of class assignment based on model or proportional assignment (Jeroen K Vermunt, 2010).They can add reason and logic to emergent different data patterns by relating them to back ground variables. The strategies opted to handle auxiliary variables in model-based clustering literature are mainly subdivided on steps of analysis. In one-step analysis measurement model of interest with a logistic regression model are simultaneously estimated by including auxiliary variables in estimation. Here the latent classes are subjected to a set of covariates or distal variables and class formation is impacted by the external variables. Whereas in step 3 approaches, we estimate model for classification with no external variables, and obtain the classification posterior probabilities of the indicators. Then these probabilities are regressed on the chosen external variables called covariates or distal outcomes in some situations. The detailed procedure of step3 classification and for the listing of possible models available for distal outcomes see,(Nylund-Gibson, Grimm, & Masyn, 2019b).

Bolck, Croon, and Hagenaars (2004)demonstrated that three-step approaches underestimate the relationships between covariates and class membership irrespective of whether we use proportional assignment or modal assignment of cases in classes. The drawback in using such approach is the chance of sustained classification error encountered of the second step leading down-ward bias in the parameter estimates. Inspired from these results Bolck et al. (2004) offered a method for correcting the three-step approach (BCH approach). Limitations associated with their approach are categorical covariates requirement for summarizing data nicely in frequency tables and cumbersome matrix multiplications requirement each time when a new set of covariates is selected. Jeroen K . Vermunt (2010)proposed a modified BCH procedure which could cover many technical difficulties faced by stand BCH, AND a new more efficient maximum likelihood based step3 method, they discussed the technical differences in each so call step-3 approach for incorporating covariates for categorical variables also they conducted s simulation study for comparing relative performance of standard one class , and variants of step 3 methods in terms of relative bias and standard errors, the scenarios were set for low class separation to high class separation, for small data to considerably big data , and the judgment of performance was based on entropy based R2 measures. Both correction methods performed very well for parameter estimates and their SEs could be trusted in given simulation set up. Jeroen K Vermunt and Magidson (2021) has discussed a novel case occurring when working with predictors or exogenous variables in model-based clustering scenarios. They raise and offer a solution for non-invariance of measurements across groups which causes step 3 analysis inapplicable. Asparouhov and Muthén (2014) has extensively discussed the application of 3-step method for explanatory variables in several different settings of model based clusters, these included basic version of latent class analysis, followed by latent transition analysis, and growth mixture modeling. The above-mentioned studies have not incorporated failure of classical assumption of local independence explicitly for such a variety of mixture models. This study reveals various cases for comparing the performance of step 3 ml method by Jeroen K Vermunt (2010) and standard 1 step approach in terms of coverage and bias. In all cases, step 3 ml outperforms for data sizes from as low 200 to as high 2000 and from low repeated to well separate latent class measurement model. The study describes software implementation of study designs for measuring the same issue. The situations are created for latent transition analysis, same with covariates , the LCCA relating to outcome variable and the case of omitted direct variables effect on prediction is measured for auxiliary variables .On the whole, this study involves various possible situations to incorporate various data settings to measure the usefulness of step 3 approaches proposed by (Jeroen K Vermunt, 2010).

J. Reinecke and D. Seddig (2011) raise concerns for finding most parsimonious models in presence of covariates since specific measurement models is not in researcher’s knowledge prior to analysis. In practice, and fitting of these models with covariates becomes infeasible in many cases . Paas et al. (2007a) discuss this issue for Markov models where choice of functional form as well variable selection in modeling is discussed. According to them, addition of covariate at first step further increases model selection within these models a daunting task. To evade these issues, the writers support for analyzing measurement model with no covariates at first step and to subsequently include the covariates later. Recently, Bartolucci, Montanari, and Pandolfi (2015)proposed stepwise analysis for covariate effects in case of Latent Markov models. A Latent Class model on the pooled data was estimated in the first step giving resultant posterior probabilities in second step of estimations whereas Markov model for transitions between adjacent time points was estimated in the third step. Their simulation study showed that this approach works well only when class separation in the first step is close to perfect. The generalization of ML type step 3 correction for latent Markov models with covariates proposed by Di Mari et al. (2016)also proved to be promising where the class separation of the cases was not very low. The three-step method with ML correction was compared with full-information ML (FIML), unadjusted three-step, and three-step with ML correction with the use of time dummies in the first step. The target measures used for the comparisons were bias, standard errors, standard deviations, relative efficiency and coverage rates. A method performs well when it yields unbiased estimates with small variability, and whenever such variability is correctly retrieved by the standard error estimates. The corrected three-step methods’ estimates are in line with FIML, except for the combination of small sample size and moderate class separation. The empirical study on financial bonds acquisition by households was tested for explaining the transitions influenced by the primary covariates such as age, household income and household size. The results were aligned to portfolio theories of asset acquisitions mention in text.

Di Mari and Bakk (2018) has extended the similar case including the direct effects of indicators. They proposed the modified approach to address the above said issue in presence of non-independence between indicators; they tested on real data from the panel version of the same GSS survey including measurements from three years: 2010, 2012 and 2014, with a sample size of 2044, 1551 and 1304 respectively. The approach proposed was for direct effect modeling in a latent Markov context. They selected six items that measured whether the respondents would allow members of different out-groups to speak in a public space. The first class, labeled 'Intolerant', has a low probability to give a positive answer on all items, while the second class has a high probability to give a positive answer on all items. The last class, labeled 'Middle', has a low probability to allow to speak for Muslims and racists, while being more tolerant toward the other groups. Zsuzsa Bakk, Fetene B Tekle, and Jeroen K Vermunt (2013a) extends in case of distal outcomes of latent class membership.They tested for two empirical applications in their paper by simulations for available methods of step 3 various versions. Results were compared in terms of standard error and the study voted for step 3 better performance based on BCH corrections also explained in (Nylund-Gibson, Grimm, Quirk, & Furlong, 2014).

2.5. Conclusion for Relevance

Above we have presented literature review for 2 broad categories of cross sectional and longitudinal data in terms of four versions of mixture models followed by the model extensions for incorporating auxiliary variables in each case of four versions: latent class clustering models, regression mixtures, growth models and Markov models. The cited literature on the extension part had few mentioned studies for being rarely applied in context of close related social science discipline. The studies addressing economic issue of quality of employment or related concept using this methodology are rarely found. Also applied empirical economic application utilizing regression mixtures, growth models and Markov models is missing in academic literature. In our work we will be incorporating various technical issues including Handling of mixed mode sparse data, dependence of observations and predictors, contribution of auxiliary variables through Step 1 and step 3 analyses for covariates and for distal outcomes. Evaluation and validation of results will be done through a rigorous mix of formal model evaluation criteria discussed in second section of chapter 4 (methodology).

CHAPTER 3

LITERATURE REVIEW OF THEORY

The review under discussion is attempted to mark research gaps with two perspectives; first is to discuss various theoretical frameworks adopted for elaborating the concept of quality of employment /job quality (used interchangeably) in literature. And secondly for developing theoretical framework for empirical application of job quality indicators applying model-based clustering).

3.1. Introduction

The traditional case of an all-lifelong quality work becomes more and more constricted through-out the globe for various kinds of economic, social, demographic and institutional shocks. The crucial task of providing quality jobs for all is by no means beyond the capacity of even developed nations. In such times concept of quality of employment have become conditional to nature of jobs having either formal or informal work arrangements. Garzón-Duque, Cardona-Arango, Rodríguez-Ospina, and Segura-Cardona (2017) discussed the scattered and confused use of the concept of employment quality. They highlighted the contextual differences based on country differences conditional to agencies involved in measurement of job quality. Further they discuss many international reports confirming common features of bad quality of job as according 'with low education levels', and scarce economic gains (job benefits) out of work.

The controversial concept can be measured form two bold perspectives further, within the formal set up of economy or within the informal set up. When measuring this concept for formal workers, the criteria of measurement would follow the working standard set for formal economy and on the contrary for informal economy we must reinvent and adjust the scope and definitions of job / work quality given the scenario. There could also be various perspectives to study Labor markets from macro to meso to micro level. To explore the concept further from conceptual differences the terms emerge in literature are of atypical, nonstandard, precarious, vulnerable and bad jobs.

Since the literature is vast and diversified. First, we present the literature supporting the diversified and contextual nature of job quality. The closely related concepts like quality of employment or work are also touched upon since job quality is the sub dimension of such concepts. The concept of job quality is broad, dynamic, contextual and multidimensional (Kalleberg, 2011). Concept of job quality (quality of employment)is discussed in literature with reference to its dimensions , outcomes and also in accordance with factors affecting job quality (Nesterenko, 2011). The concept is triggered by multiple forces at multiple levels of its operation. We are giving an overview of approaches which describe or measure various faces of quality of employment/jobs broadly including vulnerable or precarious employment as sub-faces. It is important to emphasize that Labor markets evidently differ across the developed and developing world. So, the discussed examples for developed countries may not be rationally compared to case studies of developing nations in terms of quality of employment. This leads a valid reason for not inferring one study conclusions to different context and urges for exploring the concept at case study basis.

For that we seek to investigate literature from two angles simultaneously; first is how the study under discussion operationalizes the concept and how the concept is measured in same study. Broad distinction of studies can be made from measurement perspective, confirmatory and exploratory measurement. In confirmatory analysis of quality of employment (QOE) further distinction can be made of direct and indirect approaches for measuring the construct of QOE. Direct approaches are specific to address theoretical model of quality of employment, and indirect approaches have far wider scope and may origin from different study domains. Unanimously both direct and indirectly operationalizing approaches rely on confirmatory statistical tools for measuring the quality construct**.** The alternative exploratory approaches of measurement include clustering techniques application. The studies under these divisions are presented below:

3.2. Direct Approaches

For quality of employment literature, we find that multiple and relatively diffuse concepts have developed over the past decades because of innovations and widening gap between economic conditions and economic prioritized targets across countries. First, we highlight to some extent that the terminology to describe QOE is confusing. Expressions such as ‘quality of working life’, ‘job quality’ ‘quality of employment’ are often used interchangeably and without clear definitions. This reflects the complication involved in its construct building in specific set up. we must consider manifold sides of jobs. Impacts of job quality can also be addressed in macro and micro context. Though in any case the narrative is to look about improvement in individual and overall well-being of the society, still micro macro is the broader distinction to conceptualize the idea. In this review after introducing a little about macro view of QOE, we will be focused on different debates in measurement of QOE at individual level.

We again emphasize before proceeding with literature review that Job quality/ QOE is contextual so economies around the globe from developing to developed are constrained to target it differently. For example for developing country like Pakistan, covering widened wage gap, and provision of health insurance for all working class can be a big challenge whereas for developed country assurance of good jobs can be targeted to improve overall wellbeing of individuals rather than to merely assurance of capacity to afford all basic needs with wages .

To understand the varying aspects of this construct study by Burchell, Sehnbruch, Piasna, and Agloni (2014) is a good start. The writers present excellent debate regarding the theoretical and methodological divided schools of thought for job quality concept. The article describes evolution of the concept from wage to non -wage benefits and from well- being dimensions to working conditions diversity to enforce it in a company and at national level. They brief about developments in its operationalization and measurement in other cross related disciplines. The article presents debates centering quality of employment and related concepts such as decent work and to mention the causes for the relative performance of QOE at more broad levels compared to decent work concept. They have presented academic and institutional literature scope for conceptual developments of the concepts over time. The writers concluded that because of comparable indicators from harmonized employment surveys in Europe there exists a strong bases for the analysis and measurement of QOE. Similarly the survey by Parent-Thirion et al. takes a broad and comprehensive narrative for measuring job quality for 28 European countries .This report takes job quality indices developed by European union for further profiling workers at sectorial level , organizational level and on country levels as well for European countries. The writers applied mixture modeling based on 7 indices of quality jobs and working conditions of European labor class, and five distinct clusters are found by latent class cluster analysis, where each cluster shares scores of individuals similar for each quality index. Centering on job quality indices calculated for the groups similar jobs are assigned to the same class and markedly different jobs are classified in different classes. Step 3 analyses are done for finding possible associations between distinct clusters and some outcome variables of quality jobs.

Rothwell and Crabtree (2019) conducted study about job quality situation in USA. Results for this study are based on mail surveys conducted February 2019, with a random sample of adults aged 18 and older living in all 50 states and the District of Columbia. Working adults were included in the analysis. job quality is taken as the weighted average of satisfaction on 10 dimensions of job attributes by an individual, on a one to five scale. Scores below three on the combined index indicate a "bad job." Score between three and four are "mediocre," and a score of four or above indicates a "good job." Overall, compelling evidence was found for quality differences of job situations that scored bad job compared to other groups. Hofmans, De Gieter, and Pepermans (2013)have used cross sectional and panel data to trace changes in job values and job outcomes framework developed for OECD countries. By exploiting various data source. The subject discussed is about the direct relationship between job satisfaction and job outcomes .The indictors used for measuring job outcomes are pay , hours of work, promotion, psychological and physical pressure of job, job content and social relationships at workplace. Analysis is done through regression analysis and gendered comparisons are made. Over the different aspects of the job consisting of pay, promotion, and security job Satisfaction is measured as the weighted sum of Job Outcomes. The writers concluded raised inequality between some measures of job content over the time. Similarly Erhel and Guergoat-Larivière (2010b) have applied multivariate approach to measure job quality based partially on Laeken indicators (Atkinson, Marlier, & Nolan, 2004). The dimensions defined for measuring job quality are socio- working conditions, economic security, skills and training opportunities and ability to combine work and family life. The study gives evidence for connection between labour market outcomes and job quality. Gendered differences are checked and measured, and heterogeneity of indicators between individuals is confirmed**.** Four clusters of European Union are revealed and discussed in terms of their relative performance across quality indicators, correlation between these quality indicators and employment is revealed to be strong across Europe in this study. The writers have inquired the differences in quality performance for gender differences, age differences and occupation groups across region as well. They highlighted inter- individual heterogeneity, through principal component analysis and classification.

Davoine, Erhel, and Guergoat-Larivière (2008) presented a detailed statistical framework for measuring quality of employment as a broader concept relating job quality measurement within its universe. The framework is designed and discussed under the guidelines of various agencies to provide least controversial and physically measurable indicators across various countries, and various contexts. The indicators of major dimensions of quality and its sub dimensions are presented with their relevance and availably in major data bases of the world. This study puts great effort for presenting the concept as approachable and comprehensible at operationalization level. Though the report is written with macro perspective and addresses the relevant issues of data collection and availability at national levels it guides for how to test and adjust dimensions to Meso or micro levels. This study is set in European labor market comparison of concurrences of various job quality indicators is tested by principal component analysis for 27 European countries. The empirical investigations verified against the famous hypothesis concerning tradeoff between number of jobs and quality of jobs. The paper critically evaluates the functioning of Leaken indicators and includes complementary indicators for framing job quality comprehensively. Relative importance of included indicators is tested through principal component analysis. After reducing the dimensions of data countries are mapped into four clusters by hierarchical clustering technique. Each cluster is distinguished in geographical terms with the key features of various dimensions of Leaken indicators. The progress and failure in achieving some facets of job quality have aided in differentiating the clusters and the hypothesis of heterogeneous labor market across Europe is verified here .Kalleberg, Reskin, and Hudson (2000) debated in context of nonstandard work arrangements in USA. The writers have linked non-standard work arrangements to bad job characteristics and found 1 out of 7 doing bad job in terms of no access to health insurance low pension benefits and low pay levels. Since the analysis is specific for the workers groups who fall under nonstandard work arrangements therefore various categories of nonstandard employees relationship are tested with the prevalence of bad jobs characteristics by the means of statistical model, the paper also confirms gendered and racial biases in American labor market.

(Pugliesi, 1999)have shed light on various angles for looking at the construct of job quality, they emphasize the ever-evolving nature of job quality and updating its measurement toolbox for being dynamic in nature. In addition to theoretical review the writers have supported more for measuring QOE in terms of wellbeing, health outcomes or job satisfaction. The writers have highlighted current economic debate, and emerging lines of research for the policy makers, and pointed towards the links between narrowed individual job quality indicators with macro labour market outcomes. Overall, this study is a brief about current and ongoing research scenarios related to job quality research.

Lower-Basch (2007) write in American context for explaining the causes of failures of high-quality jobs for an average American. The discussion is critical and contextual. Within this regard they highlight the dichotomy of working organizations in American economy. First kind of working organizations are pro quality jobs whilst others expect good revenues as their counter parts but provide least to their workers for reducing their cost structures. This is ironical in American labor market context. In addition to inside gaps between expectations of workers to current working conditions general. The writers have touched upon core elements of quality jobs consisting of fair pay, paid leaves to security. Each factor is discussed separately for emphasizing the contribution of quality jobs for working class towards overall status of the country effectively. Huneeus, Landerretche, Puentes, and Selman (2015)conducted a case study for Brazilian labor market. They estimated employment quality through multidimensional where range of formal employment is estimated with the help macro-economic indicators including social security contributions, job tenure and earnings. The time series data is used from the National household survey. The writers utilized quality index for measuring deprivations at disaggregated level by precise job-related features. The conclusions drawn highlighted noticeable dissimilarities between the work quality of full-time employees and the self-employed class. Affiliation with trade union and employed of public sector were more found to be having higher employment quality .Weighting of dimensions in the paper subjective and each dimension is given equal value, multivariate analysis is applied and the data constrains for limited dimensions are mentioned. Cooke, Donaghey, and Zeytinoglu (2013) support the view that job setting should not be studied as an isolated construct, rather the externalities including over all working pattern of life should be included to understand quality of work. Here work is defined all kind of paid and unpaid voluntary jobs performed by the individuals. Through a focus group interview and accordingly designed questionnaire this research seeks to understand quality of work. In contrast to general practice job quality is taken as high relative concept in this paper. Subjective evaluation of work is explained through qualitative analysis on inside job. Seven typologies emerge from such a small group by comparing Job histories, life priorities and relative choices. J. Horowitz (2016) seeks to explore bridge between job quality dimensions and subjective well-being of American Adults in year 2000. The relative contribution of picked job quality indicators is measured by structural equation modeling on the subjective wellbeing reported at individual level. The analysis is further expanded on the basis of gender and racial segregations leaving the main picture of job quality unchanged with respect to gender contrary to racial biases. Broughton et al. (2016) measure Aggregate development over time for an economy can be measured under individual contract approach, where the contract defines the nature of work. The contract defined whether you might treat bad or good at job, you are predisposed to face precariousness or not. Nergaard, Alsos, Bråten, and Jensen (2015) used a range of data sources including data from the Norwegian Labor Force Survey from 2013 and from surveys of the retail, hotel and restaurant, and cleaning industries. In the ad hoc module to the LFS, three per cent of the workers stated that they 'did not have set working hours' which was used as a definition of on-call work (precarious work ratio) in context of Norway Labor market. The share of certain types of employment relationships – typically non-standard – is then considered to be precarious**.** As discussed by De Bustillo, Fernández-Macías, Antón, and Esteve (2011) in detail , consultant who are working on projects but are highly ranked in society and have quite choice to pick and choose between projects to work on how they would be called precarious for not working merely under standard work arrangements. Because of the limits of the individual contract approach to defining precarious work there is a need for other indicators too as mentioned by (Jesnes, 2018).The method is criticized for being too subjective, and may quickly end up as a reflection on how satisfied the employee is with his or her job situation, which is not necessarily an indication of precariousness. Another point of weakness is that employees are not always aware of what rights they are entitled to. Olsthoorn (2014)uses a combination of the quality of work and the individual contract approach to defines precarious employment in order "to move beyond non-standard contracts as a single indicator for precarious employment" they defines precarious employment as those who 'earn low wages, have little job- and income security and occupy jobs that can generally be deemed low quality'. They propose two indicators for measuring precarious employment. The first is income security, constructed by use of wage, supplementary income and unemployment benefits. The other indicator is job security, constructed by use of contract type and unemployment duration. The two indicators are integrated and then tested on Dutch Labor market data.

The results call for caution though when using non-standard contracts to indicate precariousness, as highly educated people have an equal risk of being employed on a non-standard contract. Ilsøe, Larsen, and Felbo-Kolding (2017)explore by descriptive statistics a combination of the quality of work approach and the individual contract approach in their article. They study the effect of part-time work on absolute wages through analysis of collective agreements and registry data on wages and working hours of Danish employees in the cleaning, retail, and hotel and restaurant sectors from the period 2008 to 2014. The findings indicate that de facto hourly wages have increased in all three sectors since the global financial crisis. Yet, many workers in these sectors work part-time, and particularly on marginal part-time contracts (15 hours or less per week). They introduce a new aspect by also focusing on the yearly income by concept of 'living hours. They find interesting and supportive argument for presence of unexpected pattern in developed Labor markets as they conclude for Part-time work and especially marginal part-time work are associated with very low yearly income levels even in cases like Denmark.

3.3. Indirect Approaches

A basic needs approach to development is one which gives significance to meeting the basic needs of all the people. With the endowment of material needs, like clothing and shelter , the non-material needs such as political liberty and employment are also considered fragment of basic needs pie (Stewart, 1985). Quality of employment (job quality) is not much addressed by this conceptualization still there exist two important and valuable contributions. The first study is by Körner, Puch, and Wingerter (2009) who describe broad perspective of human needs including other aspects of employment; the second study is by Green and Mostafa (2012) who presented a explicitly focused job quality model utilizing this approach . These methodologies are theoretically more particularized to measure the quality of employment. Körner et al. (2009) model consists of seven facets arranged in a pyramid with the most basic need at the bottom is set and the more self-improving dimensions of QOE is set at the top. The constraint faced by this framework is its measuring capacity of national average only, making disaggregated analysis of any kind including racial or gendered not feasible even provide detailed sectorial data. Since the data employs different sources of international and national sources to develop the indices therefore, we find mainly a tool for country-level case studies rather than international comparisons. By contrast Green and Mostafa (2012) deploy a single dataset (European household data) for developing model for measuring QOE. Four scopes of job quality having earnings, working time quality, prospects of work, core quality of the job constitute basis for the model proposed. The writers presented the scheme to compare average job quality for any subgroup within the dataset, for instance by country, gender or age. Gindling and Newhouse (2012) do macro study and analyze differences amongst the self-employed of 74 developing countries. They define successful entrepreneur as the one who can be considered as employer and who does not reside in poor household. With the help of very loose classification rule they describe how worker features differ by employment status and finally categorize self-employed workers either as unsuccessful or successful entrepreneurs. Regional comparisons are also presented in this paper through aggregate averages. Basic needs are criticized from development perspective at many forums, it was the lack inherited in its scope which led to development of other sophisticated philosophical approaches for human's requirements beyond physical needs. In context of employed / worker class of society we can generalize the same argument that humans' jobs can be more than a source of earning for them, their employment status effects and is affected by interlinked socio, economic and intangible aspects of life.

Other distinguished indirect approach to address QOE is the capability approach which poses an abstract framework for measuring two core normative claims. First, the claim that the liberty to acquire well-being is of key moral position and second; that liberty to accomplish well-being is to be assumed in terms of people's capabilities. In its wider sense, the capability approach not only assesses the lives of individuals but also includes other considerations in its valuations(Basu, 1987).

In case of labor studies we find few examples including Strotmann and Volkert (2008) for capabilities and social exclusion, Bartelheimer, Büttner, and Schmidt (2011)for dynamic aspect of employment trajectories ,and for evaluation of employment profile rather more few. Recent examples include Sehnbruch, González, Apablaza, Méndez, and Arriagada (2020) who proposed a procedure for measuring the quality of employment from a multidimensional viewpoint in Latin American developing countries. The paper demonstrates that the Quality of employment can be operationalized based on the capability approach. Using the method developed by Alkire and Foster (2011) they measure quality of employment at an individual level and establish a quality threshold within selected dimension to conclude whether a person is deprived or not within each quality dimension. Same capability approach was applied by García-Pérez, Prieto-Alaiz, and Simón (2017) to measure the prevalence of precarious employment in Spanish Labor market. They establish two types of thresholds for identifying precarious jobs. First one is used to detect jobs with drawbacks in each of the facets of precariousness considered and the other is tested to find multidimensional perspective of precariousness taking all considered facets of job quality. This study has opted to adapt the adjusted multidimensional poverty rate. For similar applications on employment and the capability approach see studies by (Sehnbruch, 2006), (Lugo, 2007), and (Leβmann, 2012).

The main critique regarding this approach is related to its operationalization. Operationalizing the capability approach is perplexing in two angles: first is its inherent multidimensionality constraint and second is its idea of freedom contributing to human wellbeing. There is consensus that a multiplicity of dimensions should be considered when measuring social phenomenon, but there is no settlement about which facets are important or how to select and weight them. Another limitation is huge amount of micro-data provision for employing this approach for finding accomplishments of each person in each selected dimension.

Third indirect approach to measure job quality develops framework taking this issue as inherently multidimensional. When we work with macro-level percentages, we cannot describe about individual scores across simultaneous indicators of job quality. For example, sectorial averages of various employment status and wage levels do not inform us about within sector quality picture. In this context, the multidimensional indicators of job quality help us to build theoretical framework that emphasizing employment quality as a phenomenon which cannot be summed up through basic statistics of wage and status. Secondly individual workers characteristics of jobs are considered in this framework(Sehnbruch, 2004).With this background, Leschke and Watt (2014) have created a synthetic job quality index (JQI) for the EU27 countries. The writer has attempted to explain for European countries job quality comparisons and for the evolutions of the same countries in terms of job quality over time. They briefly address multidimensional European job quality index for measuring changes over time within the limitations of the proposed framework. Segregated analysis under various specifications for various sub-indices is discussed in the article. Bocquier, Nordman, and Vescovo (2010) utilize the same approach to develop indicators of vulnerability in employment for capitals of West Africa. The writers study interlinks between individual incomes and job quality from the main job. Rigorous methodology for making cross-country is opted by standardizing variables across each study unit. The main finding suggests increase of earnings can compensate vulnerability position at employment .The determinants of this specific side of employment quality are found by regression analysis and quantitative results are further endorsed from qualitative perspective by applying principle component analysis Similarly Bazillier, Boboc, and Calavrezo (2014) developed indicators for European countries for employment vulnerability. Overall, they compare the situation for natives and migrants regarding this facet of QOE. The results indicate marginal difference for the situation considered between these two groups. The considered variables to develop multidimensional index comprise of the nature of employment contract, size of establishment, organization type among some others. In order to have an indicative picture of the interlinks between all indicators, a multiple correspondence analysis to reduce data to lower dimensions is applied. Greenan and Seghir (2017) propose a conceptual framework to analyze job quality at the workplace.

Other related studies in various scenarios are by Jencks, Perman, and Rainwater (1988) who proposed first index of job desirability. Olsthoorn (2014) explained development of indicators of precarious employment for the case study of Netherlands. Leschke and Watt (2014)constructed a 'Job Quality Index for Europe and Bescond, Chataignier, and Mehran (2003) proposed to measure work deprivations in 40 countries by means of a simple average.

The multidimensional job quality indices reviewed so far lack from some angles expertly summed up by (Leschke & Watt, 2014).Sehnbruch et al. (2020) elaborated with many case examples for explaining generation of indicators for a country as an underestimated picture of each country since we have to give up the richness of individual employment indicators for finding single measure in this framework .Leβmann (2012) also comments in against of this approach saying "Indices presume a particular structure, which means indices select various dimensions by choice which leads to the job quality structure neither discovered nor examined".

3.4. Theoretical Contribution

Lack of the benefits associated with nonstandard jobs are for various reasons as the employers don't exploit productivity fully of the employees, so they do not provide associated fringe benefits to temporary employees. But in case of full-time jobs the absence of chosen dimensions becomes crucial. (In this thesis for exploring QOE data contains more than 83 percent people doing full time only one job, so the considered sample is somehow homogenous in job number and nature). In above cited literature we find the discussion of quality of jobs around non-standard, self -employed, or temporary workers. It seems to be understood from literature that standard work arrangements inherently contain quality factor. None of the papers in context of America and other economies addressed full time/ standard work arrangements for exploring further about the quality of work. Inspired by (James, Witten, Hastie, & Tibshirani, 2013) suggestion for some other exploratory machine learning methods application in economics, we decided to diverge from conventional practices of above mentioned direct and indirect approaches of job quality measurement. We have opted model-based clustering techniques where choice of criterion of closeness of observations is not ad-hoc and the clustering solution is obtained through rigorous statistical procedures. Further we investigated for the hypothetical heterogeneity among the total sample of people who were assumed to belong to same group in terms of job quality. We have made context and data specific evaluations by giving weightage to internal differentiation within whole sampled class.

CHAPTER 4

METHODOLOGY AND DATA BRIEFING

In first section we brief about the data sources employed in this thesis followed by the variable's introduction and analytical framework (applied for model-based or latent class clustering). In next section models' selection and evaluation criteria are discussed. In last, model specifications for each of the four modeling frameworks are mentioned.

4.1. Rationale for Selected Sample

As explained in introduction of this thesis our objectives included to utilize various mixture models for cross sectional and longitudinal data. And specific for model-based clustering analysis the empirical application is set for finding clustering of job quality segments, in similar fashion job quality segments in regression framework with relatively few and different job quality indicators are found. The other chapters utilized single variable of employment status categories for longitudinal analysis. We wanted to employ some rich data source meeting all our theoretical and analytical modeling needs. Rich data on job quality indicators was available in European macro context by "European Foundation " and the applied 3 empirical examples in macro context are based on surveys by this source. To utilize some novel data source consisting of study related queries and to further check the general perception that developed world has far better quality of work situations than developing world we opted to take USA and UK case studies as interesting and representative of job quality situations in developed world. The data employed for first analysis is by "National Longitudinal Surveys (NLS)" which is a nationally representative sample of American 8,984 young men and women of ages 12 to 16 on December 31, 1996. This sample is followed for its cross-sectional civilian class over the age. Therefore, cross section segments and longitudinal data versions can be constructed. The year taken for cross sectional analysis is 2017 which was latest available at the selection stage of data. For applying growth and Markov modeling we have taken single variable of employment status with various categories for two representative cohorts' of 1979 adults and 1997 adults from the same data source. The NLSY97 Cohort is a longitudinal project that follows the lives of a sample of American youth born between 1980-84; 8,984 respondents were ages 12-17 when first interviewed in 1997. This ongoing cohort has been surveyed 19 times to date. For detailed discussion of variables and sampling related technical queries of national longitudinal surveys see into the technical report by (Moore, Pedlow, Krishnamurty, Wolter, & Chicago, 2000).

The other empirical data employed for finding differential effects of work features on job quality in regression mixtures chapter is longitudinal and consists of 9 waves over the years by (" Institute for Social and Economic Research"). It is taken from a national representative sample of British households' data source labeled understanding society. For further technical details of data scheme look into report by "Understanding Society" (2021).

4.2. Variables briefing

Since first modeling choice is set for mixed mode multivariate data for finding job quality typologies therefore in the following, we present the chosen indicators for measuring and developing analytical framework of job quality:

4.2.1. Hourly wages

Monetary gains measured through income are the prime least controversial and most applied indicator of judging class difference in terms of jobs. Many scholars under various disciplines (sociology, economics and psychology) support for monetary gains in shaping a person's quality of life beyond job. We have had wide dispersion of wages in our data from as low below 20 $ to 1000 plus as hourly wages.

4.2.2. Schedule/Work Arrangement type

In our data we have individuals reporting standard and nonstandard work arrangements though we are not taking direct information of their occupation initially schedule and timings of jobs has direct influence on a workers physical and mental wellbeing. Many studies report this indicator as a starting point for analyzing stable and precarious nature of work. More than 80% in NLS-97 (2017) data are working in day shift, and remaining work under various arrangements including night shift, rotation, split hours and evening. So, our sample is having more respondents reporting standard work arrangements in job schedule.

4.2.3. Company Size

Firm or workplace size is very much appreciated indicator for judging the quality of employment since with high work force size organizations it is expected that individual will gain more monetary and additional job benefits. Company size varied in our selected data from as low of 15 in number for up to 25 %of total and up to 50 % we have max reported score of 50 individuals at work place and onwards we have had quit disparities in work place size covering below 1000 score of workplace size for upto 87 % and many thousands of workplace size for remaining 13 %.With such a heterogeneous sized labor force we expected quite a change in job profiles of individuals. The visual inspection of such heterogeneities could be meaningless in this specific case.

4.2.4. Health Insurance

Health insurance provided by company, or the workplace is a huge source of psychological relaxation in terms of US economy especially since the health sector is quite expensive to afford with nominal/mediocre income levels. In the data employed we have more than 80 % provided with this facility.

4.2.5. All kind Emergency Leaves

Emergencies faced by employees make them absent from work and if least provided to an employee to meet uncertainty in life, then the outcomes are bad for the individuals and sometimes for the companies as well in form of low productivity and low morale of work force. From descriptive statistics presented in appendix B we find 20% individuals reporting as low as 5 paid all kind of emergency leaves and 37% having on average 10 days for all kind of such leaves annually whereas 15 % report a very good number exceeding 50 days paid leaves. This brings quite heterogeneity regarding individual scores in terms of paid leaves.

4.2.6. Weekly Hours of work

Underemployment is widely read in various macroeconomic contexts to link it to overall economic performance of the labor market of a country. Though it is relatively less linked to job quality. Despite low unemployment rates, generally Americans worked less than 25 hours per week. Less than 35 weekly working hours are considered underemployed in Australia and slightly different thresholds are available for U.S. Why Americans are not willing to work at full rate is another interesting dimension to explore driven by multiple factors. In the data taken more than 20 % work less than 35 hours per week whereas 47% work 40 hours per week and remaining distribution is stretched for up to 100 plus excessive working hours various ranges.

4.2.7. Union Coverage

To voice and bargain for employee rights under some central body is the classical indicator of workers security from exploitative practices. In modern day measurement practice of employment related quality frameworks this indicator is still considered crucial for measuring provision of basic employee rights. Almost 20 % in data reported have privilege of some union body to support workers' rights and say in their work place.

4.2.8. Locations

Multiple locations of workplace are also very much appreciated candidate in modern studies of measuring quality of employment since the workplaces having multiple locations are expected to be progressive in business. To keep that progress sustained it is expected from such businesses to offer additional privileges to their employees. In chosen data around 75 % belong to the group who operate their business in more than 1 location.

4.3. Analytical Framework

The aim of empirical analysis conducted in succeeding chapters is to demonstrate the range and diversity in terms of job quality for two case studies of developed nations. In first case through variants of model-based clusters we explored typology of job quality using cross sectional employment data for 2017 by NLS-97. By considering a range of above listed intrinsic list of job quality indicators endorsed by personal social economic attributes (gender, degree level, housing situation) we captured the central aspects of job quality by connecting the most likely occurrence of cases together. The purpose was to assess whether groups can be identified that differ distinctly in terms of job quality. This allowed us to determine whether it is indeed possible to differentiate between these clusters, and how these groups differ from one another. Since the concept of job quality is not directly quantifiable henceforth, it was imperative to operationalize it by approximation for measuring it. The presented framework includes variables that could cover some important dimensions and factors for the concepts of interest. The framework is not based on exhaustive lists of indicators like many other reported frameworks (Ben-Ishai, 2014; Olsen, Kalleberg, & Nesheim, 2010).Also compared to Findlay, Kalleberg, and Warhurst (2013) it includes the intersection of objective job quality features conditional to working for sufficiently measuring the exclusive concept of job quality.

Following Nations. (2015) and Erhel and Guergoat-Larivière (2010a) we also adhere to the overlapping loop nature of the concept for making it workable to understand. Firstly we point out contextual features which can be further subdivided as individual contextual factors, and working place contextual factors. Individuals with different attributes of knowledge and skill and unique background impact and get affected by the same working environment differently therefore education profile and age racial and gendered differences should be included at first or final level of cluster formation. The second component of context is related to the inbuilt environment of the working place which represents many related distinctive meso and macro constraints and ideological standards of the same place. The macro context is beyond scope of the investigation therefore for the working application we included some objective features to measure it such as workplace size and multiple locations. Working time arrangements directly influence the physical and mental capacity for completing the tasks. Excessive working hours crossing even 50 + hours are quite a norm in competing societies, whereas underemployment can also cause a psychological hindrance for grooming career over life. Though multiple factors operate for being over and under employed in a society still the working hours differences can solely contribute to understand labor force profiles better. From another perspective, working time arrangements departing from standard hours also depict bad and no choice situation of employees sometimes. Therefore rotational, and splitting and indecisive working hours compared to standard working hours are added to compare for any marked work life balance differences. Third category of work conditions could include detailed factors measuring physical and mental divergences between subgroups but constrained to indicators we focused on the core indicators of wage and nonwage benefits as conventional compensation levels for which anybody takes physical or mental pains. Weekly wages, all kind of paid emergency leaves, provision of health insurance was included as prime indictors to make comparatives of job quality. For measuring say and value of employees at workplace we included presence of union representation in workplace as last representative of job quality.

Fig 4.1. Analytical framework

4.5. Model selection and Evaluation Strategy

We will explain in given section the statistics involved in model estimation and selection. The topics mainly comprise of cluster size decision driven by formal criteria of evaluations amongst competing models (generally described as class enumeration issue of mixture models). We discussed in first section the statistical tools employed for model choices (class enumeration) followed by the technical details for exploring empirical cross sectional and longitudinal data. We discussed further the modeling framework for latent class cluster analysis followed by regression mixtures, growth mixtures and Markov models. Lastly, we discussed the data choice for empirical application for various models.

First, we present the core evaluation mix opted to evaluate and validate amongst competing models in next chapters. The specific modeling choices/concerns related to Markov and growth models are presented also here. We are answering in this section for How we compared and evaluated the competing models?

For chosen variants of mixture models we have utilized the mix of standard and nonstandard practices of model selection as a combined yardstick for model evaluations. To the best of our awareness this attempt of critically evaluating models by combining several diagnostic tools including cross validation and bootstrapping for multivariate sparse data is novel.For this purpose, three kind of economic data were used to set answer for important economic empirical queries since there is no study utilizing model based clustering and regression versions of mixtures in chosen empirical setups, therefore the proposed application of exploratory methods will add to the confirmatory approaches based applied literature in econometrics.

First issue encountered in models' estimation is of class enumeration (selecting cluster division from various possible alternatives). Class enumeration issue and selection of methods is well discussed in various articles related to mixture models in regression framework. Deciding about the no of components (clusters) is an unresolved matter though in such models but this task is of utmost importance since qualitative differences highlighted by mixture models are based on the adequate composition and size of clusters. In case of dealing with real data for finding clusters the true number of clusters is unknown and have to be inferred from the data by following model selection rules of relative and absolute performance (Tofighi & Enders, 2008). Below we provide a brief note on the general model evaluation criteria for with theoretical and technical details. Since the criteria used for deciding upon the best model are diverse so where was required, we cited for the detailed background studies. In the following we present specific details for these measures:

i) Hypothesis Testing criteria

ii) Information Criteria

iii) Classification Criteria

iv) Residual Diagnostic criteria

v) Graphical /cross validation criteria

4.5.1. Hypothesis Testing Criteria for Nested models

By using the well-known likelihood ratio test statistics, we can decide between the competing models (results) in case of mixtures employing either cross sectional or longitudinal data. In standard way final model choice of increasing number of components or clusters is by carrying out hypothesis Tests successively (where null hypothesis of S clusters is tested against alternative of S+1clusters). The procedure works well for non-sparse data but for many situations under mixture models' regularity conditions are not valid, therefore an alternative opted procedure is the Monte Carlo or bootstrapping based model evaluations for finding assessment of the p-value in such cases.

Since for the sparse data standard likelihood ratio difference is not valid because sparse data does not approaches to chi square distribution asymptotically(Aitkin, 1999).To account for complex nature of mixture models selection relatively more sophisticated and monte-Carlo based methods are proposed in literature (J. L. Horowitz, 1997).These bootstrapping based methods calculate and approximated distribution by finding difference of nested models in terms of differenced log likelihoods. The specific form of test are provided in detail (K. L. Nylund et al., 2007).The bootstrap based log likelihood difference test BLRT estimates the log likelihood difference distribution to obtain a p value to validate whether s - 1 class model is rejected in favor of the s class model. Very few application of bootstrap likelihood ratio are found in literature; for applied case of ordinal latent class analysis see into (Jeroen K Vermunt & Magidson, 2013) and for some applied case studies see in (Magidson & Vermunt, 2002). In our models we applied this strategy for choosing between nested models in latent class cluster models and growth models. The steps followed to conduct this test are described by following(Karen L Nylund et al., 2007).

1. Estimate the s-1 and s models for calculating the (-2LLHD).

2. Estimate null and alternative models under the null model, Then generate a bootstrap sample for estimating the (-2LLHD) between the null and alternative models (s-1 and s class models).

3. Samples should be replicated independently for finding the true distribution of the (-2LLHD).

4. In last estimate the p value by comparing the distribution obtained in third step with the (-2LLHD) obtained in first step. The significant p value then will lead to choose alternative model in favor of null model implying addition of further clusters contributes to better explaining the data.

ii) Goodness of Fit Criteria

For deciding about most suitable model in absolute terms Pearson or likelihood-ratio based goodness-of-fit test following chi-square distribution are employed in the form of frequency tables in model based clustering applications (Bartholomew & Tzamourani, 1999).For the evaluation of models in this thesis, we have not taken this rule as a general guide to choose best model because of sparse cells. For few situations of non-sparse data(step3 models) we employed three statistics for deciding over all fit of model based on chi-squared including Cressie-Read chi-squared statistic CR2 , the Pearson chi-squared and likelihood-ratio based chi-squared statistics (Collins, Fidler, Wugalter, & Long, 1993).The particular formulae are mentioned.

$$CR^2 = 1.8 \sum_{i^*=1}^{I^*} n_{p^*} \left[\left(\frac{n_{p^*}}{\widehat{m}_{p^*}} \right)^{2/3} - 1 \right]$$

$$L^2 = 2 \sum_{i^*=1}^{I^*} n_{p^*} \log \frac{n_{p^*}}{\widehat{m}_{p^*}}$$

$$X^2 = \sum_{i^*=1}^{I^*} \frac{\left(n_{p^*}\right)^2}{\widehat{m}_{p^*}} - N$$

.

4.5.2. Information Criteria

Since each addition of class increases the likelihood value in case of mixture model. To balance the rise in fit at the cost of larger number of parameters by opting for more components/classes-based solutions information criteria are used for judging most parsimonious models. These information criteria are based on the log likelihood of the selected model. These information criteria impose different penalty according to sample size, parameters of the model, or both. Since the penalty values are differently calculated across these criteria, it is possible to have different class solution. The criteria considered in this study are based on log likelihood suggested for sparse mixture models (Magidson & Vermunt, 2002).For calculations and relative comparisons of penalty and performance of various criteria see the reviews in (McLachlan, Lee, & Rathnayake, 2019; Oberski, van Kollenburg, & Vermunt, 2013).

Andrews and Currim (2003b) studied comparatives of 7 information criteria for regression mixtures. In their conclusions Akaike's Information Criterion (AIC) with another variant (AIC3), (imposing penalty factor of 3 for each parameter) produced very low parameter bias and performed as the best model evaluation rule in varied model conditions and data structures. Results from Andrews and Currim (2003a) and Dias, Vermunt, and Ramos (2015)suggest that AIC3 is a better criterion than BIC and AIC in determining the number of latent classes in LC and FM models. Formulae for these criteria are given below:

$$\begin{aligned}
CAIC_{lg\,\mathcal{L}} &= -2\lg\mathcal{L} + (\log(N) + 1)\ npr \\
BIC_{\lg\mathcal{L}} &= -2\lg\mathcal{L} + \log(N) npr \\
AIC_{\lg\mathcal{L}} &= -2\lg\mathcal{L} + 2\ \text{npr}, \\
AIC3_{\lg\mathcal{L}} &= -2\lg\mathcal{L} + 3\ \text{npr},
\end{aligned}$$

4.5.3. Classification Criteria.

Though Information Criteria impose penalty for rise in parameters to be calculated resulted from increased mixture components (clusters).We need still to validate the voted solution for ensuring the degree of separation in selected components. To allocate units in clusters classification criteria are suggested. These criteria check the performance of a mixture model to evaluate separation of components. In these criteria, according to the nature of variables the posterior probabilities are estimated and compared to evaluate the degree of separation. The classification criteria considered in this thesis measure the ability of a mixture model to provide well-separated clusters. This set of statistics contains information on how well we can predict to which latent class cases belong given their observed indicator and covariate values, or, in other words, how well the latent classes are separated. Classification is based on the latent classification or posterior class membership probabilities. For response pattern

$$\hat{P}(l \mid \mathbf{e}_i, \mathbf{y}_i) = \frac{\hat{P}(l \mid \mathbf{e}_i)\hat{f}(\mathbf{y}_i \mid l, \mathbf{e}_i)}{\hat{f}(\mathbf{y}_i \mid \mathbf{e}_i)}.$$

The numerator and denominator are the maximum likelihood estimates for the terms appearing in the general mixture model defined in equation (1) in next section. These quantities are used to compute the estimated proportion of classifications errors (CE), as well as entropy-based measures for nominal variables:

$$CE = \frac{\sum_{j=1}^{J} w_i\left[1 - \mathrm{m}\ \hat{P}(l \mid \mathbf{e}_i, \mathbf{y}_i)\right]}{N}$$

$$R^2_{l,\ \text{entropy}} = \frac{\mathrm{er}(l) - \mathrm{er}(l \mid \mathbf{e}, \mathbf{y})}{\mathrm{er}(l)}$$

$\mathrm{er}(l)$ is the total associated with predicting latent variable l given no knowledge about covaraites e and response variables y, and $\mathrm{er}(l \mid \mathbf{e}, \mathbf{y})$ is the prediction error for all observed information from the cases). Case-specific errors are average estimate and labeled as

er $(l \mid \mathrm{e}, \mathbf{y}_i)$

$$\mathrm{er}(l \mid \mathbf{e}, \mathbf{y}) = \frac{\sum_{j=1}^{J} w_i \mathrm{er}(l \mid \mathbf{e}_i, \mathbf{y}_i)}{N}.$$

$$\mathrm{er}(l) = \sum_{l=1}^{K} \hat{P}(l \mid \mathbf{e}_i, \mathbf{y}_i) \log \hat{P}(l \mid \mathbf{e}_i, \mathbf{y}_i)$$

For computing latent errorer(l), the $w_i\hat{P}(l \mid \mathbf{e}_i, \mathbf{y}_i)$ are exchanged with the marginal latent probabilities $\hat{P}(x)$ defined as:

$$\hat{P}(l) = \frac{\sum_{j=1}^{J} w_i\hat{P}(l \mid \mathbf{e}_i, \mathbf{y}_i)}{N} = \frac{\sum_{j=1}^{J} \hat{w}_{li}}{N}$$

4.5.4. Residual Diagnostic Criteria

The Other Criteria include graphical diagnosis tools or cross-validation. We have utilized different mixes of these model selection approaches to evaluate and validate final no of clusters/ mixtures in each section. Final model was selected by observing improvement in likelihood values of models by comparing relative scores of these statistics.

Residual diagnostics to check for model fit are relatively less applied for variants of models(regression mixtures, growth mixtures and Markov models) but there is enough argument for their utilization in deciding about model unfit(Nylund-Gibson et al., 2019b). The diagnostics is mainly applied to measure the extent of associations between various possible pairs of indicators and covariates involved in study. Here we answer for, what bivariate residuals imply and how they help in deciding about the model misfit and lastly the different version of rules used to calculate these for cross sectional and longitudinal observations. Since conditional independence of indictors with respect to the hidden structure is the basic assumption of standard model-based cluster analysis therefore the failure of this assumption becomes the major cause of model misfit.

In the following we provide formulae for calculating bivariate residuals BVR between all possible pairs of covariate and indicators (Jeroen. K. Vermunt, 2005) and between pairs of indicators. The estimated values may be any real number and for a given significance level BVR report the extent of mutual association of pairs. The value of BVR >3.5 indicates failure of mutual independence assumption(Jeroen K Vermunt & Magidson, 2013).For categorical indicators, a BVR has the form of a Pearson chi-squared divided by total observations and the Sum is taken for observed values and frequency table is counted for the non-missing values.

$$BVR_{j,j'} = \frac{1}{P}\sum_{v=1}^{V_j}\sum_{v'=1}^{V_{j'}} \frac{\left[n_{v,v'} - er\left(n_{v,v'}\right)\right]^2}{er\left(n_{v,v'}\right)}$$

$$er\left(n_{v,v'}\right) = \sum_{i=1}^{I} m_i \sum_{x=1}^{K} \hat{P}\left(y_{ij} = v \mid l\right)\hat{P}\left(y_{ij'} = v' \mid l\right)\hat{P}(l \mid \mathbf{e}_i, \mathbf{y}_i)$$

In case of sparse frequency tables for mixed mode data in our analysis, the issue of empty cells in frequency table of bivariate associations gets higher. For that reason we estimated bootstrap p-values for the bivariate residuals by parametric bootstrap proposed by Oberski et al. (2013) with the 5% critical values (CV) and the Monte Carlo standard errors.

4.5.5. Longitudinal Bivariate Residuals (LBVR)

BVR are also calculated to measure associations at multiple levels of group and individuals. Since the data had multilevel structure, where individuals were nested within time units (see applications in next chapters).So we quantified between-group differences and within-group similarities in responses on the indicator concerned (Nagelkerke, Oberski, & Vermunt, 2017) BVR-group is equivalent to the BVR obtained by using the group id variable also as a nominal covariate (with its effect set equal to 0). The BVR-pairs computation for categorical indicators involves setting up the two-way cross-tabulation for the responses of pairs of observations within groups. The estimated frequencies $E(n_{m,m'})$are obtained as follows:

$$E(n_{u,u'}) = \sum_{j=1}^{J} \sum_{i=1}^{I_j} \sum_{i'<i} we_i we_{i'} \sum_{l^g=1}^{M^g} \hat{P}(y_{jit} = u \mid l^g)\hat{P}(y_{ji't} = u' \mid l^g)\hat{P}(l^g \mid \mathbf{e}_j, \mathbf{y}_j)$$

BVR-pairs equals the resulting chi-squared value divided by M · (M − 1)/2 (the number of parameters of a symmetric association) and by the average group size.

For Markov models The three types of longitudinal BVRs are BVR-time, BVR-lag1, and BVR-lag2 are estimated as described in (Jeroen K Vermunt & Magidson, 2013).These bivariate residuals at different levels indicate whether the estimated Markov model account for the time trend, the first-order autocorrelation, and the second-order autocorrelation, separately, for the variable under study. BVR-time is equivalent to the BVR obtained by using the time variable as a nominal covariate (possibly with its effects set equal to 0). For categorical indicators, BVR-lag1 andBVR-lag2 are obtained by cross-tabulating the responses at time points t−1and t (BVR-lag1) and at t − 2 and t (BVR-lag2). For BVR-lag1, Again the margins of the estimated table are adjusted to be equal those of the observed table. BVR-lag1 equals the Pearson chi-squared value for this table divided by (M −1)2 and by the average number of lag-1 responses per individual. BVR-lag2 is computed in a similar way.

$$E(n_{u,u'}) = \sum_{i=1}^{I} we_i \sum_{t=2}^{T_i} \sum_{l_{t-1}^d=1}^{M^d} \sum_{l_t^d=1}^{M^d} \hat{P}(y_{it-1} = u \mid l_{t-1}^d)\hat{P}(y_{it'} = u' \mid l^d)\hat{P}(l_{t-1}^d, l_t^d \mid \mathbf{e}_j)$$

4.6. Model Specifications

In first section we explained about how we selected amongst competing models. All of the models were evaluated by diagnostic tools discussed above for model selection in case of nested or non-nested models. Wherever it was required we differentiated the formulae for specific model category otherwise common statistics utilized across all methods are mentioned in above section.

The discussed modeling scenarios mainly consist of 4 modeling subfamilies of mixture models in regression framework (where first family of method is not set up for finding regression coefficients for different clusters of data, but the estimations utilized multinomial or ordinal regression functions from generalized linear modeling framework) second family of models is the regression mixtures and growth variants. The last family of Markov models with transitions and mixture variants uniquely addressed for heterogeneity and measurement of transitions over the time.

All these methods shared some commonalities. They utilized either categorical or mixed mode data, which emerged to be sparse in nature since multivariate cross classifications led to empty cells. All methods were serving to find theoretically hypothesized mixtures / subgroups, classes or clusters in data. In the next section relevant technical background is presented for each family of models

4.6.1.Latent Class Cluster Analysis

Following Skrondal and Rabe-Hesketh (2004) latent variable can be considered as a random variable whose actual values are not known in advance. This variable is opposite to manifest variable measured in known terms. Multiple purposes are met by targeting these variables in various methodological frame works starting from measuring errors present in considered true variables , measuring hypothetical constructs, measuring categorical latent response variables , and accounting for unobserved heterogeneity in data from unknown source for details see in (Skrondal & Rabe-Hesketh, 2004; Tuma & Decker, 2013). The benefits associated in employing latent framework are immense. To put it shortly we can incorporate various above noted issues if desired. For an example, we have worked with all these objectives in first section of models to address the issue of hypothetical quality of employment (an ideal economic construct which is not directly measured). Heterogeneity of individuals within the theoretical framework is also addressed by applying these models in same context with additional modeling changes. Classes of quality is a discrete phenomenon therefore inherently incorporated in categorical /discrete latent class cluster analysis. Variant of classical model which incorporate heterogeneity and local dependence between the residual is also modeled.

There is broad line for modeling continuous latent variables and discrete latent variables. We have worked with discrete /categorical latent variable which leads to exploration of classes or clusters in given data. Since the nature of objectives is based on methodological concerns driven from economic theory therefore by choice opted for discrete analysis for finding differences utilizing empirical exploratory economic data analysis. Although methods opted are not pure exploratory like in unsupervised learning methods, but the methods serve the purpose of finding homogenous groups within one group are met in adopted modeling set up of latent class cluster analysis.

4.6.2. Unconditional and Conditional models

We start with the equation from which special cases for pure exploratory purpose in first section of estimations, and regression variants are derived in further sections. The equation presents most general structure based on generalized linear models family (Skrondal & Rabe-Hesketh, 2004; Jeroen K Vermunt & Magidson, 2013). Here l serves as a latent variable which relates exogenous (covariates or predictors) are depicted as ex, dependent variables as y, over the index I. The relation between these variables is described as

$$f(\mathbf{y}_i \mid \mathbf{e}_i) = \sum_{x=1}^{K} P(lt \mid \boldsymbol{e}_i) f(\mathbf{y}_i \mid lt, \boldsymbol{ex}_i) = \sum_{x=1}^{K} P(lt \mid \mathbf{ex}_i) \prod_{h=1}^{H} f(\mathbf{y}_{ih} \mid lt, \boldsymbol{ex}_i) \quad (1)$$

$f(\mathbf{y}_i \mid \mathbf{e}_i)$ is taken as the probability density function estimated from $f(\mathbf{y}_i)$values conditional on $(\boldsymbol{e}_i)$ values, here the latent variable aids between the $\mathbf{e}_i$and the $\mathbf{y}_i$ variables. $P(lt \mid \boldsymbol{ex}_i)$ is the probability of belonging to a certain latent class given an individual's realized covariate values, and $f(\mathbf{y}_i \mid lt, \boldsymbol{ex}_i)$ is the probability density of $\mathbf{y}_i$ conditional and $\mathbf{ex}_i$ (the mixture densities). This implies that latent variable can be influenced by exogenous variables, and response variables possibly be affected by both exogenous and latent variables. The last $f(\mathbf{y}_{ih} \mid lt, \boldsymbol{ex}_i)$ part indicates that response variables of various segments are mutually independent given the latent and exogenous variables.

Standard latent cluster models based on local independence assumption between all indicators and with no covariates derived from the above equation is following:

$$f(\mathbf{y}_i \mid \mathbf{ex}_i) = \sum_{x=1}^{K} P(lt) f(\mathbf{y}_i \mid lt, \mathbf{ex}_i) = \sum_{x=1}^{K} P(lt) \prod_{j=1}^{J} f\left(y_{ij} \mid lt\right) \quad (2)$$

And the latent class model for n categorical indicators can be inferred as

$P(y_{i1} = u_1, y_{i2} = u_2 \dots y_{in} = u_{\mathrm{n}}) = \sum_{l=1}^{L} P(lt) \prod_{j=1}^{\mathrm{n}} P(y_{ij} = u_j \mid lt)$

Here,

$$\prod_{j=1}^{\mathrm{n}} P(y_{ij} = u_j \mid lt) = P(y_{i1} = u_1 \mid l) P(y_{i2} = u_2 \mid l) \dots P(y_{i\mathrm{n}} = u_n \mid l)$$

Within the context of the generalized linear modeling (GLM) framework the transformation of the expected value of the response variable that yields the linear predictor restricted by a regression model, is referred to as the link function(Hastie & Pregibon, 2017) .Henceforth for each kind of variable the corresponding linear predictors and regression models are available. Since mixed mode is encountered in first modeling section. Let m, denotes a particular category of response variables over cross sections and categories number is indicated byu_j. The distribution function for Nominal and ordinal dependent variables from a multinomial distribution with any category can be described as

$$P(y_{ij} = u \mid l, \mathbf{ex}_i) = \pi_{u|t,l,\mathbf{e}_i} = \frac{\exp\left(\mu^t_{u|l,\mathbf{e}_i}\right)}{\sum_{m'=1} \exp\left(\mu^t_{u'|l,\mathbf{e}_i}\right)} \quad (3)$$

$\pi_{u|t,l,\boldsymbol{e}_i}$ is the probability of any response pattern conditional to latent variable and exogenous variables. Additionally, $\left(\mu^t_{u'|l,\mathbf{e}_i}\right)$denotes the linear term that can be further restricted by a regression model according to the nature of categorical variable, typically for nominal and ordinal categorical variables multinomial and adjacent-category ordinal logistic regression model are utilized (Agresti et al., 2000; Simonoff, 2003)

The multinomial probability distribution for a single nominal latent variable conditional to covariate values is parameterized as a standard multinomial Logit model:

$$P(l \mid \mathbf{e}_i) = \pi_{l|\mathbf{e}_i} = \frac{\exp\left(\mu_{l|\mathbf{e}_i}\right)}{\sum_{l'=1}^{L} \exp\left(\mu_{l'|\mathbf{e}_i}\right)} \quad (4)$$

When estimating model for nominal variables in data, we utilized the baseline-category logit model (Agresti, 2002). For case of a nominal outcome variable by using dummy coding with the reference category denoted by m′, we get the following definition of the linear term $\mu_{u|\mathbf{e}_i}$ given in eq (5)

$$\eta_{u|\mathbf{e}_i} = \log\left(\frac{P(y=u|\mathbf{e}_i)}{P(y=u'|e_i)}\right) = \beta_{u0} + \sum_{k=1}^{K} \beta_{uk} \cdot e_{iu} \qquad (5)$$

For ordinal variables following parameterization is used(Agresti et al., 2000)(Agresti, 2002; Magidson, 1996).

$$\eta_{n|\mathbf{e}_i} = \beta_{n0} + \sum_{k=1}^{K} \beta_{\mathrm{n}k} \cdot y_n^* \cdot z_{ik} \qquad (6)$$

For further details on the N − 1 adjacent-category logits, see in ((Jeroen K Vermunt & Magidson, 2013).

$$\log\left(\frac{P(y=n+1|\mathbf{e}_i)}{P(y=n|\mathbf{e}_i)}\right) = \mu_{n+1|\mathbf{e}_i} - \mu_{n|\mathbf{e}_i} = \beta_{n0}^* + \sum_{k=1}^{K} \beta_{\mathrm{n}k} \cdot (y_{n+1}^* - y_n^*) \cdot z_{ik} \qquad (7)$$

4.6.3. Conditional models

An important extension of conditional models incorporating covariates in basic model are described in (Hagenaars & McCutcheon, 2002; Shockey, 1988).

The latent class model with three categorical indicators and two covariates can be defined as

$$P(y_{i1} = u_1, y_{i2} = u_2, y_{i3} = u_3 \mid e_{i1}^{\mathrm{c}}, e_{i2}^{c}) = \sum_{x=1}^{K} P(l \mid e_{i1}^{\mathrm{c}}, e_{i2}^{\mathrm{c}}) \cdot \prod_{t=1}^{3} P(y_{it} = u_t \mid l). \qquad (8)$$

Angolous to simple case with no covariates logistic regression functions are employed to calculate conditional probabilities as (Simonoff, 2003).

$$P(l \mid e_{i1}^{\mathrm{c}}, e_{i2}^{\mathrm{c}}) = \frac{\exp(\mu_{l|e_{i1},e_{i2}})}{\sum_{l'=1}^{K} \exp(\mu_{l'|e_{i1},e_{i2}})} \qquad (9)$$

$$here\ \mu_{l|e_{i1},e_{i2}} = \omega_{l0} + \omega_{l1}e_{i1} + \omega_{l2}e_{i2}$$

4.6.4. Direct Effects Model

Since the data utilized in first section was multivariate job quality indicators, so there was a fair chance of co-occurrences of multiple indicators in first empirical application. The correlated indicators bias the results inferred from standard latent class model. How we modeled local dependencies in LCCA model with some indicators and auxiliary variables (covariates) is explained with the given formulae. In given case, we assumed that two indicators are conditionally dependent so that the RHS of eq (8) transforms. See explanation in (Jeroen K Vermunt & Magidson, 2021).

$$P(y_{i1} = u_1, y_{i2} = u_2, y_{i3} = u_3 \mid e_{i1}^{c}, e_{i2}^{c}) = \sum_{l=1}^{K} P(l \mid e_{i1}^{c}, e_{i2}^{c}) P(y_{i1} = u_1, y_{i2} = u_2 \mid x) P(y_{i3} = \mid l, e_{i2}^{c}) \qquad (10)$$

The most general LC Cluster model is the model for mixed mode data (Brian S Everitt, 1988) ,(Lawrence & Krzanowski, 1996)and(Moustaki, 1996).

This complex model is used in our analysis by using variables of different scale types for employment data. The structure that served as the starting point was again the local independence structure that we also used for categorical and continuous variables (see equation 5). For each indicator, after specifying nature of variables as nominal, ordinal, continuous, conditional models for covariates and distal variables were set up. The marked relation of covariates and indicators was measured through direct effects modeling via joint multinomial and multivariate normal distributions. Local dependencies between pairs of categorical (nominal or ordinal) variables and between pairs of continuous variables are dealt similarly. Equation utilized for multivariate normal within classes is given below:

$$f(\mathbf{y}_i \mid l) = (2\pi)^{-\frac{K_m}{2}} |\Sigma_l|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\mu}_l)' \Sigma_l^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_l)\right\} \qquad (11)$$

4.6.7. Latent Class Model with Continuous Factor

To account for within differences of the sample population inclusion of continuous factors within general nominal latent class is suggested and applied in different contexts. This version was applied for heterogeneity testing within factor models by (Yung, 1997).The specific formulae for the particular probability structure and related discussion with various applications is given in(Jeroen K Vermunt & Magidson, 2013).

$$f(\mathbf{y}_i \mid \mathbf{e}_i) = \sum_{l=1}^{L} \int_{\mathbf{C}_i} f(\mathbf{C}_i) P(l \mid \mathbf{e}_i) f(\mathbf{y}_i \mid l, \mathbf{e}_i, C_i) \mathrm{d}\mathbf{C}_i \qquad (12)$$

Here.

$$f(\mathbf{y}_i \mid l, \mathbf{e}_i, \mathbf{C}_i) = \prod_{h=1}^{H} f(\mathbf{y}_{ih} \mid l, \mathbf{e}_i, \mathbf{C}_i)$$

C_i denotes here the score of case i on continuous latent variable.

4.6.8. Step 3 Analysis Procedures

Another approach to conduct conditional analysis is three-step LC analysis. The analysis is performed to incorporate auxiliary information in the model. Step 1 approach implies inclusion of covariates as active or inactive exogenous variables in step 1 of forming clusters. Step 1 approach implies the active role of covariates in cluster formation. The alternative methods to include such variables are broadly defined as step 3 method since these approaches are based on 3 steps. The relevant debate and advantages of utilizing such methods are described in literature review (extension section). Step 3 analysis is the alternative strategy compared to step 1 analysis for incorporating role of covariates or distal /dependent outcomes in mixture models.

Though we have applied this alternative methodology in different modeling setups and in each set up including basic cluster, regression or Markov models' technical differences emerge for estimations and algorithms naturally. For specific cases we refer to Jeroen K . Vermunt (2010) for Markov step 3 models. For growth and regression mixtures B. Muthén and Asparouhov (2002a) provide guide and case discussions. The core common structure of doing three steps is presented in the following from (Zsuzsa Bakk, Fetene B Tekle, & Jeroen K Vermunt, 2013b).

1. First, a latent class model is built for a set of response variables. This involves decisions regarding the indicators to be used, the number of classes needed, and other model features.

2. Using the final model from step 1, subjects are assigned to latent classes based on their posterior class membership probabilities, and the class assignments are appended to the data file. Class assignment can be modal (to the class for which the posterior membership probability is largest) or proportional (to each class with a weight equal to the posterior membership probability for that class).

3. Using the assigned class memberships from previous step, the association between the class membership and exogenous variables is examined with multinomial logistic regression analysis or simple cross-tabulations. The external variables can be (distal) outcomes influenced by class membership or both predictors of class membership. In case of applying proportional assignment in step 3 analysis, adjusted step-three maximum likelihood-based analysis requires expanding the data set to hold M records per entity having weights equal to

the posterior membership probabilities. To incorporate these weights more efficiently BCH adjustment based robust standard errors are proposed by (Bakk et al., 2013a).In the BCH adjustment, instead of estimating a LC model one performs the logistic regression analysis or computes the cross-tabulations or ANOVAs in the standard way. With the modification that an expanded data file with M records per entity and a specific set of weights is employed. For more details, we refer the reader to (Jeroen K . Vermunt, 2010; Jeroen K Vermunt & Magidson, 2021). Gudicha and Vermunt (2013) in simulation studies showed that the ML adjustment is the preferred approach when the external variables are covariates or categorical dependent variables. Also the suggested adjustments utilized were of BCH or ML with modal assignment of values or proportional assignment of values in step 1.Following the above mentioned studies we have employed all of these 4 types of versions for doing step 3 analysis.

4.7. Regression mixtures /Latent class Regression Models

Regression mixtures is the subfamily of finite mixtures (McLachlan et al., 2019; Wedel & Kamakura, 2000).The advantage of utilizing these methods is dual. They serve as an exploratory exercise for finding clusters or subgroups in heterogeneous data and after class enumeration the reported parameters can serve to find across classes meaningful differences. We get as a byproduct of this modeling scheme different regression relations for each cluster. We used repeated measured data from one longitudinal survey in this thesis. There were few differences in variables selection but the core to measure job quality differences was same in both sections). Given latent variable l and repeated observations of job satisfaction, P predictors $\mathbf{e}_i^p$affecting $\mathbf{r}_i$, and using R numeric or nominal covariates $\mathbf{e}_{i1}^c$affecting l.Single case total replication are denoted by$\mathbf{r}_i$, with T_i denoting as total replications.

The main differences from the Cluster section are that in Regression we make a distinction between covariates and predictors, we allow for different numbers of replications per case, we assume that the conditional densities $f\left(r_{it} \mid l, \mathbf{e}_{it}^p\right)$have the same form for each t, and we do not allow for direct effects between the multiple responses. The most general probability structure takes on the following form:

$$f\left(\mathbf{r}_i \mid \mathbf{e}_{i1}^c, \mathbf{e}_i^p\right) = \sum_{l=1}^{L} P(l \mid \mathbf{e}_i^c) \prod_{t=1}^{T_i} f\left(r_{it} \mid l, \mathbf{e}_{it}^p\right) \quad (13)$$

Restricted form involving only predictors becomes:

$$f\left(\mathbf{r}_i \mid \mathbf{e}_{i1}^p, \mathbf{e}_{i2}^p\right) = \sum_{l=1}^{L} P(l) f\left(r_i \mid l, e_{i1}^p, e_{i2}^p\right) \tag{14}$$

4.7.1. Conditional models

An important extension of the latent class Regression model for repeated observations is obtained by making class membership dependent on covariates (Kamakura, Wedel, & Agrawal, 1994). In this model, it is assumed that the probability of belonging to latent class x depends on the values of e_{i1}^c, e_{i2}^c . This is equivalent to the way covariates can be used in LC Cluster models (see section 3.2.3.). Such a LC Regression model for repeated measures is very similar to multilevel (two-level), mixed, or random-coefficients models, in which random effects are included to deal with the dependent observations problem(Agresti et al., 2000; B. Muthén & Asparouhov, 2002b). The LC Regression model is in fact, a nonparametric random-effects model ((Agresti et al., 2000);(Aitkin, 1999).All of the variants are applied on longitudinal employment data in results section 2, additional considerations are inclusion of sampling weight and complex sampling standard errors calculations, since the sample chosen is clustered sample

$$f\left(\mathbf{r}_i \mid e_{i1}^c, e_{i2}^c, \mathbf{e}_{i1}^p, \mathbf{e}_{i2}^p\right) = \sum_{l=1}^{L} P(l \mid e_{i1}^c, e_{i2}^c) \prod_{t=1}^{T_i} f\left(r_{it} \mid l, e_{it1}^p, e_{it2}^p\right) \tag{15}$$

Alternative approach opted for conditional step 1 analysis is step 3 analysis in regression framework since analysis is discussed in clustering framework in the above section. Here we don’t provide further details for doing step 3 in regression framework since the modeling set up is almost same. For comparisons see into (Jeroen K . Vermunt, 2010).

4.8. Growth mixtures/Latent class growth analysis

Here we brief about growth modeling and its specific versions utilized in our analysis of employment status change /differentials across cohorts and for finding gendered differences in growth trajectories of various classes. In the following we summarize modeling differences between conventional / standard growth modeling to the family of mixtures of growth, next we explain technical differences for these models through statistical terminology.

Underlying simple growth structure is the notion that all persons are drawn from a single population with shared parameters. This assumption is relaxed under the growth mixture framework and for mixture components varied growth parameters are feasible to calculate. This task of un-mixing the population in terms of different growth parameters is accomplished using latent categorical variables. These categorical variables allow to find trajectories or paths of change for subgroups for different groups of individual growth trajectories to vary around different group averages. The distinct growth models for each subgroup/cluster sometimes provide quite unique estimates of covariate effect. Latent class growth analysis (LCGA) is a restricted version of growth mixtures, the underlying difference of this model is the pre imposed homogenous structure of growth within each subgroup. Henceforth variance and covariance estimates for the growth factors within each class are assumed to be fixed to zero. For broad discussion of this version of models see into(Daniel S Nagin & Land, 1993)

We in our analysis have employed mixture growth model and restricted variants (latent class growth models. The specifications of these models for categorical variables are given below:

4.8.1. Standard Growth Model

A longitudinal model for categorical data that does model the individual differences is known as a generalized linear mixed model (Skrondal & Rabe-Hesketh, 2004). The model is expressed as having two levels. Level 1 describes the unit change in latent responses at each time point, and at level 2 we describe the unit change over time. The Level 1 equations are:

$$r_{\mathrm{ti}}^{*} = \beta_{0i} + \beta_{1i} a_{\mathrm{ti}} + \varepsilon_{\mathrm{ti}} \qquad \text{(G1)}$$

At Level 2, individual differences in the random coefficients from Level are represented by variability (v_{0i}, v_{1i}) around the mean intercept β_{00} and mean slope β_{01}.The individual differences are modeled as a function of an individual-level, time-invariant covariate, y_i (multiple covariates are possible) quantified by regression coefficients β_{01} and β_{01} for intercept and slope, respectively. The conditional joint distribution of the intercept and slope is assumed to be multivariate normal. In the following we only describe equations.For technical differences and detailed elaboration of the given models see into (Feldman, Masyn, & Conger, 2009)

$$\begin{aligned} \gamma_{0i} &= \beta_{00} + \beta_{01} y_i + v_{0i} \\ \gamma_{1i} &= \beta_{00} + \beta_{11} y_i + v_{1i} \end{aligned} \qquad \text{(G2)}$$

$$\mathrm{pr}(c_i = k \mid y_i) = \frac{\exp[\delta_k + \vartheta_k y_i]}{\sum_{h=1}^{K} \exp[\delta_h + \vartheta_h y_i]} \qquad \text{(G3)}$$

4.8.2. Growth Mixtures

From the standard growth model with the restriction of different growth curves of k subgroups or clusters we can add the subscript k in above sequence of equations where each class has its own variance covariance structure. By further imposing homogeneity of parameters change within each class we can acquire latent class growth curves.

$$r_{\mathrm{kti}}^{*} = \gamma_{k0i} + \gamma_{k1i} a_{\mathrm{ti}} + \varepsilon_{\mathrm{kti}} \tag{G4}$$

$$\begin{aligned} \gamma_{k0i} &= \beta_{k00} + \beta_{k01} y_i + v_{k0i} \\ \gamma_{k1i} &= \beta_{k10} + \beta_{k11} y_i + v_{k1i} \end{aligned} \tag{G5}$$

$$\mathrm{pr}(c_i = k \mid y_i) = \frac{\exp[\delta_k + \vartheta_k y_i]}{\sum_{h=1}^{K} \exp[\delta_h + \vartheta_h y_i]} \tag{G6}$$

4.9. Markov Models

This section deals with the situation where one discrete (or categorical) variable es, which we refer employment status is measured at several consecutive times, es1 es2 ………esm with realizations i, j, ... ,m. Time range taken is (6 to 16 years)and the sample size is large for 2 cohorts .The models are evaluated for one and 2 generations (cohort effects). Analysis with Markov models presented in this chapter focuses on the types of change and stability for making employment choices for two time zoned people.Though we have no measures to verify that change in employment statuses (over 4 categories different pairs) by choice or by economic constraints. Our focus remains to explore the data for possible heterogeneous groups in making transitions from one state of employment status to another, and mainly to evaluate the margin of measurement error in measuring the latent variable of employment status. The models are taking into consideration the discreteness of both space and time. They are formulated on the latent level either by postulating unknown types of change for the classes (clusters) or subgroups of individuals.

The models presented are aligned with their theoretical superiority to address issues in measurement of states (for the technical details of other restricted versions mixed Marko ,manifest Markov and others see in in (MacDonald & Zucchini, 1997).In these models we classify the dynamic nature of class membership from the static class membership in standard mixture models.We follow the general practice in literature and call the dynamic classes as latent states(Magidson et al., 2009b).The modeling terminology is adopted from(Nylund-Gibson et al., 2014) (F. Van de Pol & R. Langeheine, 1990)

For model 1 Let sdt denote the latent state at various time points, followed by the superscript d for dynamic. The time variable t runs from 0 to the last measurement occasions for the person concerned. Let us first look at a latent Markov model for a single categorical response variablees_{it}, here we denote the vector taking the responses of individual i at all-time points by es_i. Resulted three sets of probabilities from this model are the initial state probabilities $P\left(s_0^d\right)$ the transition probabilities $P\left(s_t^d \mid s_{t-1}^d\right)$and the response probabilities $P\left(es_{it} \mid s_t^d\right)$.In the given equation product $\prod_{t=1}^{T_i} P\left(s_t^d \mid s_{t-1}^d\right)$ comes from the first-order Markov assumption, indicating that the latent state at time point t is dependent on the previous connected state $t-1$, but not on the states at earlier time points. Whereas the product. $\prod_{t=0}^{T_i} P\left(es_{it} \mid s_t^d\right)$ indicates that the response at time point t depends on the latent state at the current time point, but not on the latent states or the responses at other time points(Bartolucci et al., 2014) .

4.9.1. Latent Markov model

It can be defined as

$$P(es_i) = \sum_{s_0^d=1}^{u^d} \sum_{s_1^d=1}^{u^d} \ldots \sum_{s_T^d=1}^{u^d} P\left(s_0^d\right) \prod_{t=1}^{T_i} P\left(s_t^d \mid s_{t-1}^d\right) \prod_{t=0}^{T_i} P\left(es_{it} \mid s_t^d\right) \qquad \text{(M1)}$$

4.9.2. General Latent Mixture Model

see(Bartolucci et al., 2019) and (Bartolucci et al., 2014) for introduction , procedure ,and comparative applications in different study domains This general model is opted from these.

Here we have four sets of probabilities:) $P(s \mid \mathbf{z}_i)$are class proportions which may depend on time-constant covariates, $P\left(s_0^d \mid s, \mathbf{z}_i\right)$ are initial state probabilities which may depend on classes and time-constant covariates, $P\left(s_t^d \mid s_{t-1}^d, s, \mathbf{z}_{it}\right)$are transition probabilities which may depend on classes and time-varying covariates, and $f\left(\mathbf{es}_{it} \mid s_t^d, s, \mathbf{z}_{it}\right)$are indicator distributionswhich may depend on latent states, classes, and time-varying covariates. The distribution of the indicators is modeled in the same way as in Cluster models, where it should be noted that in the default specification indicators are assumed to be independent of latent state s and $\mathbf{z}_{it}$ are mutually independent given s_t^d

$$f(\mathbf{es}_i \mid \mathbf{z}_i) = \sum_{x=1}^{K} \sum_{x_0^d=1}^{K^d} \sum_{x_1^d=1}^{K^d} \dots \sum_{x_T^d=1}^{K^d} P(s \mid \mathbf{z}_i) P(s_0^d \mid s, \mathbf{z}_i) \prod_{t=1}^{T_i} P(s_t^d \mid s_{t-1}^d, s, \mathbf{z}_{it}) \prod_{t=0}^{T_i} f(\mathbf{es}_{it} \mid s_t^d, s, \mathbf{z}_{it}) \quad \dots \quad \text{(M2)}$$

For the classes and initial states here we modeled the logit comparing the probability of moving from origin state r to destination states with the probability of staying in state r, which is referred as a transition logit. The model has a separate γ parameter for each transition logit whereas the parameters for r = s (no transition) are fixed to 0 for identification.

$$\log \frac{P(x_t^d = s \mid x_{t-1}^d = r, x, \mathbf{z}_{it})}{P(x_t^d = r \mid x_{t-1}^d = r, x, \mathbf{z}_{it})} = \gamma_{xrs0} + \sum_{p=1}^{P} \gamma_{xrsp} \cdot z_{itp} \quad \text{(M3)}$$

For modeling details of Markov or (latent transition) models see into (MacDonald & Zucchini, 1997),(F. Van de Pol & R. Langeheine, 1990) Useful books on modeling longitudinal data include (Diggle, Heagerty, Liang, & Zeger, 2002); (Hand & Crowder, 2017).

4.10. Estimation technique

We applied maximum likelihood or posterior mode techniques for estimations by a combination of EM and Newton Raphson iterations. In order to be able to deal with applications involving large numbers of time points (see, e.g.,(Dias et al., 2015).The Estep computations use a generalized version of the forward-backward recursion scheme, also known as the Baum-Welch algorithm, originally proposed by Baum et al. (1970). Details on this generalized Baum-Welch algorithm, which can deal with mixtures, covariates, and multiple response variables,are provided by (Jeroen K Vermunt et al., 2008) .For the Newton-Raphson iterations, analytic first- and second-order derivatives are computed using the forward recursion scheme described (Lystig & Hughes, 2002) based on Baum maximization technique (Baum, Petrie, Soules, & Weiss, 1970).

CHAPTER 5

HANDLING DATA BY MODEL BASED APPROACH

5.1. Introduction

In this chapter we will discuss empirical application employing latent class model-based approach and its variants. Since each variant offers some methodological improvement over the basic latent class cluster model therefore it is inevitable to explicitly present here theoretical structure of basic latent class model followed by its relevance to chosen empirical application of job quality typology building.

Latent class clustering is model based approach to handle data in probabilistic manner. It basically serves to find clusters based on common probabilistic response patterns of the units of data. The units can be micro or macro as well and the data can emerge from any distribution of exponential family. The said method calculates most likelihood of belonging to various clusters where the number of clusters are simultaneously adjusted according to statistic fit (see selection tools in chapter 4).

Latent class cluster models are one type of finite mixture models. The basic underlying assumption of these models is related to the role of latent or hidden variable to determine the observed distribution of the variables. For a given set of observed or manifest variables in Latent class cluster models, we seek for the best model which is parsimonious and classifies the distinction of individuals well. Within the found latent clusters the indicators are bound to come from same distribution. To put it differently latent class clustering is a probabilistic method of unsupervised clustering. Once identified, mathematically, the classes are homogeneous within, but distinct from each other. In these models, once the model has been fitted, the probability of class membership is estimated for each observation in each subclass or cluster. These probabilities can then be used to assign class. It is important to emphasize that in these methods, by probabilistic approach we do not assign individuals to latent classes; rather probabilities are generated for membership in all the identified classes in the model.

The last important assumption is the local independence of observed indicators. This implies the set of variables when diagnosed through their common behavior to uncover the latent construct then the link between these indictors is established through latent variable otherwise, they are least corelated. The said assumption is very restrictive and many real data settings fail to meet such restriction of no codependence. In that case the assumption of local independence can be relaxed by firstly finding the presence of such issue within the mutual pairs. Having said the basics of latent class models we would like to justify here that the theoretical plausibility of job quality to be measured as hidden rather than observed phenomenon is valid in our case. Job quality as discussed in chapter 2 is multidimensional construct and is revealed in literature from subjective and objective measures. There is no perfect tool kit of indictors exist to measure its presence because subjective experiences would remain there influenced by directly observed or unobserved contexts. Henceforth, as highlighted earlier in chapter 2, this specific issue demands more sophisticated and more reality capturing modeling techniques such as latent cluster modelling.

To allow for local dependencies and as well within-class heterogeneity Qu, Tan, and Kutner (1996) proposed a variant of standard LC model labeled as "random-effects Latent class cluster model .The idea behind including continuous factor to account for within class heterogeneity was on finite mixture variants of item response theory models proposed by(Smit, Kelderman, & van der Flier, 2000).Along with these common issue of conditional dependence and within class heterogeneity issue of sparse data is common for relatively big categories of data. Sparseness occurs when the number of observed variables or the number of categories of variables is large. The contingency tables formed for classification under in these cases result in many empty cells compared to the original sample size(Reiser & Lin, 1999) .Sparse data also results when Latent class cluster models are extended to include continuous variables. In this regard Langeheine, Pannekoek, and Van de Pol (1996) suggested to apply parametric bootstrap for estimating p values of model fit statistics. They verified the threat involved on trusting chi-squared distribution for estimations. For assessing model fit in the case of sparse data many researchers including Brian S Everitt et al. (2011a); Lanza et al. (2003) suggest to employ relative model evaluation tools for weighting both model fit and parsimony.

In many practical applications we are interested in using the latent categorical variable for further analysis and exploring the relationship between those variable and other auxiliary observed variables. Auxiliary variables also known as covariates, concurrent outcomes and distal outcomes help us to inquire for the answer of what and why in context of mixture models under study, we can find what kind of individuals belong to which class and the background information for differing response patterns becomes so crucial in some contexts where the co- occurrence of responses is vaguer for unconditional models. So, we can make class or cluster formation conditional to covariates by including their direct effect into class formation or either we can keep their role aside prior to class formation and after this step we can include those variables as predictors. Two types of analyses in this regard are conducted in this section. Firstly, we have added covariates to predict for the hidden (latent) categorical we seek to find out, secondly, we have utilized the same variable to measure its influence on the related dependent (distal) variables.

Observed variable is used as a predictor of the latent categorical variable in first type, for second case on the contrary latent variable is used as a predictor of the distal outcome. The standard way to conduct such an analysis is to combine the latent class model and the latent class regression model or the distal outcome model into a joint model that can be estimated with the maximum-likelihood estimator referred as the one-step method. Such an approach, however, can be flawed because the secondary model could affect the latent class formation and the latent class variable could lose its meaning as the latent variable measured by the indicator variables. Jeroen K Vermunt (2010) pointed out several disadvantages of the one-step method in the context of predictors (covariates) of the latent class variable and as well To answer such drawbacks, an alternative three-step approach has been developed in Vermunt (2010) by expanding on ideas presented in (Bolck et al., 2004).For further details see also (Bakk et al., 2013b). This approach is suitable for exploring relationships between the latent class variable and predictor variables. In this approach the latent class model is estimated in a first step using only latent class indicator variables. In the second step, the most likely class variable is created using the latent class posterior distribution obtained during the first step. In the third step the most likely class is regressed on predictor variables, considering the misclassification in the second step.

The technical aspects we incorporated in models include violation of standard local independence assumption of manifest variables, heterogeneity of observations in mixtures, out of sample performance of most parsimonious model, direct effects of covariates on model results and competing models for including auxiliary information into unconditional models (step 1 vs. step 3 analysis).

We have selected quality of employment (QOE) as a hidden phenomenon which is not perfectly measured by job indicators to utilize the potential of above-mentioned variants. QOE or job quality are used interchangeably here. In this chapter we evaluate various unconditional and conditional models using nls-97 cross sectional survey 2017 for finding typologies of full-time workers' class. Standard latent cluster model is tested against Latent cluster with direct effect of mutual dependence of manifest variables. The inspection is made by two methods: by observing score of "bivariate residuals" and by bootstrapping the bivariate residual. Further we tested heterogeneity of data through inclusion of continuous factors and sparse nature of data led to use parametric bootstrapping methods for selecting the final unconditional model. Conditional models are also tested with direct effect of degree, gender and ethnicity on clustering of job quality. Models with same covariate and distal outcome (job satisfaction) are applied compared by various versions of step 3 approaches.

5.2. Results Discussion

Here we are introducing main models estimated for clustering of American labor class. The sample consists of more than 83 % of full-time workers so this sample is considered as a homogenous one, remaining 17 % are distributed over part time categories. We addressed and tested the existence of possible sub-groups in terms of quality jobs. We are giving here an overview of selected models for answering specific research questions,

1. Is there exist any typologies of job quality for the given labor class?

2.What if the classical assumption of local independence of indicators in selected modeling set up fails for the data; how to address the issue of codependence of indicators?

In next part of conditional models, we answered for

What is the impact of direct effect of covariates in class formation and class size of job quality clusters?

5.2.1. Model Selection Procedure

We have selected 4 cluster case as the most parsimonious model based on lowest BIC AND AIC information criteria from standard unconditional models. Though the standard procedure of addition of one more class through improvement in log likelihood and reduction in BIC led us to select 5 cluster case (table 1), but for that case last 2 classes were negligible in their size, careful inspection of class 4 and class 5 solution showed more clear theoretically distinctive patterns for 4 cluster model. We tested various models and selected model for clusters 4-case based on relative information criteria, since classification statistics are also crucial to account for in our case since one of the objectives of this thesis is to cluster the data for quality of employment .3 cluster case was a competing model for having low classification errors, while class 5 solution was best solution reporting lowest score on relative criterion but features of last 2 classes were overlapping leading to increased classification errors for this solution. Interestingly both models of cluster 3 and cluster 4 were giving theoretically interpretable class differences.

Along with this consideration , considering the sparse nature of data , we calculated bootstrapping BVR p values and compared both 3 class and 4 class models in terms of significant measured associations between residuals of indictors(significant BVR scores), model cluster 4 outperformed model cluster 3 for both criteria of selections, -2LL was significant which implied addition of another cluster improves the model results significantly , and overall reduction in Bivariate residuals was also in big extent from total BVR 56 for cluster 3 case to 16BVR for cluster 4 case (table 1), second table presents summary for the bootstrapping where significant p value indicates adding one more cluster contributes to better classification.

Table 5.1: Unconditional Models Summary

MODELS	**LL**	**BIC(LL)**	**AIC(LL)**	**AIC3**	**Np**	**Max. BVR**	**Class.Er**	**EntroR²**
2-Clust	-45465.2	91316.9	91028.39	91077.39	49	119.5594	0.0737	0.6699
3-Clust	-44817.8	90148.26	89765.54	89830.54	65	56.5412	0.1341	0.6123
4-Clust	-44674.3	89987.53	89510.6	89591.6	81	16.7853	0.1602	0.6188
5-Clust	-44536.5	89838.19	89267.06	89364.06	97	24.7151	0.2293	0.564
3-Clust 1-CFact	-43908.2	88376.49	87958.45	88029.45	71		0.1349	0.5482
4-Clust 1-CFactor	-43787.9	88246.23	87745.75	87830.75	85		0.2304	0.5171

Table 5.2: Unconditional Models Bootstrapping

	LL	BIC(LL)	AIC(LL)	AIC3(LL)	np	Max. BVR	-2LL Diff	Bootstrap	Class. Err.
3vs4cluster	-44642.4	89923.75	89446.82	89527.82	81	25.8	332.78	0.0000*	0.1624
3cluster:Boot	-44808.8	90130.32	89747.6	89812.6	65	52.1			0.1377

5.2.2. Bivariate Residuals Comparison

In the following three tables for bivariate residuals of selected 4 cluster case are reported first table presents BVR scores under standard latent cluster model, it can be seen from table 5.3 that many cross dependencies scores are greater than 2 indicating strong associations between given job features. Since sparse mixed variable data does not follows chi square distribution, therefore the bootstrapped based BVR tables for cluster 4 and cluster 3 are reported successively in next tables (5.4:5.5). In these tables BVR scores are presented with significance scores below. From both tables of 5.3 and 5.5, it is observed that cluster 4 performs better with less significant scores.

Table 5.3: Bivariate Residuals Standard

Indicators	wage	tothr	schedule	medins	paidleaves	location	compsize	unioncv
hrwage	.							
tothrweek	2.5708	.						
schedule	7.8292	9.4462	.					
medins	2.4632	2.3409	2.4997	.				
paidleaves	12.1660	1.8504	0.4880	0.1436	.			
location	4.2271	0.1563	2.2531	0.2970	16.7853	.		
compsize	0.7550	6.5880	8.8762	0.0346	1.9096	14.2956	.	

unioncov	1.7216	0.5249	6.1829	7.3627	16.3696	0.0737	0.2016	.

Table 5.4: Bivariate Residuals By Bootstrapping case1

Dependent	**hrwage**	**tothrweek**	**schedule**	**medins**	**paidleaves**	**location**	**compsize**	**unioncov**	**jasgn**
hrwage	.								
pvalue	.								
tothrweek	0.6611	.							
pvalue	0.394	.							
schedule	7.8412	11.563	.						
pvalue	0.000*	0.000*	.						
medins	1.1193	1.3919	2.0570	.					
pvalue	0.29	0.266	0.068	.					
paidleaves	10.048	3.07197	0.4589	0.1404	.				
pvalue	0.0000	0.054	0.772	0.656	.				
location	4.7650	0.0926	2.3372	0.2014	16.757	.			
pvalue	0.044	0.782	0.086	0.648	0.000*	.			
compsize	2.3119	12.239	9.3088	0.0519	6.5852	25.840	.		
pvalue	0.132	0.002	0.000	0.822	0.01	0.000*	.		
unioncov	4.2630	1.7097	6.2492	7.1487	16.001	0.3741	1.8348		
pvalue	0.06	0.266	0.000*	0.02*	0.000*	0.61	0.226	.	
jassign	0.5441	0.0075	0.5564	0.4123	15.399	0.3779	0.5090	4.2707	
pvalue	0.624	0.998	0.69	0.774	0.000*	0.776	0.726	0.048	.

Table 5.5: Bivariate Residuals by Bootstrapping case2

Dependent	hrwage	tothrweek	schedule	medins	paidleaves	location	compsize	unioncov	jasgn
hrwage	.								
pvalue	.								
tothrweek	0.6611	.							
pvalue	0.394	.							
schedule	7.8412	11.563	.						
pvalue	0.000*	0.000*	.						
medins	1.1193	1.3919	2.0570	.					
pvalue	0.29	0.266	0.068	.					
paidleaves	10.048	3.07197	0.4589	0.1404	.				
pvalue	0.0000	0.054	0.772	0.656	.				
location	4.7650	0.0926	2.3372	0.2014	16.757	.			
pvalue	0.044	0.782	0.086	0.648	0.000*	.			
compsize	2.3119	12.239	9.3088	0.0519	6.5852	25.840	.		
pvalue	0.132	0.002	0.000	0.822	0.01	0.000*	.		
unioncov	4.2630	1.7097	6.2492	7.1487	16.001	0.3741	1.8348		
pvalue	0.06	0.266	0.000*	0.02*	0.000*	0.61	0.226	.	
jassign	0.5441	0.0075	0.5564	0.4123	15.399	0.3779	0.5090	4.2707	
pvalue	0.624	0.998	0.69	0.774	0.000*	0.776	0.726	0.048	.

The novelty of our data handling in case of mixed mode sparse employment data is revealed from next model specification, where we modeled the heterogeneity of individuals and co-association /dependence of main job indicators in the same model. We applied these strategies for comparing the performance of both models for making final choice of unconditional model. Firstly, for addressing heterogeneity of observations we included continuous factor in baseline (Model a), Model a with 1 continuous factor further improved the results (model b), there was size able reduction in BIC and AIC values and increase in log likelihood value (see last row in table 1).

Then with model b for addressing the co dependence of many indicators, we included direct effects for the highly correlated intrinsic job indictors in model b which led to compare us model c with model b. Since the bivariate residuals between these were significantly greater than 2 in model a (see table 5.3) we observed that after including direct effects (mutual association) made many pairs of bivariate residuals insignificant (see table 5.6) in following.

Table 5.6: Bivariate Residuals for model c

Indicators	hrwage	tothr week	schedule	medins	paid leaves	location	comp size	union cov
hrwage	.							
tothrweek	1.0435	.						
schedule	0.8272	**5.732**	.					
medins	0.0035	0	0.8769	.				
paidleaves	0.5824	0.759	0.8119	2.1662	.			
location	0.0197	0.862	1.5592	0.2124	0	.		
compsize	0.1764	**3.602**	0.9039	0.6544	**7.288**	0	.	
unioncov	0.7145	0.271	1.4511	2.0922	0	0.007	0.347	.

The parameters tables for model c are given in appendix A. Since our major focus of LCCA is to discuss possible cluster formation for job quality so we discuss in main results relevant statistics. From parameters inspection we briefly summarize that indicators for each of four classes are significantly different for each of the 4 clusters in all clusters and the intercepts are also different and significant measured through wald statistics. Cross comparisons show all relationship are significantly different across all possible combinations of clusters except very few exceptions. The direct effects reveal that the associations amongst combinations are significant as expected since the co-associations of these indicators were the main reason for significant high BVR values in model A. Compared to BVR of standard model we can see that after inclusion of direct effects of various workplace indicators we are left with only three high level of associations out of 64 pairs of co dependence.

Note :We could further check these BVR S under bootstrapping but we avoided for two reasons , first was the positive chances of further reduction in significant BVR s following theory discussed in (Oberski et al., 2013),according to them if sparse data under asymptotic distributions assumptions gives non-significant BVRs then the bootstrap further reduces the margin of error in calculations of bivariate associations. This could be verified from standard BVR results to BVR p value results for same case of cluster 4. Secondly, we did not want to force our model to be dependence free to incorporate the natural margin of dependence of empirical data keeping the objective of choosing model with least possible classification errors.

5.2.3. Unconditional models

For unconditional models initially based on relative fit criteria, entropy R2, and classification errors, standard Latent cluster class models with conditional independence were tested. Later models were updated with local dependence of many indicators. We joined manifest variables through allowing their mutual impact on the model and through the inclusion of continuous factor.

In the following we detail the output /statistics for the final selected model in unconditional case. In next section of conditional models, we compare the basic selected case with some covariates and compare whether the later model with covariates used as predictor of class membership changes the allocation of individuals across clusters.

In the following tables, class specific response probabilities are expressed for nominal variables and class specific means are given for continuous variable and class specific means plus response probabilities for ordinal variables are calculated in each column of the given tables. Classification is calculated through posterior probabilities. We will not discuss posterior probabilities of each response pattern for comparison and explanation. Since the data for each response patterns of more than 2500 individuals is tedious to inspect in terms of data pattern for each class therefore we picked the explanation of classes/clusters through relative easier ways in below. (For calculation class specific mean probabilities of various kinds of variables, see in technical section of methodology chapter).

5.2.3. 1.Finding the job quality typologies

We explain the categories range to make understand the interpretations. Although the categories are ordinal mostly where each highest category explains the higher value for the indicators compared to previous category. These category labels are based on scores of individuals.To explain how data is arranged we explain working hours variable here. For working hours 1st category consist of 1- 23. Here 23 various scores of working hours reported by individuals for 23 different values of working hours and the 23 categories were arranged in ascending order. Same rule is applied to company size and other variables categories. The mean earnings are not the true depiction of earning variance since averages are skewed statistics.

Table5.7: Model A Posterior Probabilities

	Clus1	**Clus2**	**Clus3**	**Clus4**
Cluster Size	0.4669	0.4304	0.0707	0.032
Indicators				
hourwage				
Mean	3431.007	1900.098	1215.966	10439.81
tothourweek				
1 - 23	0.0681	0.0715	0.5127	0.0172
24 - 26	0.028	0.0285	0.0317	0.015
27 - 27	0.5656	0.5708	0.4078	0.3621
28 - 31	0.1079	0.1073	0.0325	0.0976
32 - 35	0.1354	0.1324	0.0144	0.1849
36 - 57	0.095	0.0894	0.001	0.3232
schedule				
Reg. day	0.8623	0.7433	0.4363	0.7378
Reg. evening	0.0169	0.0341	0.1264	0.0001
Reg. night	0.0341	0.0608	0.0778	0.0154
shift rotates	0.0688	0.1198	0.2674	0.1721
shift splits	0.0088	0.0313	0.0655	0.0067
irregular hours	0.0092	0.0106	0.0265	0.0678
Med.insurance				
No	0.1264	0.2306	0.6072	0.1388
yes	0.8736	0.7694	0.3928	0.8612
paidleaves				
1 - 5	0.0237	0.0711	0.3865	0.0196
6 - 9	0.0597	0.1432	0.3519	0.051
10 - 11	0.0867	0.1664	0.1848	0.0766
12 - 14	0.0488	0.0748	0.0375	0.0445
15 - 15	0.0592	0.0725	0.0165	0.0558
16 - 19	0.1561	0.1529	0.0157	0.1521
20 - 21	0.1083	0.0847	0.0039	0.109
22 - 26	0.1658	0.1037	0.0022	0.1724
27 - 32	0.141	0.0705	0.0007	0.1516
33 - 75	0.1507	0.0602	0.0003	0.1674
Mean	5.6917	4.0946	1.0251	5.9017

location				
No	0.1576	0.3233	0.3409	0.1487
yes	0.8424	0.6767	0.6591	0.8513
companysize				
1 - 10	0.0959	0.1536	0.1295	0.085
11 - 25	0.1079	0.1698	0.1448	0.0956
26 - 45	0.0783	0.1203	0.1043	0.0695
46 - 73	0.1114	0.1647	0.1466	0.0991
74 - 104	0.106	0.145	0.1361	0.0948
105 - 124	0.1239	0.1417	0.15	0.1119
125 - 149	0.1577	0.1047	0.1602	0.1472
150 - 187	0.2189	0.0003	0.0284	0.2968
unioncoverage				
No	0.7917	0.8753	0.9062	0.9851
yes	0.2083	0.1247	0.0938	0.0149
job assignment				
DLI job	0.105	0.0647	0.0313	0.0271
Regular job	0.8919	0.9353	0.9687	0.9729
Military job	0.0031	0	0	0

5.2.3.2. Typology building from baseline model A

We explain the response pattern differences of the four clusters here. Since hourly wage is continuous so mean wage is reported for each class.The lowest proportion of class 4 consists of most likely highest earners we followed its response pattern firstly. Almost 3% highly paid people are more likely to belong to this category who work up-to 120 hours per week with variations of average working hours 36 to 100 plus. Over worked individuals also fall in this group (33%) since the last category of hours indicates working hours. More than 70 % in this data have regular day schedule, this class has more likely cases of 73% individuals who have regular day shift followed by 17 %rotating shifters to 6% irregular hours job doers. For the selected sample since medical insurance is provided to around 80% individuals. Class 4 members have likely 86% achievers in terms of medical insurance provision. Paid leaves which consist of all kind of emergency leaves, for highest number of paid leaves category, this class members have highest number of likely cases of high scores followed by subsequent mid-range categories, the probabilities of belonging to lowest initial 5 categories of upto 15 annual paid leaves is low compared to other classes. For multiple locations of the company around 85 % of this group have yes response, in terms of company size this class is distinguished compared to others, as around 45 % persons have chances to report big firm company size. since surprisingly union coverage to support employees right is not supported for more than 84% people overall, this group scores 98% for no response of union coverage followed by 97% REGULAR job doers, based on these response patterns and relative far better position for earnings, company size and other related job benefits we name this class as achievers.

The second small size class 7 % has lowest average hourly earnings (1215), highest probability of minimum working hours category 51% fall in this class followed by on middle working hours category (hours). For this group more than 56 % individuals have schedule other than regular day. Also, for denial of medical insurance for this group is highest, around 60%, more than 82 % of this group report low scores for all kind of emergency paid leaves as low as 1 to as high 10. Interestingly the company size distribution for this group is scattered up to second highest group and multiple locations of the company stands positive for 65% of individuals.

For cluster 3 (size) we have mean wages of 1900 and total hour per week are splitted majorly for mid category of more than 37 working hours to the max of over working hours of 100 plus are done by around 8% of the group members. Paid leaves score is not good for almost 50 % and other 50% is distributed unevenly for improved category of paid leaves, for highest company size category none of the group member belongs.

For cluster 1 (size) which is biggest in size, mean wages are second highest, hours worked per week are in mid category for more than 56 % followed by excessive working hours by other group members, mainly schedule is regular day.87% have yes to medical insurance, and more than 65% have paid leaves as low as 20 and as high as 75 are enjoyed by 15%. More than 84 % persons working in multiple location working firms, and 21% are working have huge firm sizes followed by good company size of more than 4000 for around 40% .We observe that this group is closely performing to the class 4 in terms of average probability distribution of individuals in many indicators but the stark difference of average earnings poses question for finding some predictors of the observed patterns for this groups and the competing class 1 (conditional analysis might help in this regard in next section).

Though we can vaguely label the classes on observed highest reported response patterns (in summary cluster 4 positions highest in terms of acquiring job related rewards followed by cluster 1.Cluster 3 performs poor in many job quality indictors whereas the responses for cluster 2 are ambiguous for many indicators in terms of QOE) but we hold to label the groups on the basis of their likely class memberships and response patterns until going through the improved models with continuous factor which addresses the heterogeneity in data and the merged model which included dependence of observations into account.

Table5. 8: Model B Probability Scores

	Clus1	Clus2	Clus3	Clus4
Clus Size	0.38	0.374	0.2321	0.0139
variables				
hrwage				
Mean	2534.97	1731.88	4404.87	14944.1
tothrweek				
1 - 23	0.1049	0.1137	0.0724	0.0135
24 - 26	0.0288	0.0295	0.0253	0.011
27 - 27	0.5608	0.567	0.5221	0.2755
28 - 31	0.1019	0.1004	0.106	0.0824
32 - 35	0.1231	0.1176	0.1461	0.1801
36 - 57	0.0805	0.0718	0.128	0.4376
Mean	42.0544	41.7013	43.7041	51.2755
schedule				
reg day	0.7583	0.7368	0.8838	0.574
reg evening	0.0531	0.0301	0	0.0002
reg night	0.1057	0.0141	0.0098	0.0327
shift rotates	0.0526	0.1671	0.0953	0.2605
shift splits	0.0192	0.0368	0.0046	0.0219
irregular hrs	0.0111	0.0151	0.0066	0.1107
medins				
No	0.1195	0.3316	0.1501	0.0993
yes	0.8805	0.6684	0.8499	0.9007
paidleaves				
1 - 5	0.0467	0.1176	0.0322	0.0305
6 - 9	0.0878	0.1787	0.0648	0.0619
10 - 11	0.1073	0.1767	0.0848	0.0819
12 - 14	0.0544	0.0724	0.046	0.0448
15 - 15	0.0615	0.0662	0.0557	0.0548
16 - 19	0.1532	0.1333	0.1487	0.1477
20 - 21	0.1009	0.0711	0.105	0.1053
22 - 26	0.1472	0.0838	0.1641	0.1661
27 - 32	0.1194	0.055	0.1426	0.1458
33 - 75	0.1218	0.0453	0.156	0.161

Mean	5.1636	3.541	5.6624	5.7289
location				
No	0.1655	0.3595	0.1876	0.0521
yes	0.8345	0.6405	0.8124	0.9479
compsize				
1 - 10	0.0904	0.1794	0.0885	0.0562
11 - 25	0.1031	0.1958	0.1009	0.0643
26 - 45	0.0763	0.1362	0.0746	0.0478
46 - 73	0.1118	0.1807	0.1094	0.0706
74 - 104	0.1121	0.1482	0.1099	0.0719
105 - 124	0.143	0.1196	0.1405	0.0951
125 - 149	0.1931	0.04	0.1913	0.1432
150 - 187	0.1702	0	0.1848	0.451
Mean	1089.35	95.0914	1164.43	2523.46
unioncov				
No	0.7424	0.9344	0.8488	0.965
yes	0.2576	0.0656	0.1512	0.035

By observing the above table, we are answering how classification of individuals was affected by incorporating heterogeneity in the baseline model. We observed that Cluster sizes and allocation of responses were changed by incorporating the assumption of possible heterogeneous response patterns within other wised assumed homogenous population.

We highlight briefly differences in average profiling of individuals compared to model a detailed pattern; class 4 stands highest earner again schedule mix is more clearly inclined towards rotating hours and irregular hours compared to baseline class 4. More evenly distributed paid leaves are observed though highest probabilities remains for highest number category of paid leaves, still company size remains same distributed and other indicators also score somehow same for this class. Class 3 is now second highest earners under this model assumptions , interestingly class sizes changed but the response patterns for particular earning category remained somehow same , here class 3 is second highest earners and their response pattern somehow matches to second highest earners under class 1 label of previous model, in terms of paid leaves this class matches the response pattern of highest earners though , one clear distinction is of the company size response average probes compared to class 1 of base line model. There we had almost same company size for this group to the highest earners. We find major change in similar patterns of highest earners and second highest earners .Size of the other two classed belonging to lowest and middle weekly earnings has increased from to

37 % compared to baseline model , while interestingly the associated response patterns remained somehow same to the baseline model.

5.2.3.3. Probability scores of model C

We further compared the profiles of model B and model C to account for dependence between the relevant job indicators, for model c we could see improved model fit and significantly reduced BVR in previous section. We compared the class profiling change in terms of size and responses that remained almost same compared to model b for model c. By row-to-row examination of ML based probability scores for both models b and c we concluded that response patterns of clusters remained same conditional to mean earnings in each case. There for we do not provide the detailed table c of probabilities here. Difference in size of clusters to baseline model A was noticeable for model B and model C. Also, regarding response patterns observed in baseline model, the lowest mediocre or highest earning class characteristics remained same though sizes shifted for models B and C. So, after careful examination of commonalities across the models B and C following the detailed response patterns conditional to mean earnings, we name cluster 4 as achievers cluster 3 as competitors cluster 1 individuals as strugglers and cluster 2 individuals as left ones. We built typology based on the response patterns for the considered employee as Strugglers, left-ones, competitors and achievers.

5.2.3. Conditional models

Next, we run conditional models and included various covariates in different sections. Urban location, family size and education level were tested in first set, but urban locations household size turned to be insignificant for differentiating the clusters therefor we considered more closer socio-economic predictors of social and (economic)job quality. Gender, ethnicity and degree were significant for 3 cluster case as well as for cluster 4. Cluster 4 again performed better so that model served as the baseline conditional model.

Table 5.9: Conditional Models Summary

Step 1 conditional models	LL	BIC(LL)	AIC (LL)	AIC3 (LL)	N	Class.Err.	Entropy R^2
3-Cluster (model A)	-44027.814	88615	88197.63	88268.63	71	0.1134	0.6696
3-Cluster (model B)	-44321.204	89202.32	88784.41	88855.41	71	0.1065	0.6853
model model A+Covariates	-44104.423	88918.5932	88388.85	88478.85	90	0.17	0.66

5.2.3.1. Step 1 Results

We observed specifically that inclusion of covariates in the model as direct effects (step 1 approach for including auxiliary variables) altered the distribution of members of cluster or not, it was marked that distribution was changed for some indicators and the class size was affected but Inclusion of predictors helped to justify the differentiated response patterns along all clusters (achievers, competitors, strugglers, left ones).

Patterns were more differentiated in terms of main indicators of varying nature such as total hours week, paid leaves, company sizes , medical insurance under this frame work. There was observed change somehow in sizes of clusters 2, 3 and 4 whereas cluster 1 remained of same size compared to model A .Further, the profile pattern was switched between cluster 1 and 2 compared to model A. Clusters 4 (successful) remained same with similar job features under modal A whilst instead of cluster 1. In this case, candidates of cluster 2 can be labeled achievers on observed response patterns since this class has highest proportions of professional and PhD degree holders. The left behind category under step 1 approach again consists of cluster 3 individuals who around 80% are not even primary educated. In strugglers group (cluster 1) on the other hand around 50% were of same low education profile but the remaining individuals were spread over higher education categories. For this sample, education level emerged as a strong class predictor of highest profile jobs. There were some gender differences across clusters; male proportion was highest for all three clusters except for the left ones though data was almost equally sampled. With relation to racial variable since black and Hispanic consist of total 24 % of racial groups it was observed from the given table the highest likely proportion of these two groups belonged to left ones and strugglers.

Table 5.10: Step 1 Probability Scores

	Clus1	Clus2	Clus3	Clus4
Clus Size	0.4826	0.2974	0.1965	0.0235
Indicators				
hrwage				
Mean	2212.38	4101.72	1421.77	11686.3
tothrweek				
1 – 22-	0.0848	0.0596	0.2054	0.0157
23 - 25	0.0282	0.0248	0.0347	0.0129
26 - 26	0.5625	0.5215	0.5889	0.3203
27 - 30	0.1065	0.1095	0.0816	0.0922
31 - 34	0.132	0.1533	0.0696	0.188
35 - 56	0.086	0.1314	0.0198	0.3709
Mean	42.5206	44.0173	38.9437	49.9504
schedule				
reg day	0.7776	0.9008	0.6028	0.6459
reg evening	0.035	0	0.0746	0.0001
reg night	0.0747	0.0062	0.05	0.0188
shift rotates	0.0824	0.0787	0.2041	0.2034
shift splits	0.0223	0.0031	0.051	0.0339
irregular hrs	0.008	0.0112	0.0175	0.0978
medins				
No	0.1482	0.1487	0.4452	0.0674
yes	0.8518	0.8513	0.5548	0.9326
paidleaves				
1 - 5	0.0517	0.0225	0.1899	0.0272
6 - 9	0.1012	0.0514	0.2568	0.0602
10 - 11	0.1251	0.0742	0.2192	0.0841
12 - 14	0.0626	0.0434	0.0758	0.0475
15 - 15	0.0682	0.0552	0.057	0.0585
16 - 19	0.1611	0.1521	0.093	0.1561
20 - 21	0.099	0.1091	0.0395	0.1083
22 - 26	0.1343	0.1728	0.037	0.1661
27 - 32	0.1013	0.1521	0.0193	0.1415
33 - 75	0.0954	0.1673	0.0125	0.1506
Mean	4.8214	5.8979	2.3786	5.6844
location				
No	0.2385	0.1752	0.361	0.1653
yes	0.7615	0.8248	0.639	0.8347
compsize				

1 - 10	0.1152	0.102	0.1779	0.0824
11 - 25	0.1291	0.1144	0.1934	0.0926
26 - 45	0.0938	0.0833	0.1347	0.0676
46 - 73	0.1329	0.1185	0.178	0.0966
74 - 104	0.1251	0.1124	0.1456	0.0923
105 - 124	0.1415	0.1294	0.1198	0.1083
125 - 149	0.1584	0.1527	0.0506	0.1354
150 - 187	0.104	0.1872	0	0.3249
Mean	720.897	1153	101.375	1863.12
unioncov				
No	0.8125	0.8239	0.9239	0.9804
yes	0.1875	0.1761	0.0761	0.0196
j-assignment				
DLI job	0.101	0.0815	0.036	0.0004
Regular job	0.8978	0.9155	0.964	0.9996
Military job	0.0012	0.0029	0	0
Covariates				
degree				
0 - 1	0.1085	0.0174	0.2709	0.009
2 - 2	0.4289	0.1168	0.5492	0.0059
3 - 3	0.1146	0.0523	0.0883	0.0312
4 - 4	0.2643	0.453	0.0814	0.2299
5 - 7	0.0836	0.3605	0.0102	0.724
Mean	2.7644	4.1246	1.9195	5.6214
race				
Black	0.1414	0.0485	0.2167	0.1356
Hispanic	0.142	0.0684	0.162	0.063
Mixed-Race	0.008	0.0093	0.0117	0.0019
Non-Black	0.7086	0.8738	0.6096	0.7995
Mean	3.2838	3.7084	3.0141	3.4653
gender				
male	0.5341	0.5927	0.4582	0.6607
female	0.4659	0.4073	0.5418	0.3393

5.2.3.2. Covariate Impact on clusters

Table 5.11: Parameters by Step 1 Analysis

Intercept	Cluster1	Cluster2	Cluster3	Cluster4	Wald	p-value
	3.2340	-0.5907	3.2991	-5.9424	110.7681	7.5e-24
Covariates	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Wald**	**p-value**
degree						
	-0.5937	0.4101	-1.2965	1.4801	263.0626	9.7e-57
race						
	-0.0770	0.3082	-0.2334	0.0022	46.8764	3.7e-10
gender						
	0.5511	-0.4814	1.3264	-1.3961	59.9739	6.0e-13

Here we highlight group differences in terms of covariates. It can be seen from the above reported table that each covariate is significant across the clusters, and the predictors affect differently the strugglers, competitors, achievers and left ones. The intercepts are quite different and of varying nature for each section of employed class. Even if we retake the likely membership of individuals based on covariate role in class formation, the difference of patterns for allocated indicators are somehow behaving in same fashion. Degree is most positively influencing job achievers, whereas race has negligible role for that cluster. Role of gender is negative for this cluster compared to opposite role of gender in being member of vulnerable group. Race and degree have negative association for the group of leftones for their current position in working class.

Next, we applied step3 approach for comparing it to the results of step1 approach, the covariates were added in third step after calculating posterior membership of individuals from unconditional model (steps explained in step 3 modeling in chapter 4).

Inclusion of covariates/external variables in many scientific studies is justified by theoretical reason independent of the modeling options and advantages described for step approaches. In this set up of job quality typologies or segregation of job market we were theoretically convinced to include most relevant possible socio-economic indicators for finding some logic behind distinctive job profiles, we were interested to relate the job profiles and classes to more natural covariates, so we included these in step 1 of class calculation and compared that either any change in composition or allocation of classes emerged. Does predictors changed the typologies, were there significant gendered or racial differences across clusters were answered through reading the profiles of each class conditional on gender, race and degree level.

The conclusion was drawn that in such data ,where theoretically inclusion of covariates or exogenous variables can guide for better pattern revealing or co-occurrences of responses , then the phenomenon which have direct and theoretically logical predictors should be explored with inclusion of all those possible predictors , this approach labeled in literature as step 1 approach to handle auxiliary information is much debated and somehow controversial B. Muthén and Asparouhov (2002b) has favored it for being less biased whilst many studies vote for alternative strategies to include role of covariates or resultant variables from latent clustering (Jeroen K Vermunt & Magidson, 2021).

5.2.3.3. Step 3 Analysis for Covariates

We selected step 3 model based on relative information criteria since absolute fit statistics were valid for these models as the data size was quite smaller, so we compared the models in terms of absolute fit and relative fit, we tested ML and BCH correction-based versions of step 3 analysis for both covariates (gender, race and degree) and distal outcomes (job satisfaction). We compared the results in model and proportional allocation of values. For covariates ML based results were good fit for the data and for distal BCH modification resulted better (see table 5.12.). First four models are covariate cases (ML and BCH proportional and modal variants) and last 2 cases are for distal (ML and BCH modal variants). CML modal is most parsimonious for covariates DBCH is for distal case. P values and degree of freedom are not reported for sparse tables.

Table5.12: Step 3 Summary for Covariates and Distal Outcomes

models	LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar	L^2	df	p-value	Class.Err.
CML	-2415.24	5114.369	4902.47	4938.47	36	73.6953	126	1	0.3503
CBCH	924.8387	-1565.78	-1777.68	-1741.68	36				0.3469
CMML	-2245.32	4774.536	4562.637	4598.637	36	123.3936	126	0.55	0.3515
CMBCH	1455.019	-2626.14	-2838.04	-2802.04	36				0.3476
DML	-5893.58	11842.38	11801.17	11808.17	7				.
DBCH	-5887.05	11829.31	11788.09	11795.09	7				.

Table5.13: Conditional Probability based membership

	Competitors	Strugglers	Left ones	Achievers
Overall	0.4427	0.4550	0.0723	0.0300
Covariates				
degree				
0	0.0622	0.7045	0.2158	0.0175
none”	0.1658	0.6616	0.1726	0.0000
GED	0.2873	0.6057	0.1057	0.0013
High School diploma	0.3872	0.5265	0.0747	0.0117
Bechlors	0.6447	0.3064	0.0111	0.0377
Masters”	0.8160	0.1393	0.0005	0.0441
PhD	**0.9563**	0.0220	0.0000	0.0217
Professional degree	**0.3534**	0.0000	0.0000	**0.6466**
race				
Black	0.3020	0.5343	0.1433	0.0204
Hispanic	0.3887	0.5258	0.0667	0.0188
Mixed Race (Non-Hispanic)	0.5973	0.3063	0.0964	0.0000
Non-Black / Non-Hispanic	0.5195	0.3976	0.0445	0.0385
gender				
male	0.4512	0.4657	0.0448	0.0383

female	0.4339	0.4439	0.1009	0.0213

The noticeable differences are linked to education differences across clusters. Highest likely cases of individuals holding professional degree are associated to achievers' groups followed by successful having highest proportion of doctorates and Bachelors. The important conclusion to support the inclusion of covariates at first step of analysis (class formation) proves better.

Table 5.14: Job Satisfaction Cluster Profiles

	Competitors	strugglers	leftones	achievers
Overall	0.4498	0.4482	0.0738	0.0282
Variables				
Job satisfaction				
Like very much	0.4854	0.4169	0.0632	0.0346
Like fairly well	0.4428	0.4556	0.0754	0.0262
Think it ok	0.3994	0.4921	0.0889	0.0196
Dislike somehow	0.3562	0.5257	0.1037	0.0145
dislikevery much	0.3142	0.5555	0.1196	0.0106

Since achievers constitute the smallest proportion of data, so the reported average likely responses of this cluster are low compared to other dominant categories of successful and strugglers. Successful report highest job satisfaction levels followed by cluster strugglers who have higher proportion of likely cases for lowest satisfaction levels (last categories of job satisfaction) compared to successful. Left ones have relatively more likely dissatisfied or neutral respondents with job quality.

5.3. Conclusion of Analysis

We have taken employment data for the year 2017 from ("National Longitudinal Surveys (NLS)"; Moore et al., 2000). Since the labor market is of developed country so initially, we expected some indicators to be highly positive for considered employed class. The nature of data was mixed mode, we could standardize the variables but for performing categorical analysis it was better capture the indicators through categories. Ideally, we could increase categories for discrete indictors, but that was complicating the visual inspection of outcomes (see Appendix A). We have taken some crucial indicators to judge state of jobs experiences later we took subjective evaluation of quality jobs as a distal outcome for evaluated clusters then the key covariates were tested as predictors of quality heterogeneity in step 1 analysis of covariates. We compared the case with no covariates (unconditional modal A) and found the change in class belonging of individuals. Further step 3 analyses supported the significant contribution of degree level, race and gender for clusters achievers, successful, strugglers, and left ones. We found some evidence of more prevalence of nonwhite and females in left ones and different degree of satisfaction scores from jobs for most successful to others.

In doing all empirical analysis we were keen about some technical issues related to mixture models generally and some specific to LCCA. We discussed and tested on our data for the violation of conditional independence assumption. For this we initially took bivariate residuals based on chi2 as diagnostic measure. Since our data is sparse which does not follows chi 2 distribution therefore for the most parsimonious model selected from bootstrapping difference log likelihood of nested models we calculated asymptotic bivariate residuals by bootstrapping, and after confirming a significant reduction in dependence of observations we incorporated the heterogeneity of observations through inclusion of continuous factor in final model and lastly for this model we again sorted out for reported co dependence of some indicators for the last model through inclusion of direct effects Validation of final models was also tested. Lastly, we applied comparison of various approaches to include the role of covariates or predictors in cluster formation, most parsimonious model was selected and elaborated.

CHAPTER 6

HANDLING DATA BY REGRESSION MIXTURES

6.1. Introduction

Regression mixtures were primarily presented in order to better accommodate for the overlooked heterogeneity in the population (Agresti et al., 2000).The specific mixture is another variant of finite mixture family with some distinctive features. As the name suggests this is the regression technique therefore it serves the basic purpose of regression scheme. It is utilized to measure quantitative or qualitative differences in parameter of interest shaped by the influence of chosen predictor in more than one groups. The effect sizes across different groups are measured here. The underlying assumption of regression mixtures is that everyone is not influenced in the same way by the same variable. Since the effect of the predictor could be same, slightly different or totally different across groups therefore various number of groups or class combinations can be tested and compared following model selection tools adopted in mixture modelling such as described in chapter 4 (see chapter 4).

In context of mixed mode data, the assumption of data emerging from various distributions is more relevant and the task of un-mixing the distributions is theoretically more comprehensible since the sampled population is inherently measured at different scales. Through generalized linear models we can model dependent variables as a function of explanatory variables using link functions described for various discrete and continuous members of exponential family Skrondal and Rabe-Hesketh (2004) but for mixed mode data(or for such a varied kind of variables) one aggregate regression fit may mask the diversified behavior of the individuals .The aggregate regression solution is also inadequate for the mix of population in general if we hypothesize unknown classes in data. Contrary to standard regression framework we can suppose a mix of subgroups to explain the structure of the overall population under mixture regression framework. To incorporate qualitative differences in the effects of a predictor variable on an outcome and vice versa to measure heterogeneity in class effects we employed latent class regression mixture (LCRM) as an explicit type of mixture model in this thesis. Regression mixture approach allowed for simultaneous estimation of regression equation after classifying individuals into distinctive classes which is of great use within the economic context of finding possible heterogeneous effect sizes in otherwise considered same effects for all. Clusters is used to label groups in regression mixtures by Brian S Everitt et al. (2011a), but we will go with the ‘class’ label for regression models following (Jeroen K Vermunt & Magidson, 2013).Regression mixture (RM) is in fact a nonparametric random-effects model (Simonoff, 2003; Skrondal & Rabe-Hesketh, 2004; Tuma & Decker, 2013).

Regression mixtures adhere to the complexities inherent in real life data measurements (Qu et al., 1996). Insignificant predictors are the general issue with the mixtures of regressions (Jeroen K Vermunt & Liesbet Van Dijk, 2001).In chapter 5, we have focused in nonparametric mixture models J. K. Vermunt and L. Van Dijk (2001)for a case study of exploring typologies of employment quality. We have addressed many methodological considerations relevant to model building in each economic context and as a byproduct of such modeling we were able to meet our theory related queries of differences in quantitative and qualitative terms for the chosen employee class. Regression models explained in this section share one common feature to the previous one. They are applied in mixtures latent framework. They are based on categorical latent variable hence the analysis is most suitable for discrete variables. Since the nature of data and latent variable is considered discrete therefore the regression mixture is labelled as latent class regression model. These models can additionally find effects differences across classes. Since these models relax traditional ordinary least squares assumptions therefore units of measurement under these models are flexible enough to incorporate mixed nature of structure if encountered. Technical background of these regression models is explained in technical section 4.3 of chapter 4, investigate regression subsection in chapter 2 for literature review.

6.2. Data Description

To complement the analysis done in previous section we have added some crucial differences of job nature for exploring their impact on job satisfaction(which is regarded as subjective indicator of job quality in this context (Clark, 2005).For measuring the differential impacts fruitfully we have taken the repeated sample of over 9 years(2000 to) from national survey of British households ("Understanding Society", 2021) .For technical details of sampling and survey design see also into (Buck & McFall, 2011). The rationale for choosing this sample is linked to availability of variables of interest over the time. The specific longitudinal design of the survey allows to access age, degree level, occupation categories, working hours length, organization size and work schedules. Here relatively different indictors (compared to analytical framework presented in previous chapter) are opted to address differential effects of job quality variation. We have intentionally taken repeated measured data to represent another case study for bringing little variety to data set ups. The concept of job quality is taken as indicated through subjective indicator of job satisfaction contrary to the previous section where various objective indicators were combined to explore regarding it. Here latent class regression mixtures (LCRM) are applied to test for any possible source of unseen heterogeneity in job quality within the sampled observations. The data set spans over 9 years and some subjective and objective indicators of QOE are picked to address the extent of difference in their impact on job quality. We have incorporated standard and nonstandard work arrangement measuring variable, as well as another major disaggregation of sample based on full and part time. Additional important predictors included for measuring their varying effects on job satisfaction were hours worked in week and company size.

How are specific dimensions of job quality linked to job satisfaction in various classes of workers? To explore this query we employed LCRM to find the best combination of classes for explaining differential impacts of work related features such as job nature (full time or part time), work arrangements (standard or nonstandard), job size, working hours on a diverse longitudinal sample of more than 8600 individuals. Class 2 solution was identified compared to 3-class solution by bootstrapping and other relative fit criteria(AIC3,BIC) suggested for particularly regression mixtures (Wedel & Kamakura, 2000). The resultant latent classes of satisfied vs. non satisfiers were responding significantly different to above mentioned indicators.

Job satisfaction is generally perceived to be influenced by core job quality features. As in previous chapter we could see the evidence of job-related indicators for profiling the job quality status of the chosen labor class. In a different modeling environment, we have tried to explore the links of subjective job quality indicators (job satisfaction level) with some objective job characteristics mentioned above. We have not taken the lengthy list of predictors in this modeling scenario since the data consists of repeated measures and there were distinctive groups with stark age and working status differences, therefore there were quite missing information when we attempted to include more work-related features. We limited the scope of variables also to measure the differences for better understanding the extent and source of differences with possible less complicated cross classifications.

In this section to meet following study objectives relevant to regression framework we will build and test empirical economic models with repeated measured labor market data:

1. To compare and evaluate different models for finding differential effects of job-related features impacting job quality.

2. To verify the differing impact of auxiliary variables across different classes by testing for the possible sources of heterogeneities in given data.

Table6.1. Summary of Unconditional Regression models

LCRM CASES	LLHD	BIC(LLH)	AIC(LLH)	AIC3(LLH)	npr	df	C.err.	EntrpR²
Basic2clu	-91172.6	182553.6	182391.1	182414.1	23	8630	0.0904	0.698
basic 3clu	-89185.1	178687.6	178440.3	178475.3	35	8618	0.1393	0.6828
2clurestrictd	-91177.2	182544.7	182396.4	182417.4	21	8632	0.0904	0.698
3clurestrictd	-89189.5	178669	178442.9	178474.9	32	8621	0.1393	0.6828

6.3. Rational behind Model Selection

From the above table we can see that first 2 models show base line 2 classes and 3 classes where the model is improved on the basic of higher log likelihood and lower information loss criteria. Further classes were not tested since beyond 3 classes class separation indicators were performing low and initially diagnostic of data revealed that there was not much spread across mediocre categories of job satisfaction .Though the case for 2 classes(Basic 2 cluster model) was most parsimonious with 23 parameters and 0.1 % classification error and relative highest score 69 % for entropy R2 but the reason for choosing it as baseline model was its higher theoretical interpretability compared to additional class case(Basic 3 cluster model). Since initial diagnostic of data showed that responses for satisfaction score were skewed and more responses were concentrated towards majorly last four categories on ordinal scale (variable details given in Appendix B).

We inspected the parameters effects across both classes followed by profiles divisions across job satisfaction categories. Since the dependent indicator had 7 categories where first 5 categories implied somehow being unsatisfied or neutral with your job and the last three categories clearly indicated for being satisfied and fully satisfied. Around 40% of subjects belonged to varying levels of being unsatisfied led by moderate satisfiers and most job satisfiers with work. The high-level satisfiers were scattered over remaining 60% division of data. Initially with no covariates class 2 models was intuitively more appealing since the pattern of being satisfied or unsatisfied were broadly divided clearly for this model. The results were further endorsed by lowest classification errors for this model.

After choosing class 2 model (Basic 2 Cluster) we found few predictors insignificant (jbft_dv) so we dropped that from further analysis and estimated the restricted versions of 2 class solution and 3 class solutions also) with further restrictions driven by economic theory. Restricted version of 2 classes was chosen as a final model for classification and validation (details of variables are provided in Appendix B).

From the table given below we can see that the parameters across the two classes are significantly different except for jbft_dv(full time job or part time job) .For other each job satisfaction predictor the p-value is found to be less than .05 implying the null hypothesis stating effects associated with that predictor are zero would be rejected. Thus, for each predictor, information of the response for that predictor contributes significantly to differentiate between the job satisfaction classes.

Table 6.2. Parameters of Unconditional Regression Model

classes	indicators			coeff	s.err	z-val	pval
Class (1)	1			0.02	0.03	0.695	0.49
Class(2)	1			-0.02	0.03	-0.695	0.49
jbsat(completely dissatisfied)		|	Class(1)	-1.2	0.26	-4.518	6.20E-06
jbsat(mostly dissatisfied)		|	Class(1)	-0.5132	0.1746	-2.9399	0.0033
jbsat(somewhat dissatisfied)		|	Class(1)	0.28	0.0902	3.184	0.0015

jbsat(neither satisfied or dissatisfied)		|	Class(1)	0.33	0.02	11.41	3.60E-30
jbsat(somewhat satisfied)		|	Class(1)	1.13	0.08	12.76	2.70E-37
jbsat(mostly satisfied)	1	|	Class(1)	0.9998	0.1839	5.4357	5.50E-08
jbsat(completely satisfied)	1	|	Class(1)	-1.0341	0.2633	-3.927	8.60E-05
jbsat(completely dissatisfied)	1	|	Class(2)	-2.4876	0.2778	-8.9532	3.50E-19
jbsat(mostly dissatisfied)	1	|	Class(2)	-2.1701	0.1955	-11.098	1.30E-28
jbsat(somewhat dissatisfied)	1	|	Class(2)	-1.1408	0.1144	-9.9702	2.10E-23
jbsat(neither satisfied or dissatisfied)	1	|	Class(2)	-0.6041	0.0563	-10.7242	7.80E-27
jbsat(somewhat satisfied)	1	|	Class(2)	1.2	0.0994	12.0749	1.40E-33
jbsat(mostly satisfied)	1	|	Class(2)	2.8047	0.1821	15.402	1.60E-53
jbsat(completely satisfied)	1	|	Class(2)	2.3978	0.2782	8.618	6.80E-18
jbsat	jbsize	|	Class(1)	-0.0068	0.0037	-1.8222	0.068
jbsat	jbsize	|	Class(2)	-0.0226	0.0041	-5.5589	2.70E-08
jbsat	jbterm_dv	|	Class(1)	-0.0031	0.009	-0.3427	0.73
jbsat	jbterm_dv	|	Class(2)	-0.0474	0.0112	-4.225	2.40E-05
jbsat	jbhrs	|	Class(1)	0.0015	0.0014	1.0134	0.31
jbsat	jbhrs	|	Class(2)	-0.0042	0.0014	-2.9078	0.0037
jbsat	jbft_dv	|	Class(1)	0.0052	0.0305	0.1714	0.86
jbsat	jbft_dv	|	Class(2)	0.0095	0.0368	0.2587	0.8
jbsat	hiqual_dv	|	Class(1)	-0.0107	0.0052	-2.0664	0.039
jbsat	hiqual_dv	|	Class(2)	0.0302	0.0079	3.8096	0.00014

6.3.1. Why was Restricted Model Selected?

For the unrestricted model Classification statistics revealed model as a good fit. We imposed further certain order restrictions to make our model more parsimonious after deleting insignificant predictor. Driven by economic theory role of hours and pay is described as positive for boosting employees moral Malik, Danish, and Munir (2012); Wanger (2017) so we imposed the increasing restrictions on both predictors. That implied to test for the hypothesis; People with higher earnings and full-time work situation are more satisfied with their jobs job satisfaction. The restricted versions for 2 and 3 classes (see table 6.1 2clurestrictd and 3clurestrictd) were tested. Results for these models in terms of lower value of information criteria (BIC and AIC3) remained same compared to basic cases also th no mark-able change in classification error and entropy R2 was observed. We preferred the restricted 2 class for being most parsimonious and equivalent in terms of lowest classification error.

Table6.3. Regression Parameters of Restricted 2 Class Model

Regression Parameters							
term				coef	s.e.	z-value	p-value
Class(1)	1			0.0264	0.0388	0.6816	0.5
Class(2)	1			-0.0264	0.0388	-0.6816	0.5
jbsat(completely dissatisfied)	1	|	Class(1)	-1.2366	0.1626	-7.6073	2.80E-14
jbsat(mostly dissatisfied)	1	|	Class(1)	-0.5351	0.1071	-4.9944	5.90E-07
jbsat(somewhat dissatisfied)	1	|	Class(1)	0.2767	0.0568	4.8699	1.10E-06
jbsat(neither satisfied or dissatisfied)	1	|	Class(1)	0.3319	0.0291	11.4131	3.60E-30
jbsat(somewhat satisfied)	1	|	Class(1)	1.1432	0.0535	21.3722	2.40E-101
jbsat(mostly satisfied)	1	|	Class(1)	1.0218	0.1158	8.8245	1.10E-18
jbsat(completely satisfied)	1	|	Class(1)	-1.0017	0.1631	-6.1411	8.20E-10
jbsat(completely dissatisfied)	1	|	Class(2)	-2.5438	0.1785	-14.2502	4.50E-46
jbsat(mostly dissatisfied)	1	|	Class(2)	-2.2081	0.1231	-17.9425	5.50E-72
jbsat(somewhat dissatisfied)	1	|	Class(2)	-1.1595	0.0826	-14.0439	8.40E-45
jbsat(neither satisfied or dissatisfied)	1	|	Class(2)	-0.6029	0.0562	-10.7358	6.90E-27
jbsat(somewhat satisfied)	1	|	Class(2)	1.2185	0.0711	17.1488	6.40E-66
jbsat(mostly satisfied)	1	|	Class(2)	2.8419	0.1044	27.214	4.40E-163
jbsat(completely satisfied)	1	|	Class(2)	2.4539	0.159	15.4375	9.20E-54
jbsat	jbsize	|	Class(1)	-0.0068	0.0037	-1.8309	0.067
jbsat	jbsize	|	Class(2)	-0.0227	0.0041	-5.5816	2.40E-08
jbsat	jbterm_dv	|	Class(1)	-0.003	0.009	-0.3368	0.74
jbsat	jbterm_dv	|	Class(2)	-0.0471	0.0112	-4.2218	2.40E-05
jbsat	jbhrs	|	Class(1)	0.0013	0.0011	1.2427	0.21

jbsat	jbhrs	|	Class(2)	-0.0044	0.0011	-4.17	3.00E-05
jbsat	hiqual_dv	|	Class(1)	-0.0106	0.0052	-2.0565	0.04
jbsat	hiqual_dv	|	Class(2)	0.0303	0.0079	3.8435	0.00012

From the above table we can see that the intercept for each category of job satisfaction for both classes are significantly different. For negative feedback categories consisting of 'completely dissatisfied' and 'mostly dissatisfied' is negative for both classes and for the categories of somewhat dissatisfied 'neither satisfied or satisfied' and 'somewhat satisfied' it stands out to be positive for class 1 and negative for class 2 implying initially class 1 contains more respondents with low satisfaction levels compared to class 2 which consists of more respondents with highest satisfaction on job levels. For extreme positive categories of satisfaction with job for class 2 we have highly significant and high size of initial response for these categories compared to somewhat contrary response for extreme satisfaction level for class 1 individuals.

The beta parameter for each predictor is a measure of the influence of that predictor on jobs satisfaction. The beta effect estimates under the column labeled class 1 suggest that class 1 is less likely to be influenced by organization size (job size) working hours and qualification. Class 1 and class 2 both are not influenced by job size (beta is approximately 0). Job size appears to be insignificant predictor for job satisfaction in case of class 1 and significant for class 2). Why we have reported the predictors which were somehow not significant, and not very much impactful on the levels of job satisfaction for replying this we take a point of departure comparing to general tradition of regression results reporting in which only significant implies good results. Interestingly though the predictors of job satisfaction were significantly different across both groups see wald= statistics in above table. But the effect sizes had explanatory power negligible indicating the exercise done to be futile at first glance.

Since the objective was exploratory where things could turn as expected or contrary. The general hypothesis of the differential impact of chosen job features was negated in this case study implying the homogenous impact of chosen features exists across both groups of satisfiers and non-satisfiers. The results urged us to look further for the possible source of difference for both classes. We addressed this query by adding subjective/ background variables since background variables or covariates come to play their role for finding the source of latent class membership in mixture models. Therefore, we did conditional analysis in next section.

Before getting into conditional models in the following we briefly discuss the model diagnostic tools including cross relationships, profiles and bivariate residuals for the above discussed model. These are included in this chapter and in further chapters to standardize the model selection procedure as much as possible to other chapters.

The relatively simple and precise way for looking into differential effects of predictor relationships across the classes is the cross-class effects in given table. The given table below shows the prevalence of cross significant relationship of each pair of predictors for meaningfully dividing the effects across classes. These relationships are conditional upon class membership, job size, job term, hours of doing job and the level of qualification. Here the size of effect is not measured but evidence exists for each predicts which significantly contributes to shaping cross class differences.

Table6.4. Regression Parameters Paired Comparisons

					Cross class comparison			Wald	d.freedom	p-value
	Class	1			Class	1	2	0.4646	1	0.5
	Job satisfaction	1	|	Class	Class	1	2	1199.711	6	2.10E-07
	Job Satisfaction	jobsize	|	Class	Class	1	2	10.4949	1	0.0012
	Job Satisfaction	jbterm_dv	|	Class	Class	1	2	7.7064	1	0.0055
	Job Satisfaction	jbhrs	|	Class	Class	1	2	18.0983	1	2.10E-05
	Job Satisfaction	hiqual_dv	|	Class	Class	1	2	31.3968	1	2.10E-08

6.3. 2. Classification of Selected Model

Profile output contains information on the class sizes, the class-specific marginal probabilities and means of the job satisfaction variable. It is clear from the first row that class 1 contains about 50% of the subjects (.5135), segment 2 contains about .4865%. Examination of class-specific probabilities shows that overall, segment 1 is least likely be completely satisfied with their work only 0.03% are completely satisfied compared to segment 2 who are most likely 28.% completely satisfied followed by 52 % likely of mostly satisfied level to 13% for somewhat satisfied level. The first four lowest levels of dissatisfaction are least reported in this class around 5 % in total compared to class 1 which has around 29 % likely cases reporting first four low score on satisfaction scale followed by highest likely cases of somewhat satisfied. Later in this chapter we will show how to classify each case into the most appropriate segment.

Table6.5. Classification Probabilities of LCRM

-

	unsatisfied	satisfied	Overall
Size	0.5132	0.4868	
Job-Satisfaction			
completely dissatisfied	0.0323	0.0088	0.0209
mostly dissatisfied	0.0635	0.0093	0.0371
somewhat dissatisfied	0.1392	0.0202	0.0812
neither satisfied or dissatisfied	0.1433	0.0271	0.0868

somewhat satisfied	0.3146	0.1307	0.2251
mostly satisfied	0.272	0.5219	0.3936
completely satisfied	0.0351	0.2821	0.1553

6.3. 3. Bivariate Residuals of Selected Model

The data is nested and clustered bivariate residuals between indicators are far higher indicating the modeling framework is not good enough to address the codependence of repeated (clustered) observations effectively and somehow justifying the danger involved in relying on predictions assuming the latent variable job quality is only the mediator between indicators. This is not such the case since the kind of job and schedule is leading to the levels of other indicators.

Table6.6. Two Level BVR of basic model

Dependent	jbsat
jbsat	.
Independent	jbsat
jbsize	0
jbterm_dv	0
jbhrs	0
hiqual_dv	0

Twolevel	jbsat
Case	1.6262
Pairs	452.6677

Whether we should discard the findings or should take these as a compromising solution for the given complex units for this we tested the out of sample performance of the chosen model by cross validation and latent class classification. The diagonal entries in the given table indicate exact classification and off diagonal entries show miscalculations. Further various types of error rates are reported based on absolute, marginal and log likelihood differences of baseline and calculated model. The divergence in rates is not much and k cross validation was also tested to check the sample performance for prediction purpose. Results in table 6.6 also reveal 10-fold validation. The model sustained the level of good classification and prediction power (classification error rate is 0.01 which is very low and indicates very good class difference and entropy R2 is around 0.7 which is also a good score compared to neutral score of 0.5).

Table6.7. Validation and Latent Classification

Latent Classification	Modal		
Latent	1	2	Total
1	4361.1458	473.3169	4834.463
2	370.1725	3448.365	3818.537
Total	4731.3183	3921.682	8653
Prediction Statistics			
Job sat			
Error-Type	Baseline	Model	R^2
Sq. Error	2.0315	1.4567	0.2829
LL	1.6136	1.4358	0.1102
Ab. Error	1.1243	0.9051	0.195
Pr. Error	0.6082	0.6059	0.0038

Classification Stat	Class		
Classification errors	0.0975		
Reduction in error(Lambda)	0.7792		
Entropy R-sq	0.681		
Standard R-sq	0.7215		

6.4. Conditional Models

An important extension of the above LC Regression model is obtained by making class membership dependent on covariates (Kamakura, Wedel, and Agrawal, 1994; Vermunt, 1997). Ours objective was to look for possible differences in two segments of given classes at various levels of occupational choices, age, gender and at various levels of overall quality of life. For meeting this objective, we separately examined 2 sets of models under step 3 analysis.

6.4.1. Step 3 Regression Mixtures

The variants in regression context were covariate proportional maximum likelihood based for individual and model case assignment followed by BCH corrections. In case of BCH corrections, the data does not follows chi2 distribution therefore criteria of **L^2** does not provides fit statistics (see chapter 4 section for discussion). Maximum likelihood (ML) based both corrections proved significantly fit. Considering lowest value of relative information tools, we opted for covariate proportional maximum likelihood based (CPML) case for 2 classes. We conducted step 3 analysis for 2 sets of covariates; first included role of occupational segregation for making satisfaction level choices. The same models also measured gendered differences across classes. The covariates included in second step 3 analysis were age and satisfaction with life to check the hypothesis of overall quality of life as a covariate for satisfactions with jobs.

Table6.8. Step 3 Regression Specification A

VARIANT	LLH	BIC(LLH)	AIC(LLH)	AIC3(LLH)	L^2	df	pvalue	C.Err.
CPML 2-Class	-40806	81755.42	81638.54	81651.54	11.2076	11	0.43	0.4542
CPBCH 2-Class	-40722	81587.55	81470.68	81483.68				0.4542
CMML 2-Class	-40850	81843.36	81726.48	81739.48	14.3696	11	.21	0.4541

CMBCH 2-Class	-40461	81065.05	80948.17	80961.17				0.4541

We included covariates to test the hypothesis that Quality of life and quality of job are interrelated or does satisfaction with over all life implies satisfaction with jobs?

Age differences have different across different satisfaction scores. Gender differences in the case A were present in making satisfaction choices. Females had relatively more chance to belong to class 2 (satisfiers) compared to males (61 % to 39%). This difference of proportions is might be because of sample size difference of both genders (58 vs. 42%).Age categories in sample are scattered from youth prime ages to very old people (16 years to 88 years).Total proportion is evenly divided in age groups and they are somehow nearly distributed across classes. The interesting finding is similar pattern of life satisfaction reporting over the years as reporting job satisfaction. The respondents for highest life satisfaction are more like to belong to job satisfiers class compared to those who are neutral and dissatisfied with life have more chances to fall in class 1. The proportion of mostly satisfied and somewhat satisfied with life is higher compared to basic model of job satisfaction presented earlier but likely patterns are somehow similar.

Table6.9. Average posterior Probability Case A

occupations	gender	unsatisfied	satisfied
self employed	Male	0.4342	0.5658
self employed	Female	0.3861	0.6139
Paid employment(ft/pt)	Male	0.5718	0.4282
Paid employment(ft/pt)	Female	0.5225	0.4775
unemployed	Male	0.6524	0.3476
unemployed	Female	0.606	0.394
retired	Male	0.2791	0.7209
retired	Female	0.2409	0.7591
on maternity leave	Male	0.5894	0.4106
on maternity leave	Female	0.5405	0.4595
Family care or home	Male	0.539	0.461
Family care or home	Female	0.4893	0.5107

full-time student	Male	0.7388	0.2612
full-time student	Female	0.6986	0.3014
LT sick or disabled	Male	0.7234	0.2766
LT sick or disabled	Female	0.6819	0.3181
Govt training scheme	Male	0.9538	0.0462
Govt training scheme	Female	0.9442	0.0558
Unpaid, family business	Male	0	1
Unpaid, family business	Female	0	1
On apprenticeship	Male	0.444	0.556
On apprenticeship	Female	0.3955	0.6045
doing something else	Male	0.5054	0.4946
doing something else	Female	0.4558	0.5442

In the following from the conditional effects of occupation categories and gender we can cross examine the likely distribution. Since the identification constraints in effect coding impose certain restriction on the sum of parameters for categorical variables therefore we have sum of parameters equal to zero in this case. We can see that people doing family business are most likely to belong satisfied class followed by retired and self -employed and on apprenticeship .Quite naturally the categories including of those individuals who are not working actively are more likely to belong to unsatisfied group. The class effects could be temporary for these individuals when they get back to work if want to. The categories included retired, unemployed, on maternity leave.

Table 6.10. Conditional Parameters for Case 1

Conditional Parameters for Classes				
Intercept	unsatisfied	satisfied	Wald	p-value
	0.0506	-0.0506	0.1681	0.68
Covariates	Cluster1	Cluster2	Wald	p-value
jbstat				
self employed	-0.2155	0.2155	163.5145	2.50E-29
Paid	0.0446	-0.0446		

employment(ft/pt)				
unemployed	0.3046	-0.3046		
retired	-0.5807	0.5807		
on maternity leave	0.123	-0.123		
Family care or home	-0.0371	0.0371		
full-time student	0.392	-0.392		
LT sick or disabled	0.3996	-0.3996		
Govt training scheme	1.1328	-1.1328		
Unpaid, family business	-1.341	1.341		
On apprenticeship	-0.1952	0.1952		
doing something else	-0.027	0.027		
gender				
Male	0.0552	-0.0552	130.0866	3.90E-30
Female	-0.0552	0.0552		

In second specifications the modeling variants were same, but the covariates were chosen different. Here CPBCH (covariate proportional) turned to be best option based on relative fit criteria reporting lowest loss of information. So, we opted this model parameters for further discussion.

Table6.11 Step 3 Regression Specification Case 2

	LL	BIC (LL)	AIC (LL)	AIC3 (LL)	L²	df	p-value	Class.Err.
CPML 2-Class	-36896.7	73935.76	73819.46	73832.46	1950.948	2171	1	0.3508
CPBCH 2-Class	-35204.7	70551.64	70435.34	70448.34				0.3507
CMML 2-Class	-35533.2	71208.72	71092.43	71105.43				0.3566
CMBCH 2-Class	-36722.4	73587.09	73470.8	73483.8	2601.163	2171	3.90E-10	0.3565

The given table reports the impact of various levels of satisfaction with life categorical impact on both classes of satisfiers and non-satisfiers. We can see that for class 1 the predicators are negligible to explain any differences whereas for class 2 these are for effective. Based on the 'Parameters' output we see that compared most satisfied last two categories the non-satisfied cases are less likely to be in class 2 than 1, this applies to all categories of no or clear satisfaction with life. Age is though significant to shape class formation but negligible followed by education role which is not explaining the likely change in categorical scores of job satisfaction. For class 2 the pattern of change is ambiguous. We concluded by these two-model specification that impact of first model parameters is clearer and more effective for class formation.

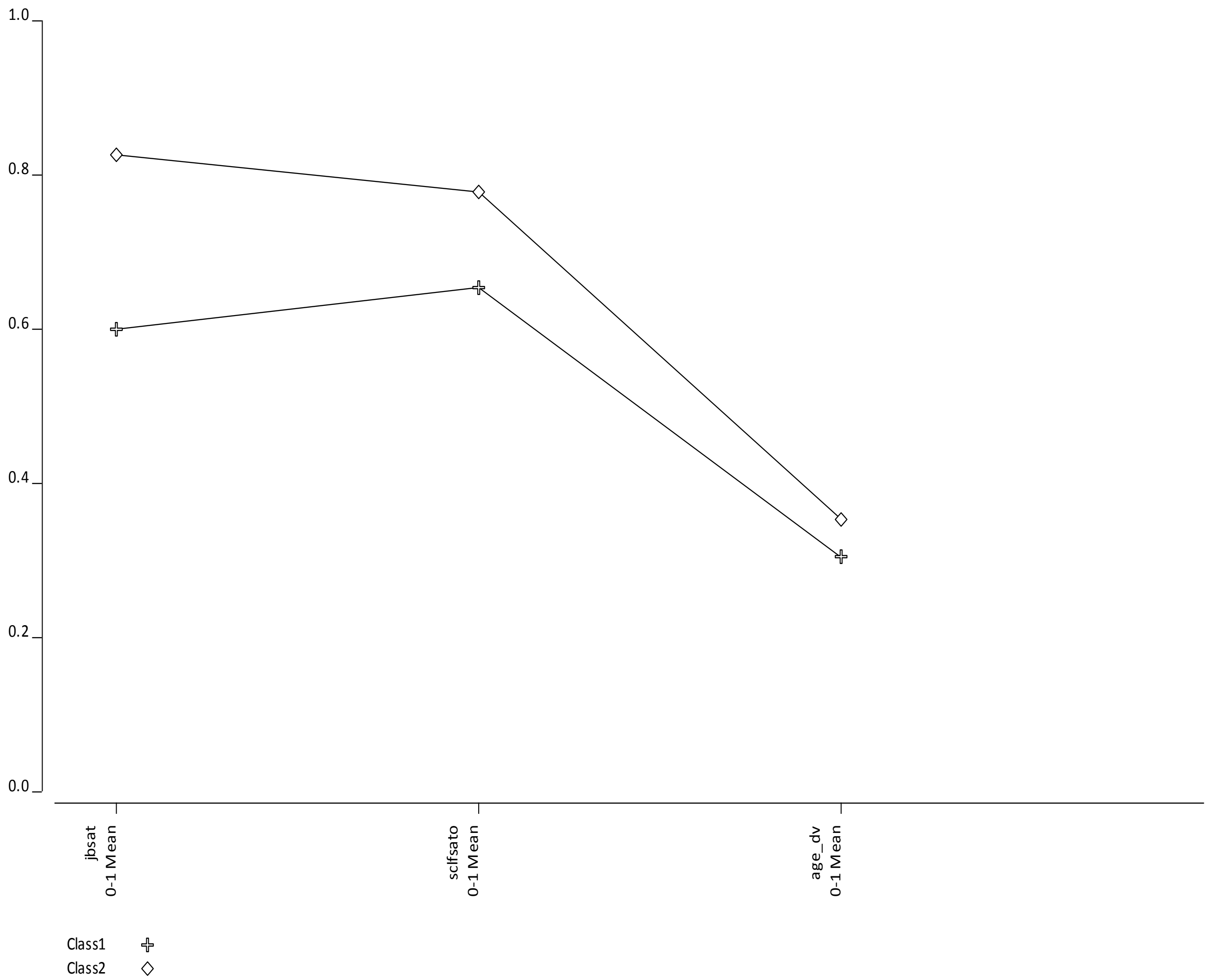
Table 6.12. Conditional Parameters for Case 2

Model for Classes				
Intercept	Cluster1	Cluster2	Wald	p-value
	0	-0.8056	61.325	4.80E-15
Covariates	Cluster1	Cluster2	Wald	p-value
scaleofsatisfaction				
completely dissatisfied	0.00	0	3907.84	5.1e-843
mostly dissatisfied	0.00	-0.5281		
somewhat dissatisfied	0.00	-1.621		
Neither Sat nor Dissat	0.00	-1.2234		
somewhat satisfied	0.00	-0.6632		
mostly satisfied	0.00	0.2658		
completely satisfied	0.00	1.3417		
age_dv				
	0.00	0.0191	404.4106	6.00E-90
hiqual_dv				
Degree	0.00	0	186.6861	2.00E-38
Other higher	0.00	0.0446		

A level etc	0.00	-0.1329		
GCSE etc	0.00	-0.3566		
Other qual	0.00	-0.2209		
No qual	0.00	-0.3452		

The given fig is based on the Profile output of the above give conditional model. These are average differences across classes. On the x-axis class-specific means for numeric covariates are reported as well as on the y axis the class-specific probabilities of being in a certain range of covariate values are displayed. These means and probabilities are obtained by aggregating and re-scaling posterior membership probabilities. In the above modeling setups we used these covariates to describe (rather than to measure) the latent classes and also to reduce classification error.

Fig6a. Profile difference for conditional case



6.5. Conclusion

In this chapter we have conducted conditional and unconditional analysis for longitudinal data featuring job satisfaction level subjective to job intrinsic and extrinsic features .We have taken diversified sample of UK population to explore the possible presence of heterogeneous sub populations within the larger group of individuals. The data was sparse when latent framework was applied to tabulate cross relations of included 5 to 8 indicators for 8000 plus individuals. For that reason, absolute fit diagnostics became invalid and through relative fit measures and cross validation the unconditional model was finalized. For conditional models' occupation and subjective scores on quality of life turned to be the best indicative of current standing of individuals on job satisfaction ladder.

CHAPTER 7

HANDLING DATA BY GROWTH MODELING

7.1. Introduction

What purpose growth modelling approaches serve, where conventional approaches lack and the usefulness of alternative mixture approaches in measuring development or change is discussed in the following followed by the empirical case study's results and discussion.

As the word growth suggests, under this modelling framework we attempt to estimate the change, development or progress in the variable of interest over a given time. Since change implies being dynamic therefore contrary to static cross-sectional framework, repeated measured data or longitudinal data is prerequisite for such analysis. In specific scenarios change between only two time points can also be measured by score change calculation. The change can be measured in continuous or discrete sense related to the theoretical query. The continuous change is measured in terms of mean and variance leading to quantitative difference measurement and discrete change can be measured in terms of classes.

The conventional growth approaches labelled somewhere as random effects multilevel modelling are based on average change as a sufficient representative for the whole sample under study. This framework is the assumption that all individuals are drawn from a single population with common parameters. In growth mixture modelling, latent subgroups or unknown subsets of population are assumed to exist within the data. To address and measure the difference in subsets of population mostly mean change, variance change, and covariance statistics are calculated and compared for the subgroups. Naturally this implies the usual imposition of normally distributed data from which the subgroups emerge. The assumption of normal distributed data is always challenging and controversial to hold in case of real-life data sets since survey-based data sets mostly provide skewed responses. Another reason of non-normal distributed data could be the class or cluster search before conducting the analysis. This leads to latent class variant of growth mixture modelling. The objective of class differences is accomplished in this case by using latent trajectory classes (i.e., categorical latent variables).This technique individual growth trajectories for each subclass varying around different means (Andruff et al., 2009).

Further for following the developmental course of individuals over the time for exploring in depth pattern differences in subgroups (of otherwise considered homogenous group) constraints the researchers on two grounds: first to acquire extensive dataset consisting of either repeated measure of variables (preferably longitudinal data, sometimes including multiple cohorts) secondly to specify the modeling techniques accordingly. Among such techniques for measuring change or growth patterns in explorative sense longitudinal latent variable growth curve models are suggested in literature (MacCallum & Austin, 2000).Nevertheless, conformist growth methodologies hypothesize a single population for all individuals and a single growth path sufficient for estimating the change in parameter of interest for whole population. Besides it is presumed that impact of external/auxiliary variables on growth factors is same for everyone in that peculiar approach. However, many real-life issues theoretically validate the categorical responses of individuals leading to distinctive subpopulations (e.g., socio-economic classes and employment status categories).

The main technique of exploring change over time does not changes everyone equally is growth mixture modelling. These models do not rely on the assumption that all individuals under study are drawn from a single population, which is the main limitation of more common latent growth models.compared to their cross sectional counterpart ‘latent class cluster model’ these models in longitudinal version explore the clusters for whom the rate of change or effect sizes are different .In this specific format these models are extension of latent class regression models where the predictor is mainly time or age and the effect size of time or age across various populations is postulated and measured. Thus, longitudinal heterogeneity through the identification of unobserved subpopulations in the sample under study is tested. The population heterogeneity in these models is captured by the inclusion of a categorical latent variable that identifies subgroups of individuals, typically referred to as latent classes

Latent growth modeling approaches such as latent class growth analysis (LCGA) and growth mixture models (GMM) have been much advocated for their effectiveness in categorizing homogeneous subgroups inside the grander heterogeneous group and for finding important classes with respect to growth patterns. The results of many studies presented in growth subsection of literature review suggest theoretical contentions of heterogeneity of growth trajectories existing within the larger population. In addition, many findings suggested that describing an entire population using a single growth trajectory oversimplifies the complex growth patterns related to continuity and change among members of different groups. To address limitations of standard or average growth approach, following Daniel S Nagin and Tremblay (2005); (Karen L Nylund et al., 2007) amongst many latent class growth and mixture modeling seems to be the most appropriate method for fully capturing information about inter individual differences in intra individual change taking into account unobserved heterogeneity within a larger population.

7.2. Summary of Models

We have estimated latent class growth models with variants for the empirical categorical employment status. Initially to save computational burdens and to observe any meaningful patterns of distinctive growth trajectories we have tested relatively simple versions of growth models on 2 cohorts with relatively small sample size (3 years unbalanced, 6 years balanced, and 10 years balanced time periods). All time slots had prime age periods (range between 16 to 25 years). Initial diagnostic suggested choice for 3 or 4 class models as the best fit (see Appendix C). We wanted to compare between the simple to complex structures of such models over youth to post adulthood for testing the hypotheses of likely development for various employment courses (let say from being inactive to employed, or from inactive to unemployed, from unemployed to employed and vice versa). Therefore, by extending age years (16 to 35) we separated analysis for 2 cohorts (see appendix B). It made ease on computation times and still we could include cohort effects later in conditional models.

Starting from standard homogenous growth we ended up on the growth mixture model with missing data. During the model's evaluation process with other mentioned absolute and relative model evaluation criteria explained in chapter 4, additionally after handling the class enumeration problem the selected model was further tested by bootstrap likelihood ratio test (BLRT)since we were again having sparse data. Competing models were also compared by cross validation outputs. Latent class growth model with 3 growth trajectories emerged as the best choice for trajectories classification (see graphs in Appendix C).

7.2.1 Linear latent growth models

Initially we started with the combined generalized linear model (GLM) with a hierarchical linear model (HLM) known as random-effects growth models in literature (Skrondal & Rabe-Hesketh, 2004). Following basic structure of these models repeated measurements on individuals are expressed as a function of time (look into methods section).We have measured individual differences in employment status when time equals zero and change in the various categories (employed, unemployed, out of work force) over time was modeled by permitting the intercept and slope coefficients to vary across individuals. The intercept and slope(s) are, therefore, referred to as random coefficients, random effects, or (latent) growth factors. To avoid over extraction of trajectories we started with random sets for all models since this option reduces the chance of over extraction of trajectories which is major issue encountered in growth mixtures(Jeroen K Vermunt, 2017).

Mainly three kinds of modeling option were selected for analyzing longitudinal data of 2 cohorts(79 and 97 of national longitudinal surveys by (United States Department of Labor). We labeled the classes based on most suggested model in light of model selection criteria opted in our analysis: relative fit between nested/non nested, parsimony, interpretability (theoretical validity) ,low classification errors, ease of convergence, high entropy R2, lowest two level bivariate residuals and finally by cross validation).

Starting from the basic linear growth model we can observe a size able decrease in log likelihood based absolute and relative model fit statistics. It is to be noted that imposed structure is different, so models are not nested generally. First case shows growth model where only intercept differs across 2, 3 and 4 class specifications. So, we can pick one best performing for this case of nested models. In second specification we vary only slopes or effect sizes across age for different possible class solutions. In third case inspired from first 2 specifications in favor of close options for 3 or 4 class as best fit we only tested for mixture variant for 3 and 4 classes. Most appreciated cases were checked for out of sample performance by cross validation and lastly RE specification showing latent class growth mixture was further tested for class 3 and class 4 solution including missing data. We can compare from the given summary table extent of bivariate residuals, entropy R2 and level of classification errors across models. In the following we separately discussed the specifications in terms of model performance to make final choice.

Table 7.1. Growth Models Specifications

	LL	**BIC(LL)**	**AIC(LL)**	**AIC3(LL)**	**Np**	**df**	**Max. BVR**	**Class.Err.**	**Entropy R^2**
1 class linear growth	-87628.5	175293	175265	175269	4	8110		0	1
2 intercept class specific LCGM	-75134.3	150340.6	150284.6	150292.6	8	8106	0	0.0248	0.8967
3 intercept class specific LCGM	-69989.7	**140087.4**	140003.3	140015.3	12	8102	0	**0.0244**	**0.9191**

4 intercept class specific LCGM	-68413.6	**136971.3**	136859.3	136875.3	16	8098	0	**0.078**	**0.8284**
2class random slope LCGM	-74708.9	149498.9	149435.8	149444.8	9	8105	0	0.0229	0.903
3classrandom slope LCGM	-69434.4	**138994.9**	**138896.8**	**138910.8**	**14**	**8100**	**0**	**0.0252**	**0.9182**
4classrandomslope LCGM	-67777.4	135725.9	135592.8	135611.8	19	8095	0	0.0764	0.8296
3 class GMM	**-63543.1**	**127239.3**	**127120.3**	**127137.3**	**17**	**8097**	**19.3333**	**0.232**	**0.479**
4 class GMM	-63435.6	127069.2	126915.2	126937.2	22	8092	15.8681	0.3311	0.3779
validation1	-67777.4	135725.9	135592.8	135611.8	19	8095	0	0.0764	0.8296
validation2	-69434.4	138994.9	138896.8	138910.8	14	8100	0	0.0252	0.9182
3 class GMM missing	-63525.8	127204.6	127085.6	127102.6	17	8097	0	0.2426	0.4498
4 class GMM missing	-63440.2	127078.3	126924.3	126946.3	22	8092	0	0.2153	0.4604

7.2.2 LCGM with Random Intercept

In this specification category-specific intercepts were tested for latent class growth analysis (see 2-4 intercept class specific lcgm). Within this modeling framework although 4 class solution could be selected on the base of lowest information criteria high entropy R2 and low classification errors, but class 3 solution emerged as more parsimonious with highest entropy R2 (91%) led by lowest classification error amongst competing alternatives (Look into highlighted comparisons of these statistics in table 7.1).

In the following we report the parameters and longitudinal path opted for each category by 3 distinctive classes, and lastly report two level bivariate residuals (results for some competing models are given in appendix C). Here we basically investigated significance differential effects of age (prime age life course) on shaping employment patterns of individuals by finding mean change for various classes for various employment statuses. In case of perfect heterogeneity, development of various phases of employment overtime should be distinctive across subgroups of the selected individuals which is not the case here. Though the initial starting differences in employment status is seen by the different intercepts across classes.

Table 7.2. Regression Scores for Random Intercept

classes					coefficent	s.e.	z-value
Class(1)				0	.	.	.
Class(2)				1.0394	0.029	35.8499	1.80E-281
Class(3)				-1.4881	0.055	-27.0619	2.80E-161
es(employed)		given	Class(1)	0	.	.	.
es(unemployed)		given	Class(1)	0.3749	0.0482	7.7792	7.30E-15
es(out of work force)		given	Class(1)	2.9428	0.0556	52.9708	7.7e-612
es(m.employed)		given	Class(1)	-1.9454	0.2007	-9.6925	3.20E-22
es(employed)		given	Class(2)	0	.	.	.
es(unemployed)		given	Class(2)	-1.2573	0.0418	-30.0471	2.40E-198
es (out of work force)		given	Class(2)	-0.1128	0.0519	-2.1714	0.03
es (m.employed)		given	Class(2)	-2.8572	0.1541	-18.5368	1.00E-76
es (employed)		given	Class(3)	0	.	.	.
es (unemployed)		given	Class(3)	-0.5831	0.0813	-7.1761	7.20E-13
es (out of work force)		given	Class(3)	0.2402	0.0832	2.886	0.0039
es(m.employed)		given	Class(3)	3.5175	0.0847	41.5384	4.1e-377
es	age			-0.0226	0.0005	-47.478	5.5e-492

Random intercepts /fixed effects also known as (latent class regression model) is significantly different for each class implying initial position of employment status is different for each class. Since for issue of identifiability of parameters we have employed dummy coding where first category labeled “employed” is taken as baseline from which the change in other categories would be compared. Across the three classes we can see that over the given age duration of 16 years, class 1 individuals are most likely to be out of work force compared to baseline status of being employed, for this class mean change from employed to out of work force equals to 2.94. The change is also positive, big and significant for case of being unemployed compared to being employed. For class 2, individuals are very unlikely to be military ‘m. employed’ from being regular full employed though in class 3 individuals are most likely to be military or m. employed from baseline status of being employed. The chance of being unemployed after being employed is lowest for the class 3. In summary employment choices are picked differently across these three classes. The average effect of linear time/age is though significant to shape these trajectories but insignificant, therefore the extended cases of square or cubic trajectories are not calculated further.

Bivariate residuals are extreme higher at second level of association, which imply the present level of autocorrelation are not well accounted by these models.

Table7.3. BVR for 3 class random intercept LCGM

Dependent	**es**
es	.
Independent	**es**
age	0.0000
Twolevel	**es**
Case	1.5175
Pairs	333.1598

In given profile table we have the advantage to compare the likely pattern of change in course of all 3 employment choices for each of three classes. Class 1 reveals the change (0.22 % to 0.51 %) in having first status (employed) over the 16 years. (We are explaining in terms of first and last point's change otherwise it is possible to read the change in each class for each 3 categories of employment status for each year). For class 2 this change margin is initially quite higher (0.81 to 0.93) compared to class 3 which have individuals who likely had the probability to remain in this category like class 1 but with higher range of change in their status of being employed over time (0.28 to 0.71). Class 3 emerges as different for category 1 status change over the time since its individual rise gradually for likely to be employed over the time. Similarly other categories reveal major differences over the prime age life course for this class. For category of 'out of work force' we had class 1 reporting highest proportion of likely cases that is 68% followed by steady decline in this status up to 41 % in last reported years. In summary the response patterns for growth of various employment status categories suggests class 1 (24 % size) have more likely cases who had grown over the time for being employed, and more likely cases who initially and finally ended up with being out of labor force whereas class 2 has individuals more likely to steadily remain employed around whole life span considered followed by class 3.

These distinctive patterns tempted us to label the classes, but we left this task until further specification diagnostics applied on improved model assumptions of random slope inclusion and mixture version of latent class models.

Table7.4. Profile-Longitudinal of employment status

status	Time	1	2	3	Overall	Observed
Emp. status						
employed	1	0.2248	0.8189	0.2824	0.6424	0.5903
	2	0.2412	0.8314	0.3102	0.6568	0.6256
	3	0.2584	0.8432	0.3392	0.6708	0.6656
	4	0.2762	0.8542	0.3693	0.6846	0.7002
	5	0.2948	0.8646	0.4003	0.6981	0.7297
	6	0.314	0.8743	0.4319	0.7114	0.744
	7	0.3336	0.8833	0.4636	0.7243	0.7472
	8	0.3534	0.8916	0.4948	0.7367	0.7554
	9	0.3737	0.8994	0.526	0.7489	0.7619
	10	0.3943	0.9066	0.5568	0.7607	0.7704
	11	0.4175	0.914	0.5899	0.7734	0.7743
	12	0.4432	0.9214	0.6251	0.7869	0.7798
	13	0.4705	0.9285	0.6602	0.8006	0.7833
	14	0.4883	0.9329	0.6825	0.8093	0.7876
	15	0.5051	0.9368	0.7026	0.8173	0.7884
	16	0.5153	0.9391	0.7144	0.822	0.8203
unemployed	1	0.0834	0.0595	0.0402	0.0643	0.08
	2	0.0834	0.0563	0.0411	0.0621	0.0782
	3	0.0832	0.0532	0.0419	0.06	0.0789
	4	0.0829	0.0503	0.0425	0.0579	0.0663

	5	0.0825	0.0474	0.0429	0.0558	0.0561
	6	0.0819	0.0447	0.0432	0.0538	0.0477
	7	0.0812	0.0422	0.0433	0.0518	0.0455
	8	0.0804	0.0398	0.0432	0.05	0.0424
	9	0.0794	0.0375	0.0429	0.0481	0.0388
	10	0.0783	0.0353	0.0424	0.0463	0.0393
	11	0.077	0.033	0.0417	0.0444	0.0387
	12	0.0753	0.0307	0.0408	0.0423	0.0375
	13	0.0734	0.0284	0.0395	0.0401	0.0328
	14	0.0721	0.027	0.0387	0.0388	0.0309
	15	0.0708	0.0257	0.0378	0.0375	0.0327
	16	0.07	0.0249	0.0372	0.0367	0.0295
out of work force	1	0.6897	0.1185	0.0581	0.2561	0.2779
	2	0.6735	0.1096	0.058	0.2458	0.2504
	3	0.6566	0.1012	0.0577	0.2358	0.2173
	4	0.6392	0.0934	0.0572	0.2261	0.1978
	5	0.6212	0.0861	0.0565	0.2165	0.1832
	6	0.6026	0.0793	0.0555	0.2071	0.1823
	7	0.5838	0.0731	0.0544	0.1981	0.1857
	8	0.565	0.0674	0.053	0.1894	0.1819
	9	0.5458	0.0621	0.0515	0.1808	0.1829
	10	0.5263	0.0572	0.0498	0.1725	0.176
	11	0.5046	0.0522	0.0478	0.1635	0.1743
	12	0.4806	0.0472	0.0454	0.154	0.1717
	13	0.4553	0.0425	0.0428	0.1444	0.1742
	14	0.4388	0.0396	0.0411	0.1382	0.1715
	15	0.4234	0.037	0.0395	0.1325	0.1679
	16	0.414	0.0355	0.0385	0.1291	0.1458

7.2.3 LC Growth Models with Random Slopes

In this chapter we have not ended up with one best fit model since each of the variants discussed above address development(change)in employment status under different assumptions so from three set of underlying assumed structures one best fit from each set was selected and compared in terms of interpretation of change over time. Henceforth, after selecting best from first set i.e., 3 class model next models was set up for included random slope to incorporate the change in growth of various classes around the mean value of change. We tested for whether the unique slope parameter brought further insight in understanding differential effect of age on given classes.

Under this specification, age lacks explanatory power for explaining class differences since the effect sizes of age are negligible. This implies the mean level of change is not much different across the three classes and random slopes is not suitable specification for this case. As far the conditional effects of time are concerned, we find some changed effect sizes naturally. For class 2 individuals we have highest likely change of being out of labor force over the time after being employed, this effect is large and significant for class 3 as well and lowest positive for class 2.Class 2 individuals are most unlike to be unemployed over the age and class 1 individuals have highest chance to be unemployed after being employed over the age , the effect size is around .5 and significant. See the statistics of individual and across the classes reported through wald (=) and wald (0).

Table7.5. Regression Scores for LC growth model

classes				**coefficent**	**s.e.**	**z-value**	**p-value**	**Wald (0)**	**df**	**p-value**
Class(1)	**1**			0.00	.	.	.	2917.50	2	3.0e-634
Class(2)	**1**			1.58	0.06	26.2464	7.9e-152			
Class(3)	**1**			2.59	0.05	45.4167	2.2e-450			
es(employed)	**1**	**|**	**Class(1)**	0.00	.	.	.	14552.12	9	3.1e-3148
es(unemployed)	**1**	**|**	**Class(1)**	0.65	0.13	4.92	8.5e-7			
es(out of work force)	**1**	**|**	**Class(1)**	1.88	0.15	11.86	1.8e-32			
es(m.employed)	**1**	**|**	**Class(1)**	5.80	0.20	27.68	9.8e-169			

es(employed)	**1**	|	**Class(2)**	0.00	.	.	.			
es(unemployed)	**1**	|	**Class(2)**	-1.47	0.07	-20.81	3.4e-96			
es(out of work force)	**1**	|	**Class(2)**	0.48	0.08	5.46	4.7e-8			
es(m.employed)	**1**	|	**Class(2)**	-5.85	0.26	-21.85	7.6e-106			
es((employed)	**1**	|	**Class(3)**	0.00	.	.	.			
es(unemployed)	**1**	|	**Class(3)**	0.13	0.07	1.81	0.070			
es(out of work force)	**1**	|	**Class(3)**	1.69	0.09	18.43	6.5e-76			
es(m.employed)	**1**	|	**Class(3)**	0.24	0.20	1.18	0.24			
es	**age**	|	**Class(1)**	-0.03	0.00	-28.04	4.9e-173	2753.83	3	4.3e-597
es	**age**	|	**Class(2)**	-0.00	0.00	-0.42	0.67			
es	**age**	|	**Class(3)**	-0.04	0.00	-43.19	1.2e-407			

Further cross class effect of age to designate individuals to employment course is highly significant from the given table of paired comparisons.

Table 7.6. Paired comparisons

term				comparison			Wald	df	p-value
Class	1			Class	1	2	688.8715	1	7.90E-152
				Class	1	3	2062.679	1	2.2e-450
				Class	2	3	1254.057	1	1.10E-274
es	1	|	Class	Class	1	2	1889.448	3	1.8e-409
				Class	1	3	1216.47	3	2.00E-263
				Class	2	3	699.9851	3	2.10E-151
es	age	|	Class	Class	1	2	543.4041	1	3.40E-120
				Class	1	3	8.5902	1	0.0034
				Class	2	3	941.3757	1	9.90E-207

The profile section suggests somehow similar likely pattern of growth over time for various categories of Employment status. To summarize the likely cases of growth for ES category 1 for class 1, it is changing positively over time suggesting the rate of being employed is positive. For unemployed category there are more likely cases for whom growth in being unemployed is low. For ‘out of work force’ we observe somehow similar pattern whereas being inactive the reported rates are low at initial youth years to middle years and finally declined growth rate , this suggests overall more individuals of this class are economically active. For class 2, unemployed categories are steady over the time with low starts and ends whereas out of work force individuals remained part of this group for more than 50 % more or likely all the time.

Class 3 had opposite developmental course for individuals being fully employed compared to other two classes, it had highest reported likely cases of being active labor at youth years and had cases of such individuals from initial probability of 76 % with the positive change upto 90% being employed over time. This naturally suggested decline in growth of other 2 categories for this group. As we can see for category out of workforce there was persistent decline in terms of size of 15% to 1 % over age. We label the classes on the base of common response patterns as Mediocre active, Mostly Inactive, Active.

Table 7.7. Longitudinal Change Patterns of Classes (Random Slopes)

		Class				
	Time	1	2	3	Overall	Observed
Employment status						
employed	1	0.1722	0.3561	0.7624	0.6296	0.5903
	2	0.205	0.3564	0.7897	0.6504	0.6256
	3	0.2418	0.3568	0.8145	0.6696	0.6656
	4	0.2824	0.3571	0.8369	0.6874	0.7002
	5	0.3265	0.3574	0.857	0.7038	0.7297
	6	0.3736	0.3577	0.8749	0.7187	0.744
	7	0.4221	0.3581	0.8906	0.7323	0.7472
	8	0.471	0.3584	0.9044	0.7445	0.7554
	9	0.5203	0.3587	0.9165	0.7556	0.7619
	10	0.5687	0.359	0.9272	0.7656	0.7704
	11	0.62	0.3593	0.9374	0.7754	0.7743
	12	0.6725	0.3597	0.9469	0.7849	0.7798
	13	0.7225	0.3601	0.9553	0.7934	0.7833
	14	0.7532	0.3604	0.9603	0.7986	0.7876

	15	0.7797	0.3606	0.9645	0.8029	0.7884
	16	0.7947	0.3608	0.9668	0.8053	0.8203
unemployed	1	0.0372	0.0794	0.0742	0.0736	0.08
	2	0.0396	0.0794	0.0677	0.0692	0.0782
	3	0.0417	0.0794	0.0615	0.065	0.0789
	4	0.0435	0.0793	0.0557	0.0611	0.0663
	5	0.045	0.0793	0.0503	0.0573	0.0561
	6	0.046	0.0793	0.0453	0.0539	0.0477
	7	0.0466	0.0793	0.0407	0.0507	0.0455
	8	0.0468	0.0793	0.0366	0.0478	0.0424
	9	0.0464	0.0793	0.0328	0.0452	0.0388
	10	0.0456	0.0793	0.0293	0.0427	0.0393
	11	0.0441	0.0792	0.0259	0.0403	0.0387
	12	0.0421	0.0792	0.0226	0.0379	0.0375
	13	0.0395	0.0792	0.0196	0.0356	0.0328
	14	0.0377	0.0792	0.0178	0.0343	0.0309
	15	0.0359	0.0792	0.0162	0.0331	0.0327
	16	0.0347	0.0792	0.0153	0.0324	0.0295
out of work force	1	0.0618	0.5635	0.1564	0.254	0.2779
	2	0.0632	0.5632	0.1369	0.2404	0.2504
	3	0.0642	0.5629	0.1194	0.2282	0.2173
	4	0.0646	0.5626	0.1037	0.2172	0.1978
	5	0.0643	0.5623	0.0898	0.2074	0.1832
	6	0.0635	0.562	0.0775	0.1988	0.1823
	7	0.062	0.5617	0.0668	0.1912	0.1857
	8	0.0599	0.5614	0.0576	0.1846	0.1819
	9	0.0574	0.5611	0.0496	0.1787	0.1829
	10	0.0543	0.5608	0.0426	0.1736	0.176
	11	0.0506	0.5604	0.036	0.1688	0.1743
	12	0.0462	0.5601	0.0299	0.1642	0.1717
	13	0.0415	0.5597	0.0247	0.1602	0.1742
	14	0.0384	0.5595	0.0216	0.1579	0.1715
	15	0.0355	0.5592	0.019	0.1559	0.1679
	16	0.0338	0.5591	0.0176	0.1548	0.1458

7.2.4. Growth Mixture Results

In this section we present the last variant of growth models relatively for some vivid reasons. Going back to the summary table 7.1 presented in start of chapter if we had to make only one choice of the most suitable representative model of given data then random effects version models having fixed variance covariance structure within same class were best fit in terms of lowest value for relative fit statistics for four cluster case followed by 3 cluster case. Since clustering remained main objective thoroughly in this thesis so we gave more weightage to low classification errors otherwise in growth literature it is very much recommended to choose model based on relative fit criteria. Following this criteria and very important diagnostic tool of two-level bivariate residuals significant low scores we compared between 4 cluster and 3 cluster competing cases and compared results through low classification errors in each model. Further to make selection we applied bootstrapping to confirm final choice. Validation statistics revealed lower prediction errors for 3 cluster case and bootstrapping absolute fit for sparse tables suggested the model fits well (see Appendix C).

For second cohort (nls_97) we applied these last variants and compared the development/change course of employment status for two different time zone individuals. Below we present cross validation comparison tables followed by detailed discussion for differential effects of random intercepts, slopes, variances and covariance .In last we present the proliferation of individual class trajectories through profiles and expected change values. Additionally graphs are provided to make comparisons visual for 3 and 4 cluster change patterns.

Table 7.9.Regression Score for Growth Mixture

term				coeff	se	z-value	p-value
Class(1)				0	.	.	.
Class(2)				1.596	0.072	22.1508	4.10E-6
Class(3)				2.0199	0.0626	32.2845	6.10E-6
es(employed)		|	Class(1)	0	.	.	.
es(unemployed)		|	Class(1)	6.0362	0.3534	17.0807	2.10E-65
es(out of work force)		|	Class(1)	8.7094	0.4683	18.5993	3.30E-77
es(m.employed)		|	Class(1)	15.2694	0.6892	22.1564	2.70E-2
es(employed)		|	Class(2)	0	.	.	.
es(unemployed)		|	Class(2)	2.2309	0.2214	10.0758	7.10E-24
es(out of work force)		|	Class(2)	3.1981	0.2808	11.3895	4.70E-30

es(m.employed)		|	Class(2)	0.6436	0.5004	1.2861	0.2
es(employed)		|	Class(3)	0	.	.	.
es(unemployed)		|	Class(3)	-0.4708	0.1382	-3.4067	0.00066
es(out of work force)		|	Class(3)	2.2526	0.1697	13.2745	3.30E-40
es(m.employed)		|	Class(3)	-5.9817	0.8933	-6.6965	2.10E-11
es	age	|	Class(1)	-0.1051	0.0039	-26.7122	5.20E-65
es	age	|	Class(2)	-0.0672	0.003	-22.4052	3.10E-23
es	age	|	Class(3)	-0.0374	0.0018	-20.3032	1.20E-91
Variances							
term				coef	s.e.	z-value	p-value
u0				1.8349	0.0298	61.5691	9.1e-826
u1				0.0217	0.0003	71.7924	6.9e-1122
Covariances / Associations							
term				coef	s.e.	z-value	p-value
u0	u1 (chol)			-0.0757	0.0012	-62.7738	2.7e-858
Variances / Covariances continuous latent							
term				coef	s.e.	z-value	p-value
u0				3.3668	0.1094	30.7845	2.10E-65
u0	u1			-0.139	0.0044	-31.6245	3.10E-35
u1				0.0062	0.0002	33.163	1.10E-38

From the above table we can see that the included continuous random effects for separating classes more effectively are highly significant and with very low standard errors, variances of intercept shown through an aggregative u0 are high in magnitude compared to the u1 which measures average change around mean values for each cluster. Covariance is highly significant though negligible in size.The breakdown of continuous latent term in last decomposed matrix form shows the effects of considering different means and variance and covariance structure for the data in hand is significantly applicable. After the continuous part of mixture framework we come to discuss the usual class element by reading the effect of distinct slopes and distinct intercepts for each class. We observe each of the class had different position to take development from one state to another this is read through the conditional effect of time for each category for each class, also the effect of the only considered age (random slope) is negative and significant for each class though low in magnitude. This negative effect makes one thing clear that whatever the change faced for ES categories for the given three classes ultimately over time there were declines in affiliations to these patterns with low sizes. Also the parameters for the four categories in each class are highly significantly different reported through (wald equal)statistics.

The two level bivariate residuals value far below 2 indicates significant values of conditional independence at groups and individual level, this low reported value signals the model fit and suitability to study the measured change in structure (ES change). Since the higher values of these cross-dependence indicators implies model misfit in last versions.

Table 7.10.BVR Score for LCGM

Dependent	es
es	.
Independent	es
age	0
gender	19.3333
Two level	es
Case	0.1827
Pairs	0.0883

From comparing the results of above specification of latent class growth model to other 2 specifications in terms of model performance we concluded the last one as the best approach to read change in employment status over time.

7.3. Step 3 Variants

In this section we report conditional models based on solely step 3 approach (four variants).The usual specifications were divided for proportional and model assignment of values in step 3 and based on standard maximum likelihood based corrections and BCH corrections (see discussion in chapter 2).3cluster case with BCH corrections based on modal assignment turned to be best in terms of relative error and parsimony .The reported likelihood is also maximum relevant to this model (CPBCH 3-Clu).

Table7.11. Case specifications for Step 3 Analysis of growth model

		LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar	df	Cl.Er	Entr R^2
Model2	Step3-Covariate-Proportional-ML 3-Cluster	-96423.16	192939.31	192862.3	192870.3	8	4	0.29	0.21
Model3	**Step3-Covariate-Proportional-BCH 3-Cluster**	-79506.55	159106.08	159029.1	159037.1	8	4	0.29	0.21
Model5	Step3-Covariate-Modal-ML 3-Cluster	-91195.89	182484.78	182407.8	182415.8	8	4	0.29	0.20
Model6	Step3-Covariate-Modal-BCH 3-Cluster	-80765.08	161623.15	161546.2	161554.2	8	4	0.29	0.20

Further the average probabilities of the three selected classed based on the conditional effects (gender and ethnicity) are reported. We can see that active employed are highest in size and surprisingly female dominantly belong this cluster and amongst inactive male black are more likely to remain in this group over the age. Finally we checked the conditional belonging of these variable categories for the three classes

7.13. Case Profile by Step 3 Analysis

Average Probability			
	Mediocre active	Mostly Inactive	Persistent Active
Overall	0.0867	0.3673	0.546
Covariates			
gender			
M	0.1633	0.5832	0.2535
F	0.0126	0.1584	0.829
ethnicity			
Black	0.1062	0.447	0.4468
Hispanic	0.0711	0.3104	0.6185
Nonblack/Non-Hispanic	0.081	0.3414	0.5775

Estimated Values-Model				
		Cluster		
gender	ethnicity	Mediocre active	Mostly Inactive	Persistent Active
M	Black	0.1945	0.6564	0.1491
M	Hispanic	0.1345	0.5192	0.3463
M	Nonblack/Non-Hispanic	0.1557	0.564	0.2804
F	Black	0.0203	0.2434	0.7364
F	Hispanic	0.0073	0.1004	0.8923
F	Nonblack/Non-Hispanic	0.0101	0.1299	0.8601

7.4. Conclusion of Empirical Application

By employing longitudinal data over the prime working age of American labour class we attempted to explore growth patterns differences in taking employment course. To serve the objective and for testing the hypothesis of heterogeneous subpopulation within the larger population we employed methods which could measure inter individual differences in intra individual change over time. Three mainstream modeling variants of growth modeling were tested and elaborated in terms of model performance. Lastly status typology was built based on the consensus of model variants.

Though class sizes appeared different under above discussed growth variants but typically three patterns of change were observed under each specification. The highest proportion had those likely cases that remained and grow to be employed and least likely cases of being inactive or unemployed. We labeled that cluster as persistent active. The relative smaller cluster had initially lower reported cases of being employed which persistently growing in likely to be employed and highest cases of individual to be unemployed following remarkable slows in such status over age and distinctive highest and consistent cases of being out of labour force over the age. On part of such contributors, we labeled the second cluster as mostly inactive. Third segment had lowest number of likely cases with different response patterns of giving lowest employment starters cases rising over age to be employed but low in proportion to other clusters we named this cluster as mediocre.

CHAPTER 8

HANDLING DATA BY MARKOV MODELS

8.1. Introduction

When same individuals are asked for responses over the same questions (as done in panel studies) the responses from same measurement units become clustered naturally. The issue of auto correlation is quite common in such data-based models and handled from various econometric perspectives.The purpose of choosing and applying Markov models in this chapter is to utilize the offered modeling flexibilities of these models for incorporating measurement error(present in survey data), possible heterogeneity and the autocorrelation in time dependent data simultaneously. Mixture versions of latent Markov models serve to find all said purposes. The most general structure of these models (see methodology) and its variants provide flexible modeling options for longitudinal data handling. In this chapter we discuss mainly those models which incorporate the transitions of units over time or age effectively. The possibility of measurement error calculations accounting for unobserved heterogeneity between subjects for latent subgroups is unique compared to other methodologies treating longitudinal data presented in earlier chapter. Here we utilize the same data to make comparative notes for theoretical findings and methodological improvements if any. Our criteria of finding the best fit for data clusters (states called in this modeling environment) would be based on additional considerations of longitudinal performance of the models measured through longitudinal bivariate residual errors presence at various levels and prediction accuracy (measured through cross validation).

Since the competing modeling family was of growth models in the last chapters and the analysis in this chapter is based on the same data utilized in last chapter therefore, we justify here working further with chosen modeling setup. In the previous section the reason that first two growth variants failed to provide an adequate fit is because of the non-capacity of those models for fully explaining the first order autocorrelation in longitudinal data. Latent Markov models differ from traditional latent class models in that they contain transition probability parameters which accounts for first order autocorrelations directly by allowing the latent category that a person is in to change over time. Such dynamic latent categories are called states. A 2-state latent Markov (LM) model differs from a 2-class latent growth model in 2 primary ways: while persons in the 3-class latent growth model belonging to persistent active, inactive or mediocre class is considered to remain in that class for whole slot of 16 years. Compared to this, in latent Markov models a person in one class may shift to another class at some other time and the exact probability of moving from one state to another is obtained. Compared to Latent class growth models, these models can explicitly include the probability of changing from 1 state at previous time to some other state at next time (transition probabilities) as model parameters. Because these transition probability parameters account for first order autocorrelation directly it is probable that these models provide a better fit to longitudinal data. The models discussed in this chapter are Markov, latent class Markov, manifest Markov , time homogenous Markov ,time heterogenous Markov and mixture latent class Markov. Each of these are unique in perspective of time dependence for influencing current and next state of individual units. Starting from the basic manifest Markov we can say the current and future state of individual units is not influenced by any other factor other than the given variable information itself. The Markov model is based on the assumption of current state dependence on very previous state. The hidden or latent class Markov assumes the state of dependence on the very previous unit of change but the element of change as hidden or latent in nature. The time homogenous or time heterogenous transition structure imposes the effect of equal or different effect across sub populations for moving from one stage to another. The most general structure of latent mixture Markov is based on pattern of change and states of change difference between subpopulations .For details of these variants see into(Bartolucci et al., 2019).

8.2. Theoretical Significance

The basic notions involved in describing the employed and unemployed are standard. People doing some work on regular basis jobs are called employed and contrary to this segment people who are jobless seeking available for work are considered unemployed. The labor force is usually counted by these two segments. People who are neither employed nor unemployed are not in the labor force. If we have another category where people are not working and not even looking for work to be called unemployed than this makes more extreme case than out of labor force, the category represents inactive part of the economy. Discouraged workers are people of working age who stop looking for work. They usually stop looking because they are discouraged after not finding a job. Because they are not actively looking for work these people are excluded from the labor force and are therefore not counted in the official unemployment rate. Finding the transitions from one status of employment to another by dynamic analysis can bring more inside picture of employment situation compared of stagnant statistics in particular time measured.

8.3. Results Summary

We have done separate analysis for various modeling setups explained in methods chapters since we had longitudinal information for 2 cohorts with gender and racial information therefore for total data we made cohort comparisons purposely for state change behavior over time and conducted the analysis further with step 3 approach to trace the state change differences over gender and racial groups. Time homogenous models are the models where we assume that time effects all population considered equally contrary to these models under

Time heterogeneous structure, we assume time effects each class differently for moving from one state to another. In most general structure of Mixture latent Markov models the heterogeneity of observations is measured by mixture component (see methodology section 4);autocorrelation by Markov component , and measurement error through latent part of the model (Paas et al., 2007b).We have considered most general structure (mixture latent Markov) and its special variants :Latent Markov models ,Mover stayer models / with covariates for 2 age cohorts and for prime working age. Since unemployed or being inactive affects current and future employment choices of individuals in the peak working years so we precisely focused to observe the transitions of individuals from working to non-working categories**.** In the following the summary table is given with explanations followed.

Table 8.1. Summary of Markov Variants

total	**LL**	**BIC(LL)**	**AIC**	**AIC3**	**N**	**df**	**MB VR**	**Cl.Er**	**EntrR^2**
1CReg	-139058.12	278154.96	278124.2	278128.2	4	15951		0	1
2CReg	-119585.25	239257.60	239188.5	239197.5	9	15946		0.0336	0.857
3CReg	-111488.99	223113.47	223005.9	223019.9	14	15941		0.0364	0.877
4CReg	-109081.40	218346.68	218200.8	218219.8	19	15936		0.1009	0.771
5CReg	-107511.77	215255.81	215071.5	215095.5	24	15931		0.1276	0.750
3State 2CMark	-102895.11	205945.72	205822.2	205838.2	16	16602	0	0.0768	0.678
4State 2CMark	-100808.03	201859.02	201666.0	201691.0	25	16593	0	0.1703	0.561
3State 2CMark	-102142.82	204528.60	204335.6	204360.6	25	16593	19.8	0.0662	0.718

3State 2CMark	-103607.94	207488.00	207271.8	207299.8	28	16590	2.30	0.1055	0.641
4State 1CLat Mark	-100646.94	201556.27	201347.8	201374.8	27	16591	0	0.1627	0.562
4State 1CMe niMak	-104446.53	209038.84	208923.0	208938.0	15	16603	0	0.0661	0.745
4State 1CHet LatMa rk	-100263.48	200905.98	200604.9	200643.9	39	16579	1.26	0.1532	0.573

The table details Markov specifications from 1 class R to 5 class R is same specification of simple regression to mixture regressions. Starting from the null model of mutual independence between responses at each of the 16 time points we find sizeable improvement in results by increasing the class size up to 5 class but the inspection of 4 and 5 class revealed that the last 2 classes for these models are negligible in size and theoretically inseparable. Before evaluating Markov models these simple regressions with age as a predictor served as a baseline guess for the possible heterogeneous segments of people with possible different employment status over life. We made a closer inspection in predicted employment categories values for 3 class and 2 class case it was evident that data mostly contains employed people over the considered life span. Though we were not targeting to find the predictors of various choices of ES categories over time, but we wanted to find those segments who moved either to being inactive or discouraged over age and were further interested to see the transition structure background conditional to cohorts racial and gendered differences therefore after a preliminary guess of class structure through categorical regression analysis we tested for various states and class structures on the 2 age zones data. Since the data was huge in size so contingency tables were sparse. It is suggested in uni-variate Markov literature by Bartolucci et al. (2015) to work with states equal to the categories(i.e. observed states are as equal as true states). In next models with relatively small sample size (single cohort analysis) we worked with more different cases by considering the condensed responses for first category. And for keeping the union of theoretically different four states we postulated for 3 and 4 states combination suitable to represent the data. In above table we have tested for 3 state latent Markov case, further with covariates of gender, cohort and ethnicity, 3 class mixture latent Markov unconditional, 3 class mixture latent Markov conditional, and for 4 states usual homogenous and heterogeneous time specifications for 1 class latent ,and manifest Markov models. Below we discuss the models and present the initial state probabilities of each state followed by transition probabilities and discuss the measurement error calculated under each discussed variant. We also report longitudinal bivariate residuals to compare the model's performances for accounting first order autocorrelation in given data.

8.3.1. Model 1: Manifest Markov

We started with this simplest version (4state 1cmenim labeled in table 8.1) of categorical change in employment status for its simplicity and as a base model for making the complicated cases understandable, Manifest Markov are a special variant of the general mixture model (Nylund-Gibson et al., 2014). The imposed structure considers the variables are measured perfectly, though the model (4state 1class manifest Markov) fits better to the mixture version of 3 state cases with homogenous and heterogeneous transitions. Still, we do not find this model appropriate to represent our data due to measurement error present in survey items and more precisely the construct of employment status has quite a margin to be measured inaccurately. (Since the measurement of employment status is biannually, therefore the individual employment experiences are not accurately accounted for during the unreported spell).

From the given table it is seen that the initial probability of belonging to the given four employment statuses are 63% ,6%,25%, and 4% in sequence for being employed, being unemployed, being out of labor force and military employed (m. employed). (We would not be discussing this special case of employment since first three categories are of main interest). At the age of 21 around75% have had some job experience or work background, the transition table shows very interesting inter-status change Probabilities. Given a case is in 'employed' state (State 1) at time t-1, the probability of remaining in this state at time t is .89. Given a case is in the other state (being unemployed) at time t-1, the probability of remaining in first category of ES employed state at next time only is .18. This shows change from first state to second state is .72. Similarly given a case in the state of out of labor force at time t-1, the probability for an individual to remain in employed is .63%. Therefore, transitions from employed to out of labor class are 37%. Similarly for the last category from first state is calculated and is lowest (1-.81) =19% amongst all transitions.

Measurement model for manifest case reports no measurement errors, since the indicators are assumed to be perfect representatives of employment statuses. Bivariate residuals between time and es are quite big in size (see 8.3.) which shows that first order auto correlation is not well addressed under this model specification.

Table 8.2. Estimated values Manifest Markov (Model A)

	Initial State probability			
	employed	unemployed	Out of work force	m. employed
	0.6325	0.0685	0.256	0.043
	State			
State[-1]	employed	unemployed	Out of work force	m. employed
Employed	0.8951	0.0339	0.0688	0.0022
unemployed	0.6047	0.1877	0.204	0.0036
out of work force	0.2914	0.0726	0.6332	0.0027
m. employed	0.1262	0.0271	0.0336	0.8131
	observed			
True State	Employed	unemployed	out of work force	m. employed
Employed	1	0	0	0
unemployed	0	1	0	0
out of work force	0	0	1	0
m. employed	0	0	0	1

The bivariate 2-way residuals are reported for the given case these are quite higher than 2 at second levels of time nesting.

Table 8.3.BVR for Manifest Markov (Model A)

Indicators	Emp-status
Emp-status	.
Longitudinal	es
Time	7.1406
Lag1	0.8662
Lag2	29.1913

8.3.2. Model 2: Heterogeneous Latent Transition Model (Model B)

Second interesting case amongst 4 state models is the time heterogeneous Latent transition model (4state 1C LM labeled in table 8.1). Initial state probability for being employed for this specification is low compared to manifest Markov i.e.,48% unemployed and out of labor force states as initial starts.

The transition table is quite informative to trace cross category and within category transition over the age since it becomes complicated to read each cross category to formalize the transition scores in heterogeneous Markov. We provide an overview based on base model results. Here age is the covariate which is not affecting each class equally. For class 1, the individuals are more consistent to remain in this state whereas for class 2 over the time the probability to remain in this state fall. Since the probability to belong to one state at one time sum to 1 over the row or over the same time therefore diagonal readings of transitions based on previous time to the next time suffice to understand the non-diagonal categories transitions. From the diagonal pattern observation over the whole age zone considered, we see that for the people who have started at state 1 of being employed, they remained consistent over time. We observe consistent transition on diagonal entries which implies the probabilities to switch from one state to another are stable over the age. This implies one class remains consistent in their state; this led to test for the hypothesis of movers and stayers. Additionally, the measurement part of the results helped us to see the superficial distinctive category representation of individuals between being unemployed or out of labor force. The bold probabilities in third section show the highest prevalence of overlapping of being unemployed and out of labor force. In next modeling scheme we merged these categories to account for measurement error comparisons.

Table.8.4. Estimated probabilities for Latent Markov 4 state (Model B)

		Initial State probability			
		employed	unemployed	Out of work force	m.employed
		0.481	0.224	0.2513	0.0438
		Transition State probabilities			
age	State[-1]	employed	unemployed	Out of work force	m.employed
21	Employed	0.9335	0.0289	0.0337	0.0038
21	unemployed	**0.1466**	0.7098	0.1384	0.0052
21	out of work force	0.2041	0.0665	0.714	0.0153
21	m.employed	0.0964	0.0252	0.1045	0.7738
22	Employed	0.9404	0.0266	0.0302	0.0028
22	unemployed	0.1309	0.7335	0.1319	0.0038
22	out of work force	0.1952	0.0686	0.7258	0.0104
22	m.employed	0.0925	0.0224	0.0931	0.792
23	Employed	0.9464	0.0245	0.0269	0.0021
23	unemployed	0.1165	0.7556	0.1253	0.0027
23	out of work force	0.1862	0.0706	0.7361	0.0071
23	m.employed	0.0884	0.0199	0.0828	0.8089
24	Employed	0.9518	0.0226	0.0241	0.0015
24	unemployed	0.1033	0.7761	0.1187	0.0019
24	out of work force	0.1774	0.0725	0.7453	0.0048
24	m.employed	0.0844	0.0177	0.0734	0.8245
25	Employed	0.9566	0.0208	0.0215	0.0011

25	unemployed	0.0915	0.795	0.1121	0.0014
25	out of work force	0.1687	0.0744	0.7536	0.0032
25	m.employed	0.0805	0.0156	0.065	0.8389
26	Employed	0.9609	0.0191	0.0192	0.0008
26	unemployed	0.0808	0.8126	0.1057	0.001
26	out of work force	0.1603	0.0763	0.7612	0.0022
26	m.employed	0.0766	0.0138	0.0575	0.8522
27	Employed	0.9648	0.0175	0.0171	0.0006
27	unemployed	0.0712	0.8288	0.0994	0.0007
27	out of work force	0.1522	0.0781	0.7682	0.0015
27	m.employed	0.0727	0.0122	0.0508	0.8643
28	Employed	0.9682	0.0161	0.0152	0.0004
28	unemployed	0.0626	0.8436	0.0933	0.0005
28	out of work force	0.1444	0.0799	0.7747	0.001
28	m.employed	0.069	0.0107	0.0447	0.8755
29	Employed	0.9713	0.0148	0.0136	0.0003
29	unemployed	0.0549	0.8573	0.0874	0.0004
29	out of work force	0.1369	0.0817	0.7807	0.0007
29	m.employed	0.0654	0.0094	0.0394	0.8858
30	Employed	0.9741	0.0136	0.0121	0.0002
30	unemployed	0.0482	0.8698	0.0818	0.0002
30	out of work force	0.1297	0.0835	0.7863	0.0005
30	m.employed	0.0619	0.0083	0.0347	0.8952
31	Employed	0.9766	0.0124	0.0107	0.0002
31	unemployed	0.0422	0.8813	0.0764	0.0002
31	out of work force	0.1228	0.0853	0.7916	0.0003
31	m.employed	0.0585	0.0073	0.0305	0.9037
32	Employed	0.9789	0.0114	0.0096	0.0001
32	unemployed	0.0368	0.8917	0.0713	0.0001
32	out of work force	0.1163	0.0871	0.7965	0.0002

32	m.employed	0.0553	0.0064	0.0267	0.9116
33	Employed	0.9809	0.0105	0.0085	0.0001
33	unemployed	0.0322	0.9013	0.0665	0.0001
33	out of work force	0.11	0.0889	0.801	0.0001
33	m.employed	0.0522	0.0056	0.0235	0.9187
34	Employed	0.9828	0.0096	0.0076	0.0001
34	unemployed	0.0281	0.91	0.0619	0.0001
34	out of work force	0.104	0.0906	0.8053	0.0001
34	m.employed	0.0493	0.0049	0.0206	0.9253
35	Employed	0.9844	0.0088	0.0067	0.0001
35	unemployed	0.0245	0.9179	0.0576	0
35	out of work force	0.0983	0.0924	0.8092	0.0001
35	m.employed	0.0464	0.0043	0.018	0.9313
		Measurement probabilities			
	State	Employed	unemployed	out of work force	m. employed
	employed	0.9705	0.0115	0.018	0
	unemployed	0.0687	**0.031**	**0.9003**	0
	Out of work	**0.5963**	**0.2285**	**0.1751**	0.0001
	m.employed	0.0286	0.0002	0.0043	0.9669

However, from the longitudinal bivariate residuals of Model B we can see that time trend is captured well and the 1st and 2nd order autocorrelations are adequately addressed as the scores are non- significant (less than 3.84).These Lag1 and Lag2 L-BVRs represent a substantial reduction in the amount of 1st and 2nd order autocorrelation(compare to table 8.3).

Table8.5. Longitudinal Bivariate Residuals for Model B

Indicators	empstatus
empstatus	.
Covariates	es
age	1.2608
Longitudinal	es
Time	1.8602
Lag1	3.1044
Lag2	0.6511

8.3.3. Model 3: Mover Stayer (Model C)

Since it was quite evident from discussed probability that being employed persistently remained in highest proportion over the time therefore it was logical to postulate for the special variant of mixture models. In mover stayer case where one class remains on the same change pattern conditional to the previous states and other moves from one state to another. In this section we explain two special cases of these models one with no covariates and the second for basic comparisons for 2 cohorts.

8.3.3. 1.Unconditional case

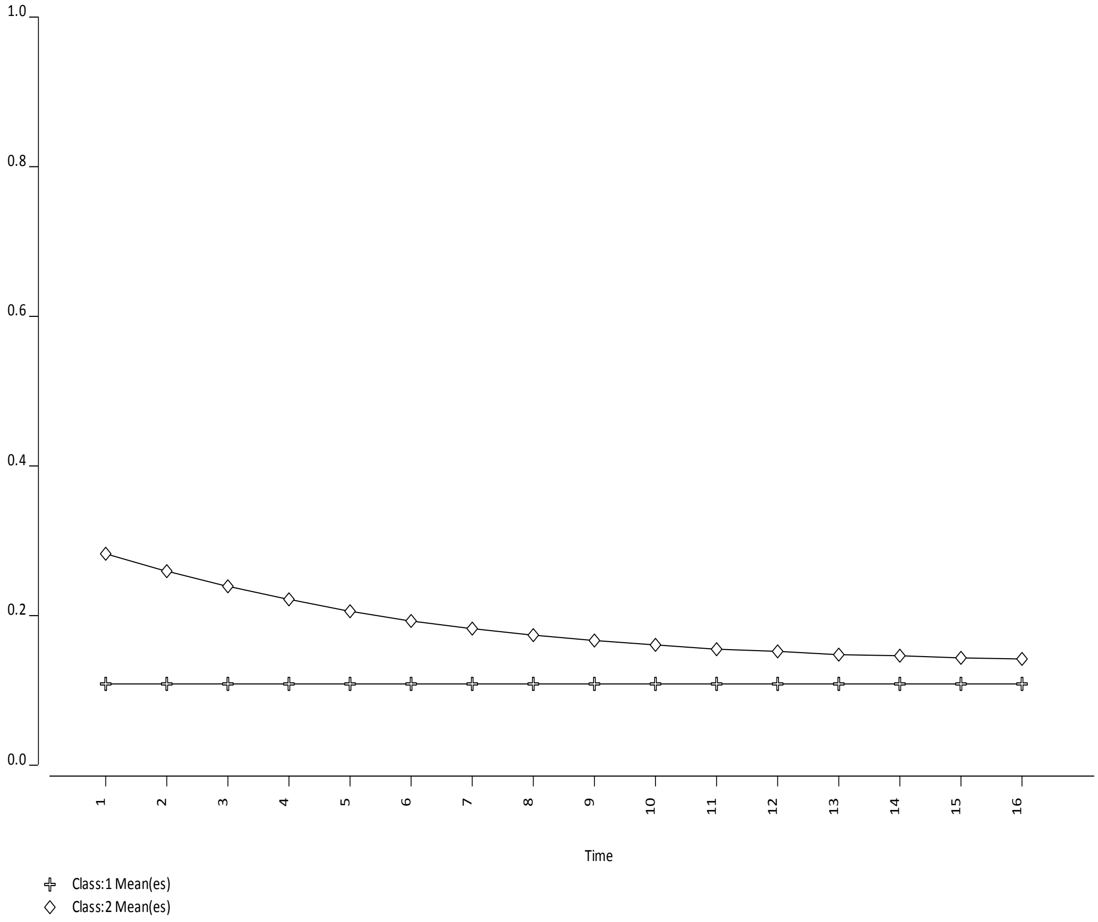
The table and graphical view based on average profile for unconditional mixtures helps to understand the conditional probability distribution of the 2 classes, in first row we have class size where first class has more likely cases of employed individuals and relatively less cases of being discouraged labor force compared to class 2.

Table.8.5. Profile for Mixtures (Model C)

Profile						
	Class		**State**			
	1	**2**	**1**	**2**	**3**	**Overall**
Size	0.3638	0.6362	0.7609	0.2168	0.0223	

ES						
Employed	0.8258	0.7005	0.9502	0.1063	0	0.746
unemployed	0.0398	0.056	0.0223	0.1527	0	0.0501
out of work force	0.1241	0.2145	0.0275	0.7409	0.0051	0.1816

Fig 8.1.Mixtures of movers and stayers



Following table in usual reports the class sizes along with the initial probability or remain in the class of employed, unemployed or discouraged at the age of 21, the segregation of transitioning from one state to another over the age span is now described for both classes and conditional to time in the given table 8.6.

Since the true states are less than the observed states so the choice of excluding one state is justified in the overlap nature of job status measurements, the status of being employed is mutually exclusive to other categories but the status of unemployed or being inactive is somehow ambiguous in survey measurement it was revealed through v high measurement reported for being unemployed in 4 status case for assumed one population. To check the hypothesis, we merged the 2 categories of being unemployed and out of labor force as a true status of being inactive, the state 2 and state 3 were merged as being inactive which reduced the measurement remarkably compared to the above models (10% low compared to 90%).

From the transition table we can see that stayer's class does not move over the categories or employment status whereas for movers we can find inter-status change probabilities. Here Given a case is in 'employed' state (State 1) at time t-1, the probability of remaining in this state at time t is .9 2 and Given a case is in the other state (being inactive implying out of labor force or unemployed) at time t-1, the probability of remaining in employed state at next time is .74. This shows change from employed state to the second state is .26. The bivariate residuals are still high but less than manifest Markov.

Table 8.6. Estimated values model for unconditional mover stayers

Class		**Class**		
		stayers	**movers**	
		0.3638	0.6362	
		employed	inactive	m.employed
stayers		0.8538	0.1358	0.0104
movers		0.5019	0.4357	0.0624
	Transition probabilities			
Class	State[-1]	employed	inactive	m.employed
stayers	employed	1	0	0
stayers	inactive	0	1	0
stayers	m.employed	0	0	1

movers		employed		0.9234	0.0729	0.0037	
movers		inactive		0.249	0.7477	0.0033	
movers		m.employed		0.1639	0.0529	0.7831	
		Measurment probabilities					
State		employed	unemployed	out of work force	m.employed		
employed		0.9502	0.0223	0.0275	0		
inactive		0.1063	0.1527	0.7409	0		
m.employed		0	0	0.0051	0.9949		

8.3.3. 2.Conditional Mover Stayer (Model D)

Draw backs of step 1 approach were highlighted in last section of chapter 2 where many studies were mentioned suggesting for usefulness of step 3 approach , the particular study suggesting including direct effect of covariate at first step was proposed by B. Muthén and Asparouhov (2002a) who suggested to include covariates also at first step of class enumeration if theoretically researcher justifies for the contribution of covariates for class formation . In this section we present conditional models (standard 1 step approach) followed by step 3 Markov models with direct effect of some important covariates in step1.

In the following, the same mover stayer version discussed above is tested with cohorts' effects. Gendered and racial differences were also observed for the transitions between employment statuses. The included covariates were highly significant. Very interesting comparisons of the relevant class sizes conditional to the cohort gender and race differences are reported in the given table. Overall class 2 is differing for being stayer or mover in both cohorts for males and for the ethnicity differences including the gendered differences are also stark for same cases. Over the cohort's males are more likely to be stayers in both cohorts except for race 3 and race 4 for 97 cohort. Here of these age period male individuals of are more likely to be transitioning from employed to be inactive. Females are also more persistent in being the initial status over the course of time with few exceptions of race 3 and race 4 individual, who are more likely to move from being employed to be inactive group.

Table 8.7. Average Probabilities for Conditional Mover Stayer

gender	cohort	ethnicity	Movers	Stayers
1	**79**	**black**	0.3296	0.6704
1	**79**	**Hispanic**	0.3959	0.6041
1	**79**	**Nonblack/non-Hispanic**	0.4662	0.5338
1	**97**	**black**	0.4228	0.5772
1	**97**	**Hispanic**	0.494	0.506
1	**97**	**Nonblack/non-Hispanic**	0.5655	0.4345
1	**97**	**Mixed race**	0.6343	0.3657
2	**79**	**black**	0.2329	0.7671
2	**79**	**Hispanic**	0.288	0.712
2	**79**	**Nonblack/non-Hispanic**	0.3503	0.6497
2	**97**	**black**	0.3114	0.6886
2	**97**	**Hispanic**	0.3761	0.6239
2	**97**	**Nonblack/non-Hispanic**	0.4455	0.5545
2	**97**	**Mixed race**	0.5171	0.4829

For model D the following tables presents the estimated values of probability membership to different statuses over age for staying in same class or moving. Overall pattern of transitions is consistent to unconditional case, but the diagonal entries reveal that covariates have impacted substantially the movement chance to be inactive compared to un accounted role of covariates in measuring transitions.

Table 8.8. Conditional Mover Stayer over Age

Initial status probability					
		Class	**Employed**	**unemployed**	**Out of work force**
		stayers	0.9161	0.0754	0.0085
		movers	0.4649	0.4691	0.066
age	**Class**	**Transition probabilities**	**1**	**2**	**3**
21	**stayers**	**employed**	1	0	0
21	**stayers**	**inactive**	0	1	0
21	**stayers**	**m.employed**	0	0	1
21	**movers**	**employed**	**0.8145**	0.1702	0.0153
21	**movers**	**inactive**	0.3616	**0.6294**	0.0089
21	**movers**	**m.employed**	0.1908	0.0748	0.7344
22	**stayers**	**employed**	1	0	0
22	**stayers**	**inactive**	0	1	0
22	**stayers**	**m.employed**	0	0	1
22	**movers**	**employed**	0.8365	0.1519	0.0116
22	**movers**	**inactive**	0.3365	0.6572	0.0063
22	**movers**	**m.employed**	0.1841	0.0662	0.7497
23	**stayers**	**employed**	1	0	0
23	**stayers**	**inactive**	0	1	0

23	**stayers**	**m.employed**	0	0	1
23	**movers**	**employed**	0.8562	0.135	0.0088
23	**movers**	**inactive**	0.3119	0.6837	0.0044
23	**movers**	**m.employed**	0.1774	0.0584	0.7642
24	**stayers**	**employed**	1	0	0
24	**stayers**	**inactive**	0	1	0
24	**stayers**	**m.employed**	0	0	1
24	**movers**	**employed**	0.8737	0.1197	0.0067
24	**movers**	**inactive**	0.2882	0.7088	0.003
24	**movers**	**m.employed**	0.1706	0.0516	0.7778
25	**stayers**	**employed**	1	0	0
25	**stayers**	**inactive**	0	1	0
25	**stayers**	**m.employed**	0	0	1
25	**movers**	**employed**	0.8892	0.1058	0.005
25	**movers**	**inactive**	0.2654	0.7325	0.0021
25	**movers**	**m.employed**	0.164	0.0454	0.7906
26	**stayers**	**employed**	1	0	0
26	**stayers**	**inactive**	0	1	0
26	**stayers**	**m.employed**	0	0	1
26	**movers**	**employed**	0.9029	0.0933	0.0038
26		**inactive**	0.2437	0.7548	0.0015

	movers				
26	**movers**	**m.employed**	0.1573	0.04	0.8027
27	**stayers**	**employed**	1	0	0
27	**stayers**	**inactive**	0	1	0
27	**stayers**	**m.employed**	0	0	1
27	**movers**	**employed**	0.915	0.0822	0.0028
27	**movers**	**inactive**	0.2232	0.7758	0.001
27	**movers**	**m.employed**	0.1508	0.0351	0.8141
28	**stayers**	**employed**	1	0	0
28	**stayers**	**inactive**	0	1	0
28	**stayers**	**m.employed**	0	0	1
28	**movers**	**employed**	0.9257	0.0722	0.0021
28	**movers**	**inactive**	0.2039	0.7954	0.0007
28	**movers**	**m.employed**	0.1444	0.0308	0.8247
29	**stayers**	**employed**	1	0	0
29	**stayers**	**inactive**	0	1	0
29	**stayers**	**m.employed**	0	0	1
29	**movers**	**employed**	0.9351	0.0634	0.0016
29	**movers**	**inactive**	0.1859	0.8136	0.0005

29	**movers**	**m.employed**	0.1382	0.0271	0.8348
30	**stayers**	**employed**	1	0	0
30	**stayers**	**inactive**	0	1	0
30	**stayers**	**m.employed**	0	0	1
30	**movers**	**employed**	0.9433	0.0555	0.0012
30	**movers**	**inactive**	0.1691	0.8306	0.0003
30	**movers**	**m.employed**	0.1321	0.0237	0.8442
31	**stayers**	**employed**	1	0	0
31	**stayers**	**inactive**	0	1	0
31	**stayers**	**m.employed**	0	0	1
31	**movers**	**employed**	0.9505	0.0486	0.0009
31	**movers**	**inactive**	0.1535	0.8462	0.0002
31	**movers**	**m.employed**	0.1262	0.0208	0.8531
32	**stayers**	**employed**	1	0	0
32	**stayers**	**inactive**	0	1	0
32	**stayers**	**m.employed**	0	0	1
32	**movers**	**employed**	0.9568	0.0425	0.0007
32	**movers**	**inactive**	0.1392	0.8607	0.0002
32	**movers**	**m.employed**	0.1204	0.0182	0.8614

33	**stayers**	**employed**	1	0	0
33	**stayers**	**inactive**	0	1	0
33	**stayers**	**m.employed**	0	0	1
33	**movers**	**employed**	0.9624	0.0371	0.0005
33	**movers**	**inactive**	0.1259	0.874	0.0001
33	**movers**	**m.employed**	0.1149	0.0159	0.8692
34	**stayers**	**employed**	1	0	0
34	**stayers**	**inactive**	0	1	0
34	**stayers**	**m.employed**	0	0	1
34	**movers**	**employed**	0.9672	0.0324	0.0004
34	**movers**	**inactive**	0.1138	0.8861	0.0001
34	**movers**	**m.employed**	0.1095	0.0139	0.8766
35	**stayers**	**employed**	1	0	0
35	**stayers**	**inactive**	0	1	0
35	**stayers**	**m.employed**	0	0	1
35	**movers**	**employed**	0.9714	0.0283	0.0003
35	**movers**	**inactive**	0.1027	0.8973	0
35	**movers**	**m.employed**	0.1043	0.0121	0.8836

For the measurement error we can see that like previous models the status of being employed either full or military is distinctive and well measured with relatively quiet less margin of error There is further reduction observed in measurement error when the second and third status is merged into inactive category. The fourth status is though not of theoretical interest and its initial state probability is quite low in each variants discussed so far but we have not deleted it for taking categories completely in measurement.

Table 8.9. Conditional Mover Stayer Measurement model

True State	employed	unemployed	out of work force	m. employed
employed	0.9482	0.0239	0.0278	0
inactive	**0.0741**	0.7729	0.153	**0.0001**
m.employed	0	0	0.004	0.996

The following table reports that latent construct of employment status is independent to cohort as the reported BVR is below 3 .84 implying insignificant impact of cohort effect). Further over time the associations between times are not well captured by this specification since the BVR are substantially higher than standard level.

Table 8.10. Bivariate Residuals of Conditional Mover Stayer

Indicators	**es**
es	.
Covariates	**es**
age	0
gender	0
cohort	2.3064
ethnicity	0
Longitudinal	**es**
Time	4.7123
Lag1	21.9405
Lag2	21.324

8.3.3. 3. Step 3 Markov Approach

In the following we conducted the steps described in section 4 methodology for various variants of step 3 analyses. Much is explained in earlier sections for other mixture models. Here by following the steps as mentioned in previous contexts first we estimated the partially unconditional model where age was included as a predictor of class enumeration. Usual diagnostics reveled in this case that 3 cluster case was most suitable to be representative of the considered two populations. Then from the obtained posterior classifications of individuals across clusters in step 2 we conditioned the cluster in step 3 on the given covariates. Variants were based on modal and proportional assignment for ML based and BCH based corrections and the transitions were conditioned on age and cohorts and the combinations separately. The variant where the transitions were dependent only on time turned to be best amongst the options given the highest log likelihood value and lowest values for the relative information criteria.

Table 8.11. Step 3 Markov Models

Markov	3 Step 3 states Markov	LLHD	BIC(LLHD)	AIC(LLHD)	AIC3(LLHD)	Npr
Variant1	3State ML PR	-108436	216949.53	216887.8	216896	8
Variant2	3StateML MD	-79502	159140.38	159032.3	159046	14
Variant 3	3StateBCHPR	-108383	217134.47	216841.2	216879	38
Variant4	3StateBCHMD	-108398	216931.85	216823.8	216838	14

The transitions are well transparent in the given table where measurement model is based on variant 2 of model assignment. Direct Comparison to step 1 measurement error probabilities presented earlier is not suitable because in step 1 we did proportional assignment of observations across clusters whilst in step 3 analysis model assignments is made due to sparse nature and high frequency tables. Still, we can see quite accurate measurements of the constructs /true states compared to the table measurements of step 1(table 8.8). This was mentioned somewhere in last section of literature review of methods that inclusion of covariates at step 1 is quite advantageous if the class differentiation is theoretically meaningfully driven by the covariates. In the case discussed above we can safely assume that time zone can be the valid predictor of class separation so their inclusion also at first step 1 for class enumeration enhances to bring accurate results in third step of analysis.

Table 8.12. Profile by Step 3 Markov Model

		State change probabilities		
age	**State [-1]**	**1**	**2**	**3**
21	**employed**	0.9994	0.0005	0.0001
21	**inactive**	0	1	0
21	**discouraged**	0	0.0001	0.9999
22	**employed**	0.9997	0.0003	0.0001
22	**unemployed**	0	1	0
22	**discouraged**	0	0.9999	0.0001
23	**employed**	0.9998	0.0001	0
23	**unemployed**	0	1	0
23	**discouraged**	0	1	0
24	**employed**	0.9999	0.0001	0
24	**unemployed**	0	1	0
24	**discouraged**	0	1	0
25	**employed**	0.9999	0	0
25	**unemployed**	0.0001	0.9999	0
25	**discouraged**	0.0001	0.9999	0
26	**employed**	1	0	0
26	**unemployed**	0.9999	0.0001	0
26	**discouraged**	0.0004	0.9996	0
27	**employed**	1	0	0
27	**unemployed**	1	0	0
27	**discouraged**	0.0011	0.9989	0
28	**employed**	1	0	0
28	**unemployed**	0.9999	0	0.0001

28	**discouraged**	0.0034	0.9966	0
29	**employed**	1	0	0
29	**unemployed**	0.9996	0	0.0004
29	**discouraged**	0.0107	0.9893	0
30	**employed**	1	0	0
30	**unemployed**	0.9982	0	0.0018
30	**discouraged**	0.0325	0.9675	0
31	**employed**	1	0	0
31	**unemployed**	0.9927	0	0.0073
31	**discouraged**	0.0947	0.9053	0
32	**employed**	1	0	0
32	**unemployed**	0.9707	0	0.0293
32	**discouraged**	0.2458	0.7542	0
33	**employed**	1	0	0
33	**unemployed**	0.8898	0	0.1102
33	**discouraged**	0.5039	0.4961	0
34	**employed**	1	0	0
34	**unemployed**	0.6628	0	0.3372
34	**discouraged**	0.7599	0.2401	0
35	**employed**	1	0	0
35	**unemployed**	0.3237	0	0.6763
35	**discouraged**	0.9079	0.0921	0

8.4. Single Cohort Analysis

8.4.1. Model Selection

Since the data is highly sparse so we followed the usual criteria of model selections with special emphasis on the longitudinal performance of the model measured through longitudinal reduced chances of codependence between the variables considered. Visual inspections also helped us to finalize the model. If the additional categories were not separable over some range, we preferred the more parsimonious and logically meaningful model. Final selection was made based on Bootstrap likelihood ratio test. Models with possible states and classes are labeled in the given table.

Why we conducted separate analysis on single cohort was driven by 2 reasons. With reduced size of data, we wanted to check other variants of models (homogenous and heterogeneous transitions, mixture variants) and wanted to impose further model selection techniques (validation and bootstrapping ; these techniques took too long when we tested on both cohorts jointly).

For the single cohort data 3 state 2 class model turned to be best performing in terms of low BVR, low classification errors and score for entropy R2. In the table same state class combinations are also tested with covariates. The selected model choice is verified by bootstrapping the difference between higher and lower states and classes. The pair of difference between 3state mixture and 3 state one class population favored for the addition of class (i.e., mixture). As we can see from the table homogenous versions of models are more fit to the data structure. When we imposed the heterogeneous transitions(see models with prefix HET) bivariate residuals started to rise with an increasing trend indicating the imposed structure as an incorrect choice. Though the last variants are better in terms of classification and entropy R2, but the BVR for the Het7 is high while number of parameters are highest for the last HET1 with high classification errors and with sizeable reduction in entropy R2.

Table 8.13. Markov models specifications for single cohort

		LL	BIC(LL)	AIC(LL)	AIC3(LL)	Npar	Max. BVR	Class.Err.	Entropy R^2
HOM1	1Stat1CMark	-88415.1	176857.2	176836.2	176839.2	3	0	0	1
HOM1	2State1C Mark	-71159.4	142400	142336.9	142345.9	9	0	0.0467	0.7753
HOM1	3-State 1-Class Markov	-62993.2	126139.7	126020.5	126037.5	17	0	0.0476	0.807
HOM1	1-State 2-Class Markov	-88415.1	176866.2	176838.2	176842.2	4	0	0	1
HOM1	2-State 2-Class Markov	-68806.3	137729.8	137638.6	137651.6	13	0	0.0446	0.7906
HOM1	**3-State 2-Class Markov**	**-62577.2**	**125388.7**	**125206.4**	**125232.4**	**26**	**0**	**0.0478**	**0.8049**
HET1	2-State 1-Class Markov	-68885.4	137870	137792.9	137803.9	11	222.3577	0.0439	0.7974
HET1	3-State 1-Class Markov	-62660.2	125527.8	125366.5	125389.5	23	137.1955	0.0467	0.8123
HET1	4-State 1-Class Markov	-61205.6	122762.7	122489.2	122528.2	39	3.0146	0.1262	0.6716
HET1	2-State 2-Class Markov	-68611	137357.3	137252.1	137267.1	15	222.267	0.0421	0.8035
HET1	3-State 2-Class Markov	-62400.4	125089.2	124864.8	124896.8	32	136.1034	0.0461	0.8131
HET1	4-State 2-Class Markov	-61004	122503.8	122118.1	122173.1	55	0.2689	0.1076	0.6793

The reason for making final choice for 3 state cases under homogenous and heterogeneous model specifications was partly due to the subjective choice of making single cohort analysis aligned to previous section and due to the objective constraint of making final pick based on least measurement error. Also, the status of being unemployed and out of labor force is very inaccurately measure under 4 state specifications like previous section. Henceforth we confirmed the selection by parametric bootstrap, the table reports significant p value suggesting in favor of 2 subpopulations with different course of transitions.

Table 8.14. Model selection by bootstrapping

3-State 2-Class Markov/ 3-State 1-Class Markov			
No of cases	8206	8206	
No of time units	131296	131296	
(NPR)	26	17	
BLRT		p-value	s.e.
LLHD of null	-62993.2		
Diff in NPR	9		
-2LL Diff	832.088	0.00	0.001

The last table of this section reports standard three sets of probabilities for given two classes where class 1 consists of more of those individuals who over the given life course remained more active into labor market and class 2 individuals who though were low in proportion but faced different likely scenario of transitioning from being active to inactive labor unit of economy. We can see from the given table that the initial probability to start employment career is quite low for class 2 compared to class 1, and from the transition section its evident that given state of being employed in previous time leads to higher chance of being inactive for class 1 individuals compared to class 2. Lastly reported are measurement error probabilities in similar fashion to previous section.

Table 8.14. Distribution across Mixture Markov

		Class			
		1	2		
		0.7177	0.2823		
		State [=0]			

	Class	employed	inactive	m.employed	
	1	0.6565	0.275	0.0685	
	2	0.3721	0.5954	0.0325	
		State			
Class	State[-1]	employed	inactive	m.employed	
1	employed	0.9816	0.0176	0.0008	
1	inactive	0.3777	0.6138	0.0085	
1	m.employed	0.1896	0.0437	0.7667	
2	employed	0.8297	0.1698	0.0005	
2	inactive	0.1139	0.8856	0.0005	
2	m.employed	0.0069	0.0196	0.9735	
	State	employed	unemployed	out of work force	m. employed
	employed	0.9442	0.035	0.0208	0
	inactive	0.0979	0.1091	0.7929	0.0001
	m.employed	0.0384	0.0136	0.0049	0.9431

8.5. Empirical Application Background

We have started with relatively smaller sample size (6 years of ES data for both cohorts of 1997 and 1979) .ES employment status data is though measured at regular intervals in survey data but actual employment status can be measured on continuous scale if resources allow. Further for testing the continuous change or stability in employment status we found 3 state Markov models as the best fit to data .Considering this in main text we presented the separated analysis on cohort basis and extended the time range from 6 to 16 years consisting of prime youth to post adulthood period(16 to 35 years). We selected the specific age range expecting major transitions between various categories of employment and primarily concentration of change towards being employed at least in any midrange of good deal of time (16 year). The findings did not support our assumptions and were discussed in results section. The models employed tested various variants of Markov specifications starting from simplest Markov (to measure change over time in manifest variable employment status) to latent Markov (to measure unknown types of change in various classes), and finally for finding measurement error, mixtures of Markov models were applied on the same data. Each of the models had different specifications. For one cohort, we also tested about the hypothesis of invariant role of age across classes by estimating homogenous time vs. heterogeneous time effects on latent states. The most parsimonious model selected on various diagnostic criteria like Latent class cluster models in first section was further compared with competing models by parametric bootstrap. The finalized models are discussed fully in results section. Finally, to incorporate the cohort effects followed by gender biases or racial difference Step 3 analyses was conducted on most favored model.

CHAPTER 9

CONCLUSION

9.1. Conclusion

In first chapter of analysis (chapter 5) we applied latent class cluster analysis for exploring the chances of any existing typologies based on job quality profiles in American individuals. We tested employment data for the year 2017 from ("National Longitudinal Surveys (NLS)"; Moore et al., 2000). The nature of data was mixed mode, we could standardize the variables but for performing categorical analysis it was better capture the indicators through categories. We have taken some crucial indicators to judge state of jobs experiences later we took subjective evaluation of quality jobs as a distal outcome for evaluated clusters then the key covariates were tested as predictors of quality heterogeneity in step 1 analysis of covariates. We compared the case with no covariates and found the change in class belonging of individuals. Further step 3 analyses supported the significant contribution of degree level, race and gender for clusters achievers, successful, strugglers, and left ones. We found some evidence of more prevalence of nonwhite and females in left ones and different degree of satisfaction scores from jobs for most successful to others.

In doing all empirical analysis we were keen about some technical issues related to mixture models generally and some specific to LCCA. We discussed and tested on our data for the violation of conditional independence assumption. For this we initially took bivariate residuals based on chi2 as diagnostic measure. Since our data was sparse which does not follows chi 2 distribution therefore for the most parsimonious model selected from bootstrapping difference log likelihood of nested models we calculated asymptotic bivariate residuals by bootstrapping, and after confirming a significant reduction in dependence of observations we incorporated the heterogeneity of observations through inclusion of continuous factor in final model and lastly for this model we again sorted out for reported co dependence of some indicators for the last model through inclusion of direct effects Validation of final models was also tested. Lastly, we applied comparison of various approaches to include the role of covariates or predictors in cluster formation, most parsimonious model was selected and elaborated.

In next chapter by employing latent class regression mixture on longitudinal data we could find prevalence of job satisfiers a non-satisfiers in British households sample (see chapter 6).

We were interested to find the differences in effect sizes of some important job satisfaction indicators. Two groups solution was marked by absolute, relative and bootstrapping model selection techniques. Interestingly the predictors of job satisfaction were significantly different across both groups, but the effect sizes had explanatory power negligible indicating the exercise done to be futile at first glance. Since the objective was exploratory where things could turn as expected or contrary. The exercise done only in this case (compared to other applications) did not support the presence of heterogonous segments with response to chosen indicators. Also, the general hypothesis of the differential impact of chosen job features was negated in this case study implying the homogenous impact of chosen job features existed across both classes. The results urged us to look further for the source of difference could be addressed by subjective background variables since background variables/ covariates come to play their role for finding the source of latent class membership. Therefore, we did conditional analysis with step 3 approaches. Through variants of Step 3 models in regression case we found occupational and gendered differences in two classes of satisfied vs. non satisfied class on job.

In next chapter 7 of growth models in mixture framework y employing longitudinal data over the prime working age of American labor class we attempted to explore growth patterns differences in taking employment course. To serve the objective and for testing the hypothesis of heterogeneous subpopulation within the larger population we employed methods which could measure inter individual differences in intra individual change over time. Four mainstream modeling variants of growth modeling were tested and elaborated in terms of model performance. Lastly status typology was built based on the consensus of model variants.

Though class sizes appeared different under above discussed growth variants but typically three patterns of change were observed under each specification. The highest proportion had those likely cases who remained and grow to be employed and least likely cases of being inactive or unemployed. We labeled that cluster as persistent active. The other relative smaller cluster had initially lower reported cases of being employed which persistently grow in number to be employed over the age. For that cluster, individuals being unemployed following remarkable slows in such status over age and finally consistently becoming part of out of labor force were observed for this cluster. On part of such developments, we labeled the second cluster as inactive. Third segment had lowest number of likely cases with different response patterns of having lowest employment cases which grew to be employed relatively low to persistent active ones, therefore we named this cluster as mediocre active over age.

In the next chapter with same data extending to growth mixture modeling we opted for the second major family of longitudinal categorical data analysis. Here our objective was to find the transitions between various employment statuses over the 16 prime years of working life in USA. We selected the specific age range expecting major transitions between various categories of employment. We primarily looked for concentration or change towards being employed at least in any good deal of time range. The models employed tested various variants of Markov specifications starting from simplest Markov (to measure change over time in manifest variable of employment status) to latent Markov (to measure unknown types of change in various classes and finally to test measurement error mixtures of Markov models were applied on the same data. Each model had different specifications. We tested the hypothesis of invariant role of age across classes by estimating homogenous time vs. heterogeneous time effects on latent states. The most parsimonious model selected on various diagnostic criteria (like LCCA in first section were further compared with competing models by parametric bootstrap). Finally, to incorporate the cohort effects followed by gender biases or racial differences Step 3 analyses was conducted on the most parsimonious model in last. The results revealed significance of racial differences and cohorts for making transitions in employment statuses over time.

9.2. Policy Implications

Broadly we can point out two policy pointers from the core of four modeling schemes applied in empirical data. First is regarding job quality clusters and Markov models of employment status; second is related to model building in mixture models.

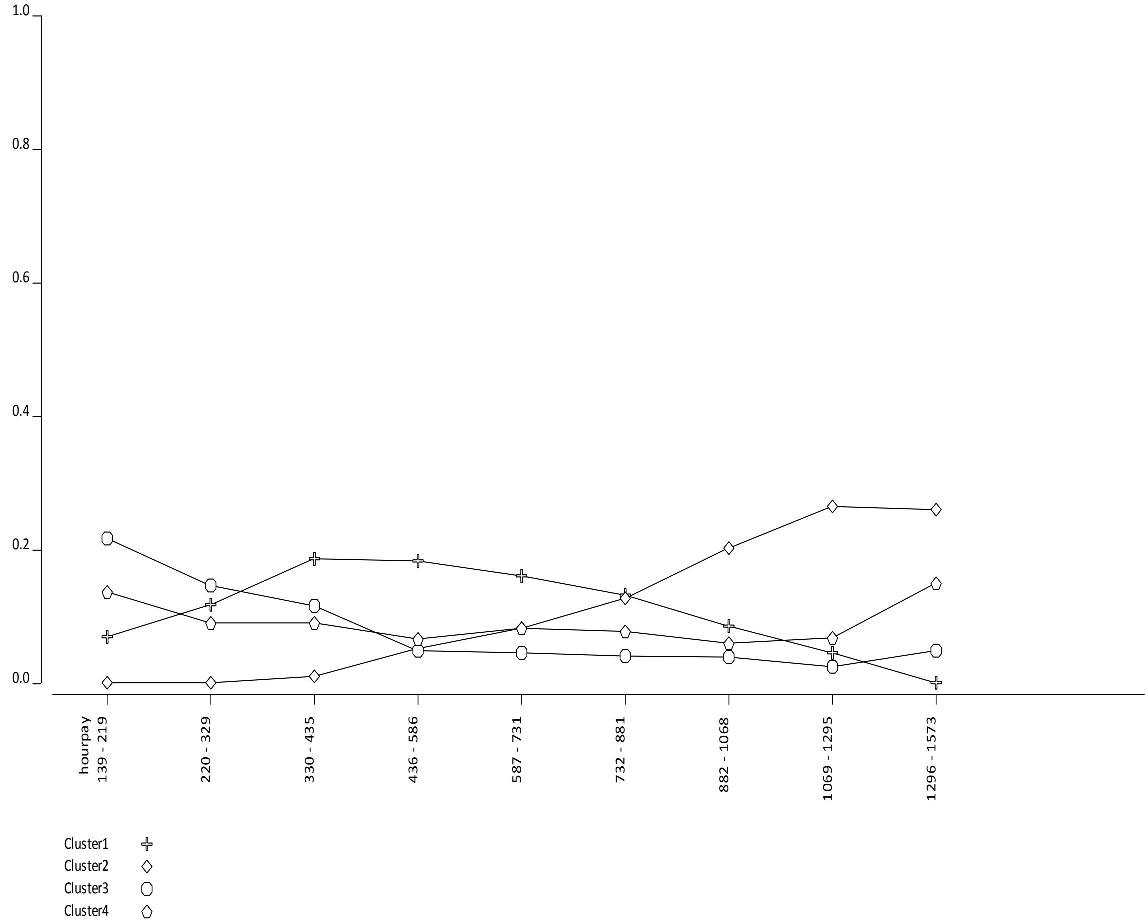
Finding of typologies in other sciences like criminology and psychology is very much in practice since the cluster formation of a like objects minimized the burden of policy targets and focused group interventions can be made. About specific scenario taken in this thesis, clustering of job quality is itself rewarding to look into inside job quality situation of the individuals based on objective characteristics. Since the concept is latent in nature and one average like other economic estimates hides the differentiated experiences of the individual therefore in particular context of micro units focused qualitative analysis should be employed to bring the specific features related to underprivileged quality profiles. Also, when we measure the employment status through Markov models then these models also produce the measurement error probabilities of chosen survey items. The lowest measurement error indicates closest measurement of movement observed in various groups over the categories. These modeling strategies can save the burden and cost of finding employment status each year through survey items and their usefulness to measure survey items is very much appreciated in many scientific domains.

The second pointer is towards model building in mixtures. Especially with respect to model-based cluster analysis we suggest to test and validate the assumptions of standard cluster model before discussing the clusters or typologies for class differences.

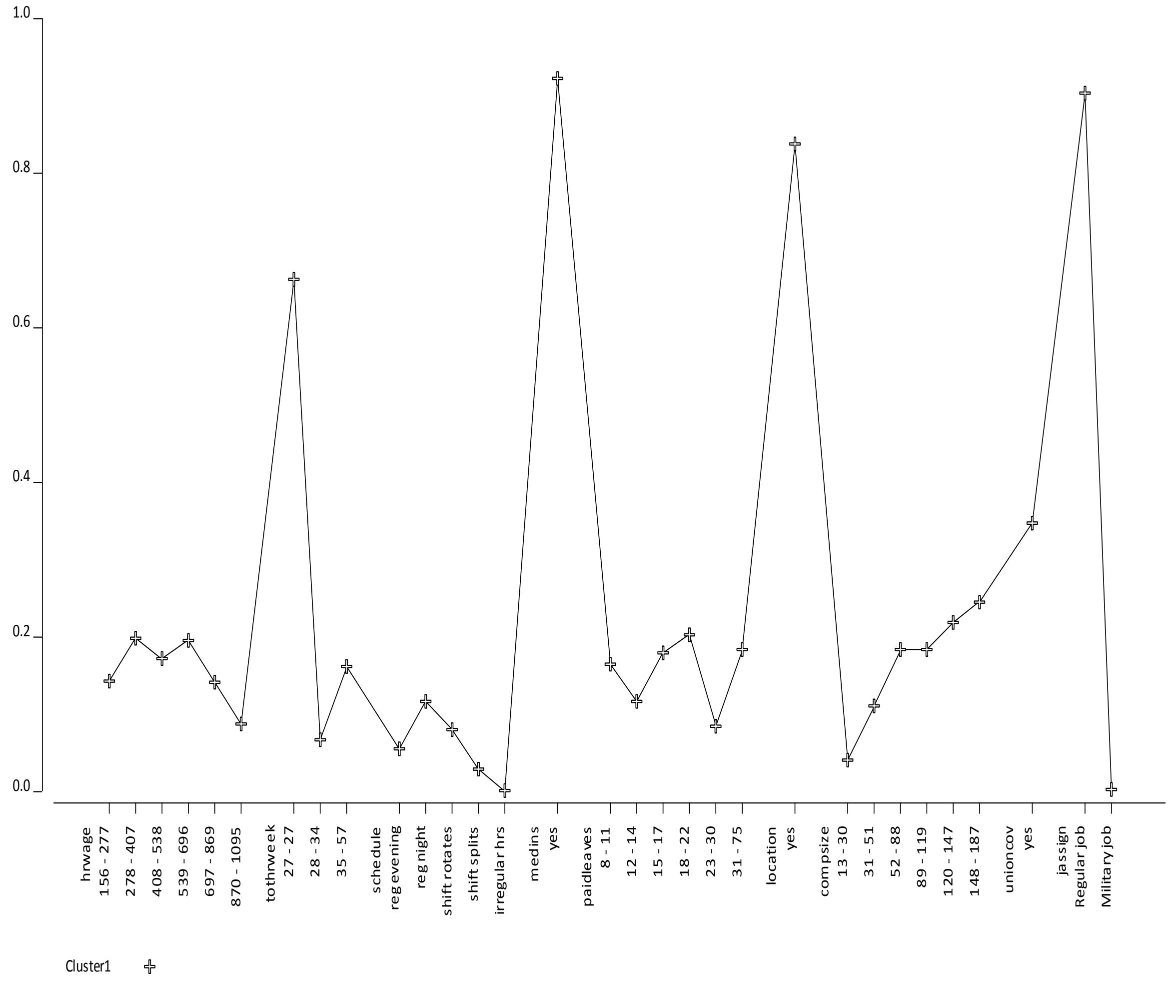
APPENDIX A

A. VISULAIZATION OF CLUSTERS

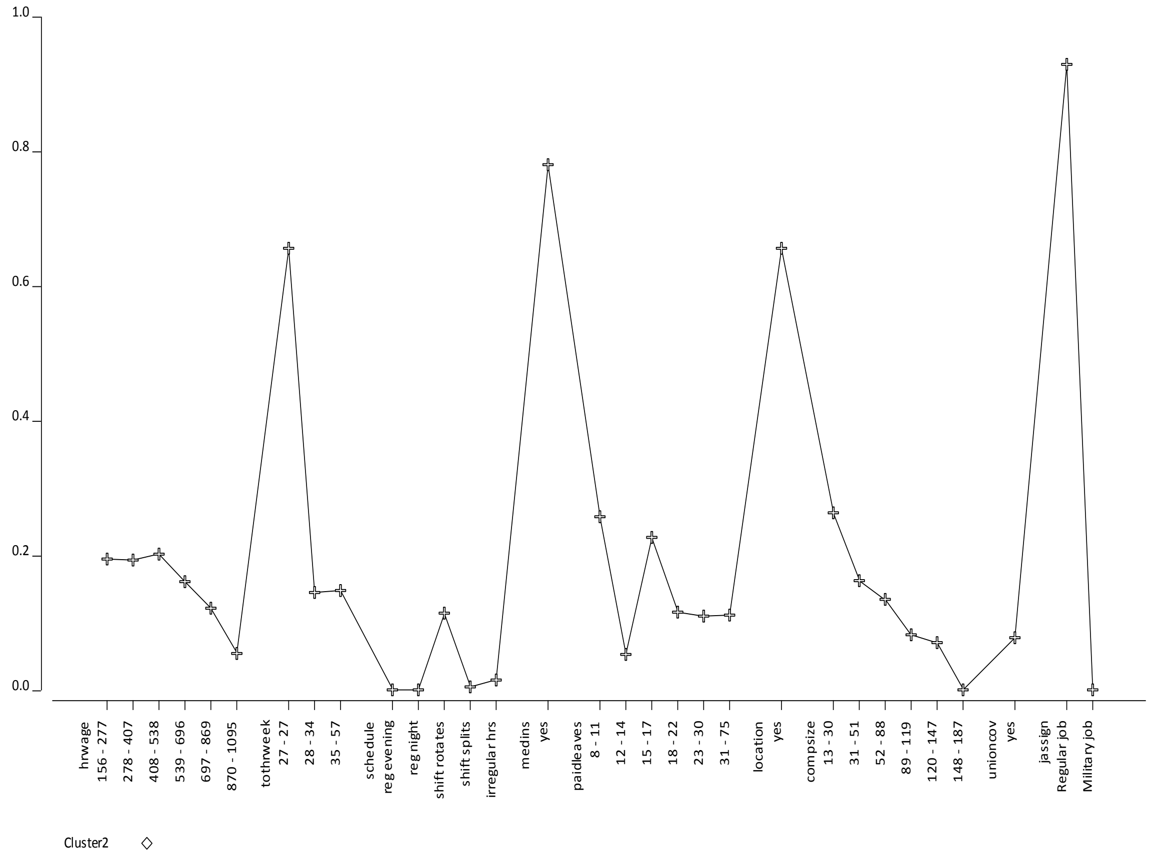
In the following we are providing some graphs for endorsing differences of job quality indicators through visual inspections.



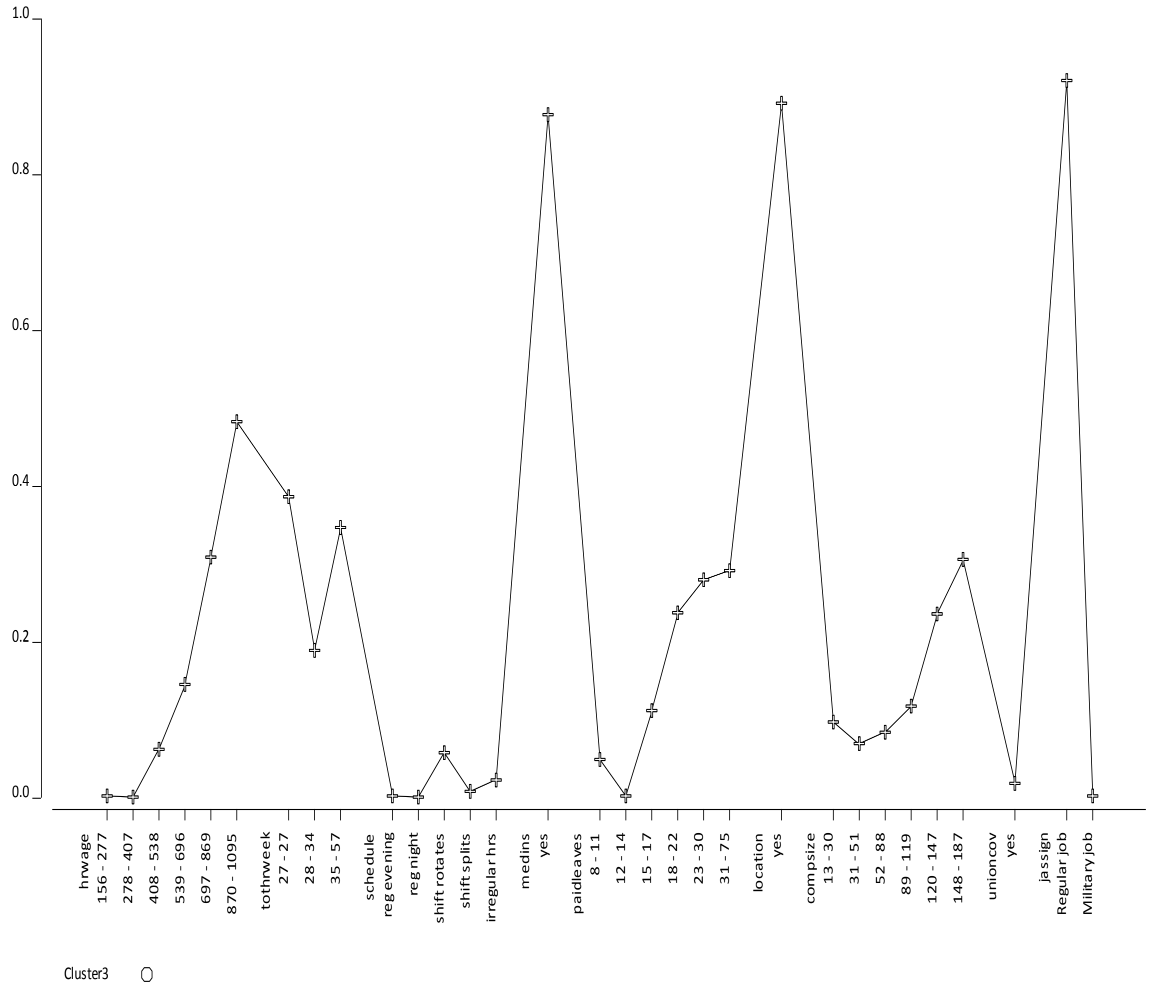
FigA1.wage differences across clusters



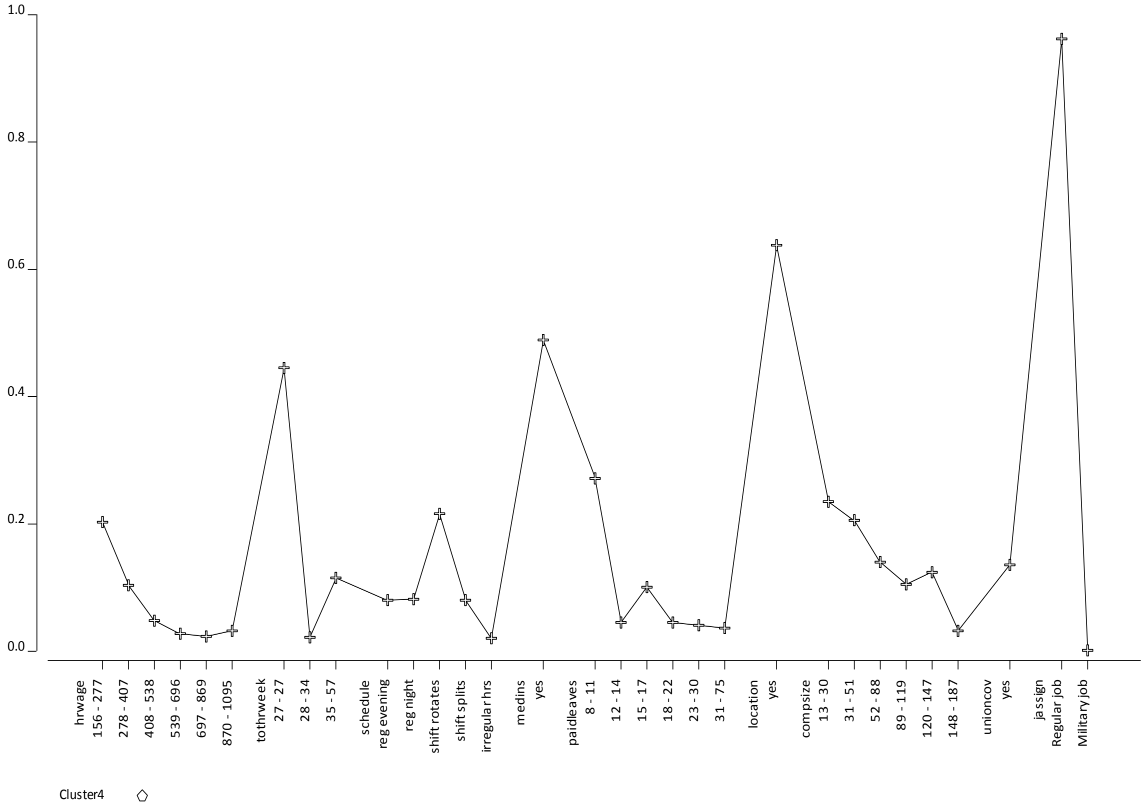
FigA2.Indicators distribution for Successful



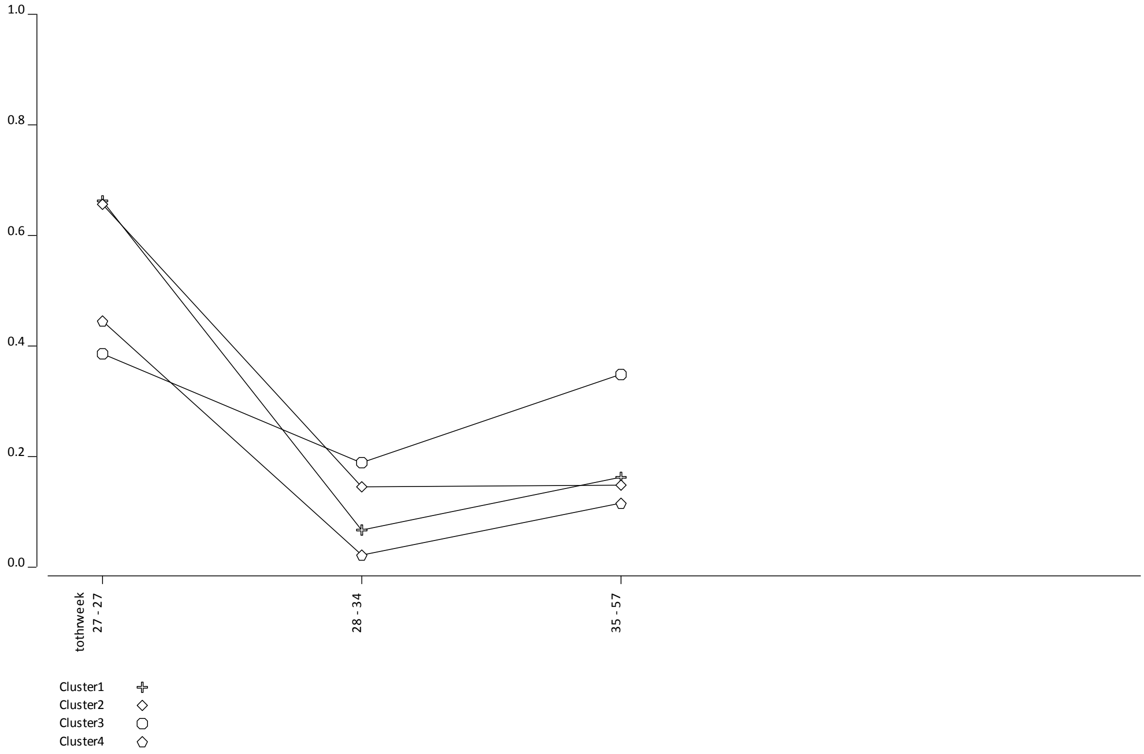
FigA3.Indicators distribution for Strugglers



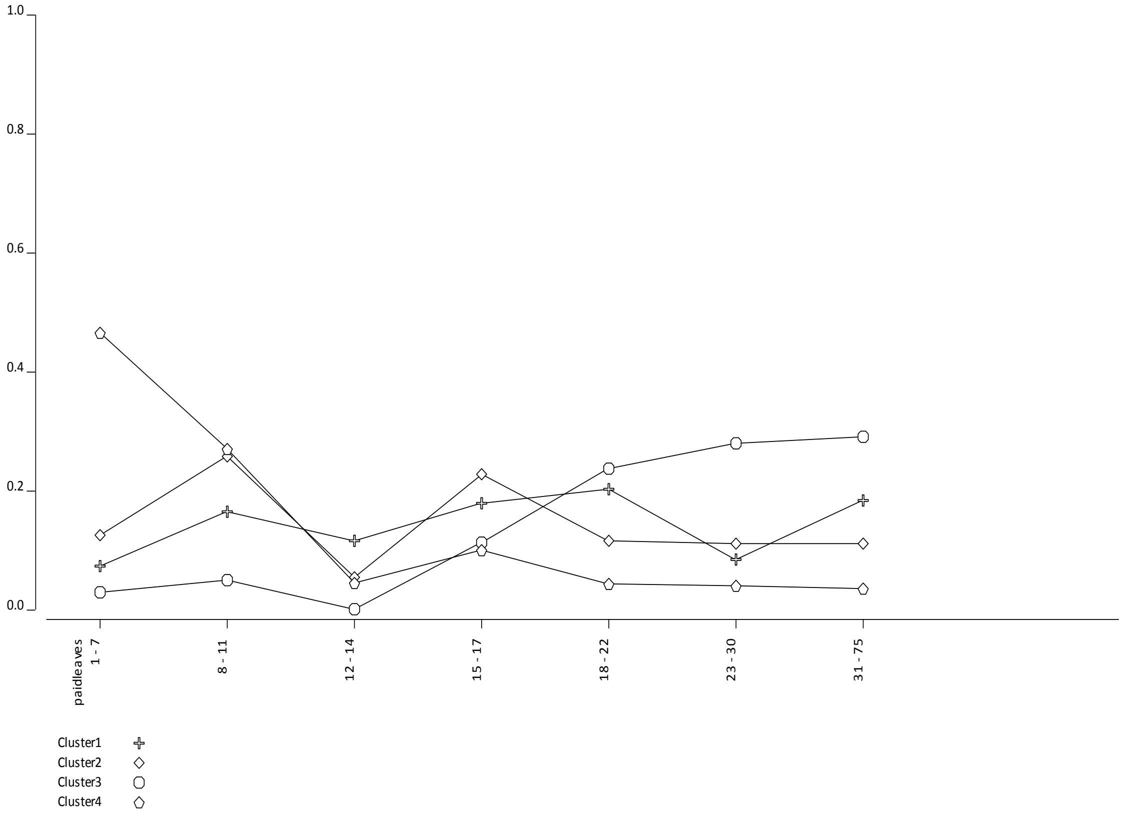
FigA3.Indicators distribution for Left Ones



FigA4.indicators distribution for Achievers



FigA5. Working time distribution across clusters



FigA6. Paid leaves distribution across clusters

Appendix B

B1.Variable Detail for latent class models

Variable Detail for latent class models	
9 Indicators	
hrwage	Continuous
tothrweek	Ordinal
1-22	28.134948
23 - 25	37.719512
26 - 26	40
27 - 30	44.422131
31 - 34	49.692053
35 - 56	61.397321
schedule	Nominal
reg day	1
reg evening	2
reg night	3
shift rotates	4
shift splits	5
irregular hrs	6
medins	Nominal
No	0
yes	1
paidleaves	Ord-User
1-5	1.1871921
6-9	6.1613833
10-11	9.9401709
12-14	12.113095
15-15	14
16 - 19	15.72549
20 - 21	19.92
22 - 26	23.098413
27 - 32	28.913043
33 - 75	73.158491
location	Nominal
No	0
yes	1

compsize	Ord-Fixed
1-10	5.7699387
11-26	18.201133
26 - 45	35.509363
46 - 73	63.980926
74 - 104	121.24771
105 - 124	251.62983
125 - 149	649.71866
150 - 187	5312.5418
unioncov	Nominal
No	0
yes	1
jassign	Nominal
DLI job	0
Regular job	1
Military job	4
3 Covariates	
degree	Num-Fixed
0	0
none”	1
GED	2
High School diploma	3
Bechlors	4
Masters”	5
PhD	6
Professional degree	7
race	Num-Fixed
Black	1
Hispanic	2
Mixed Race (Non-Hispanic)	3
Non-Black / Non-Hispanic	4

gender	Num-Fixed
male	1
female	2

B2: Descriptive statistics for latent class models

gender 1
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 0 1400 2000 2756 3000 121795 1793
gender 2 Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 0 1174 1719 2290 2600 72115 1562
race 1 Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 0 1100 1500 1935 2151 43750 847
race 2 Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 0 1300 1850 2308 2600 57692 723
race 3 Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 0 1153 1732 3844 3008 66667 29

race 4

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	1420	2100	2885	3293	121795	1756

degree 0

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	960	1250	1605	1700	25000	162

degree 1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	1025	1400	1787	1984	57692	176

degree 2

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	1200	1625	2071	2250	121795	450

degree 3

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	1350	1850	2418	2656	56250	97

degree 4

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	1800	2500	3107	3667	64038	154

degree 5

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	2234	3200	3828	4487	68182	49

degree 6

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
930	2738	3314	3702	4314	8413	4

degree 7

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1625	3846	6073	8273	10000	66667	9

medins 0

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
42	1249	1700	2168	2500	56250	24

medins 1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	1671	2238	2771	3261	54651	116

unioncov: 0

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0	1325	1863	2418	2789	56250	143

unioncov: 1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
110	1670	2240	2570	3200	12000	36

Job satisfaction 1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.10	23.00	37.00	31.84	39.00	80.00	13

Job satisfaction: 2

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.10	26.00	37.00	32.85	39.00	96.00	24

Job satisfaction: 3

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	28.00	37.00	33.33	39.00	96.00	49

Jobsatisfaction: 4

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	25.00	37.00	32.46	39.00	97.00	57

Jobsatisfaction: 5

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.10	27.00	37.00	32.95	39.00	97.00	112

Jobsatisfaction: 6

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.10	25.00	37.00	32.51	39.00	97.90	255

Job satisfaction: 7

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.20	20.00	35.00	30.19	38.00	97.90	159

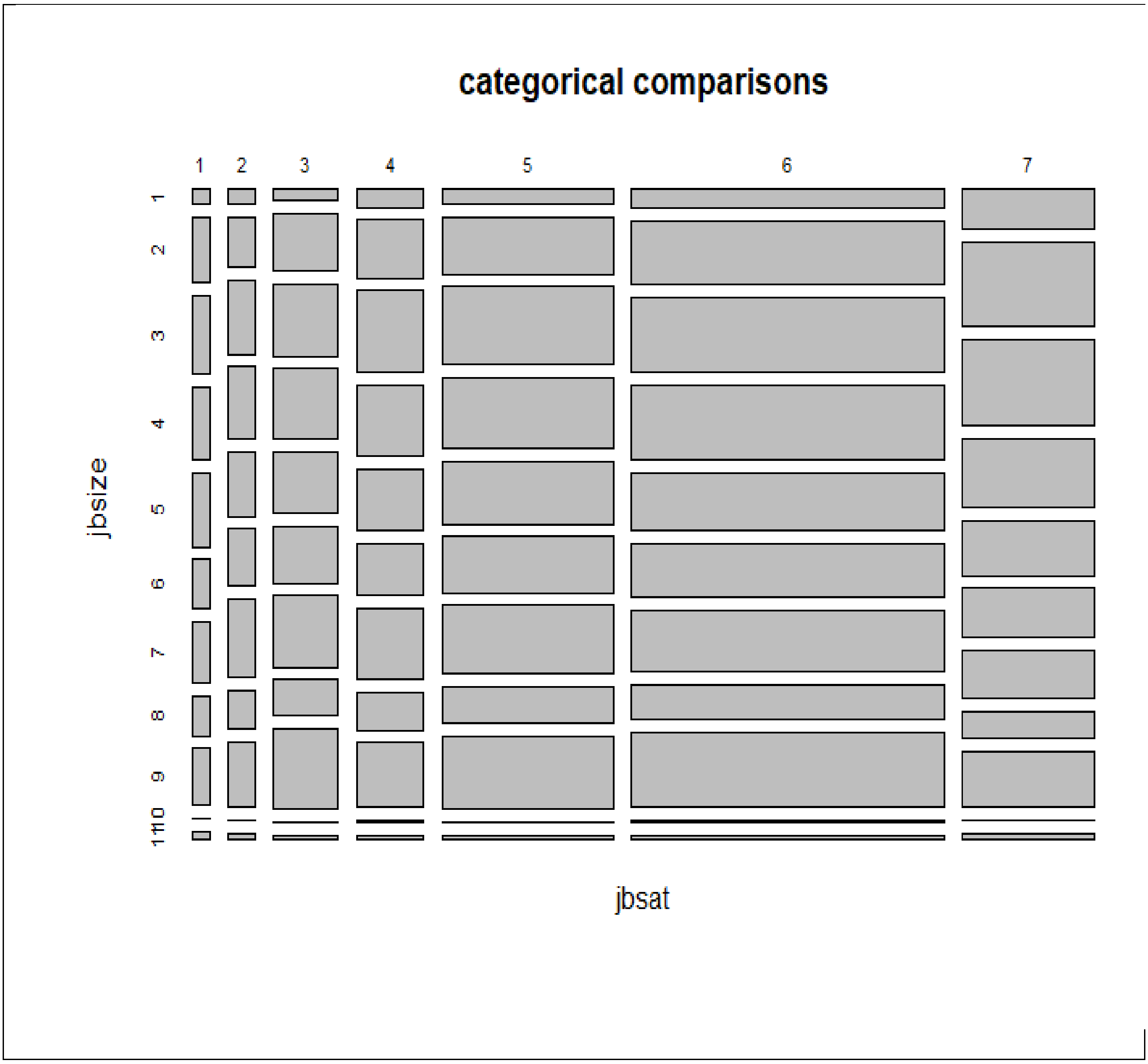


Table B3: Employment status pro portion in cohorts

cohort	employed	unemployed	out of labor force	militry employed
79	81339	5708	21732	2763
97	50916	3432	11758	1575

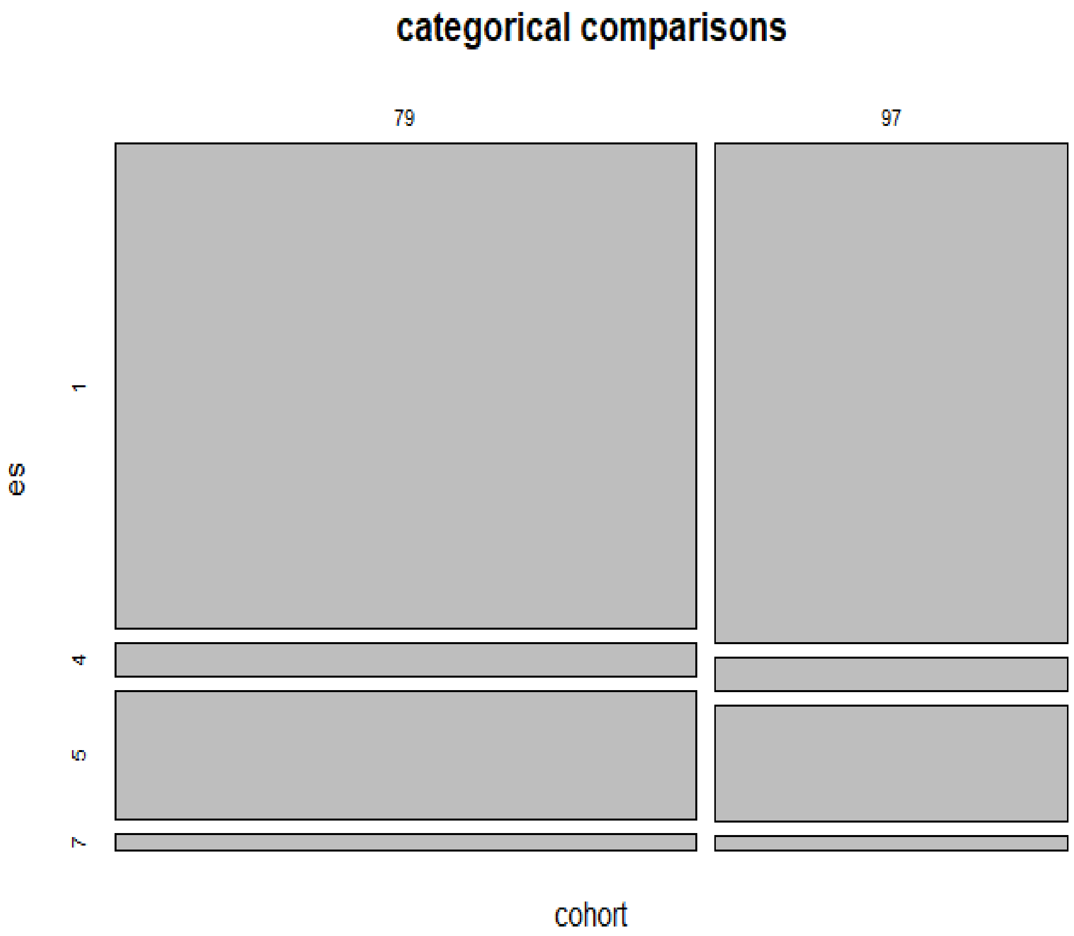


Table B4. Variables information for data employed in regression mixture

PSU ID	psu	3081
Stratum ID	strata	1599
Case ID	pidp	8653
Dependent		
jbsatis	Ord-Fix	7
cdissatis	1	1
mdissatis	2	2
somedissatis	3	3
neither sat or dissat	4	4

somewhat satisfied	5	5
mostly satisfied	6	6
completely satisfied	7	7
Independent		
jbsize	Num-Fix	11
1 - 2	1	1
3 - 9	2	2
10 - 24	3	3
25 - 49	4	4
50 - 99	5	5
100 - 199	6	6
200 - 499	7	7
500 - 999	8	8
1000 plus	9	9
fewer than 25	10	10
25 or more	11	11
jbterm_dv	Num-Fixed	6
permanentjob	1	1
seaswork	2	2
contractfixedt	3	3
agencyhiring	4	4
casual work	5	5
not permanent	6	6
jbhrs	Num-Fix	256
0	0	0
0.1	0.1	0.1
0.2	0.2	0.2
0.5	0.5	0.5

1	1	1
1.2	1.2	1.2
1.5	1.5	1.5
2	2	2
2.5	2.5	2.5
3	3	3
...		
88	88	88
89	89	89
90	90	90
91	91	91
92	92	92
95	95	95
96	96	96
97	97	97
97.9	97.9	97.9
hiqual_dv	Num-Fixed	6
Degree	1	1
Other higher	2	2
A level etc	3	3
GCSE etc	4	4
Other qual	5	5
No qual	9	9

APPENDIX C

Table C1 Parametric Bootstrapped Score for 3 class solution

3-Class Model				
Number of cases	8114			
Number of replications	111542			
Number of par	17			
Chi-squared Statist			Bootstrap	
Degrees of freedom	8097	p-value	p-value	s.e.
L-squared (L^2)	65885.79	2.0e-8866	0.216	0.0184

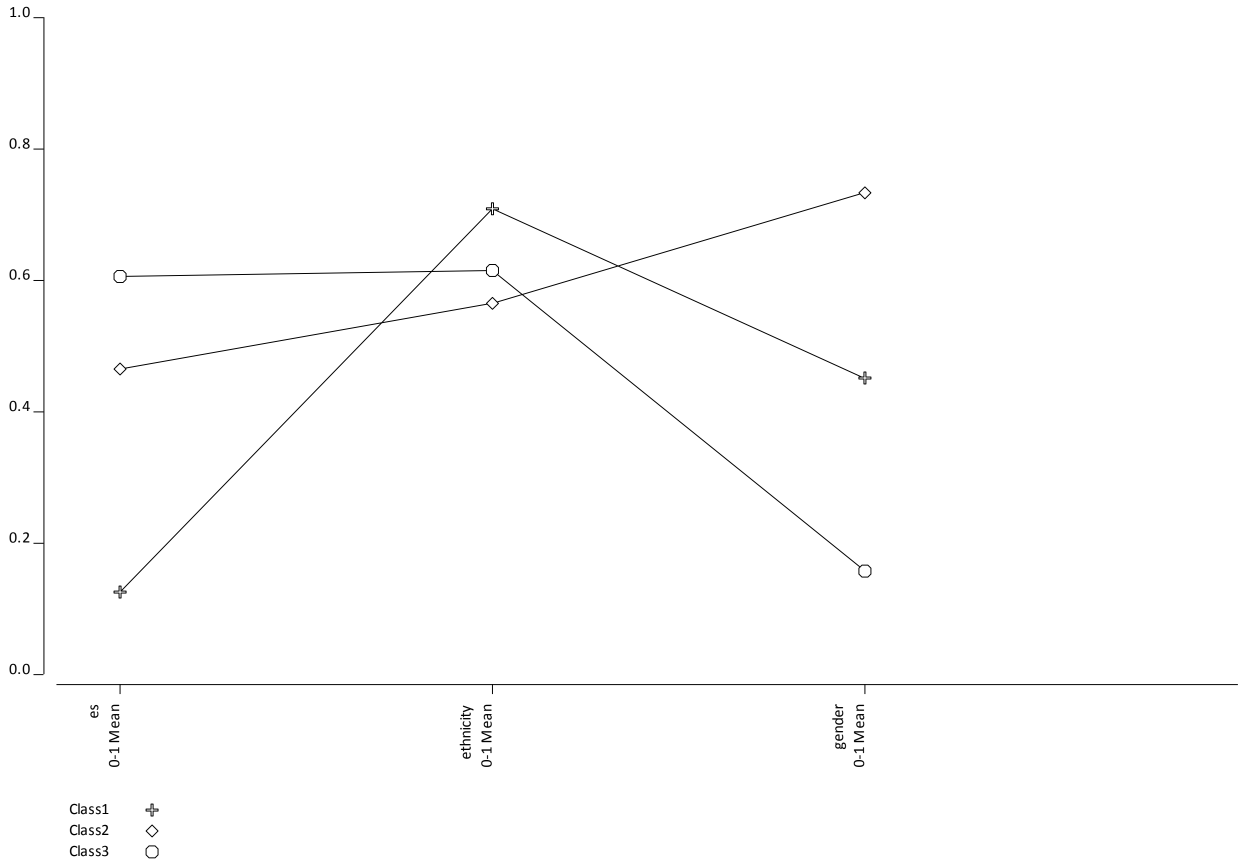


Fig C1 conditional profile for 3 class solution

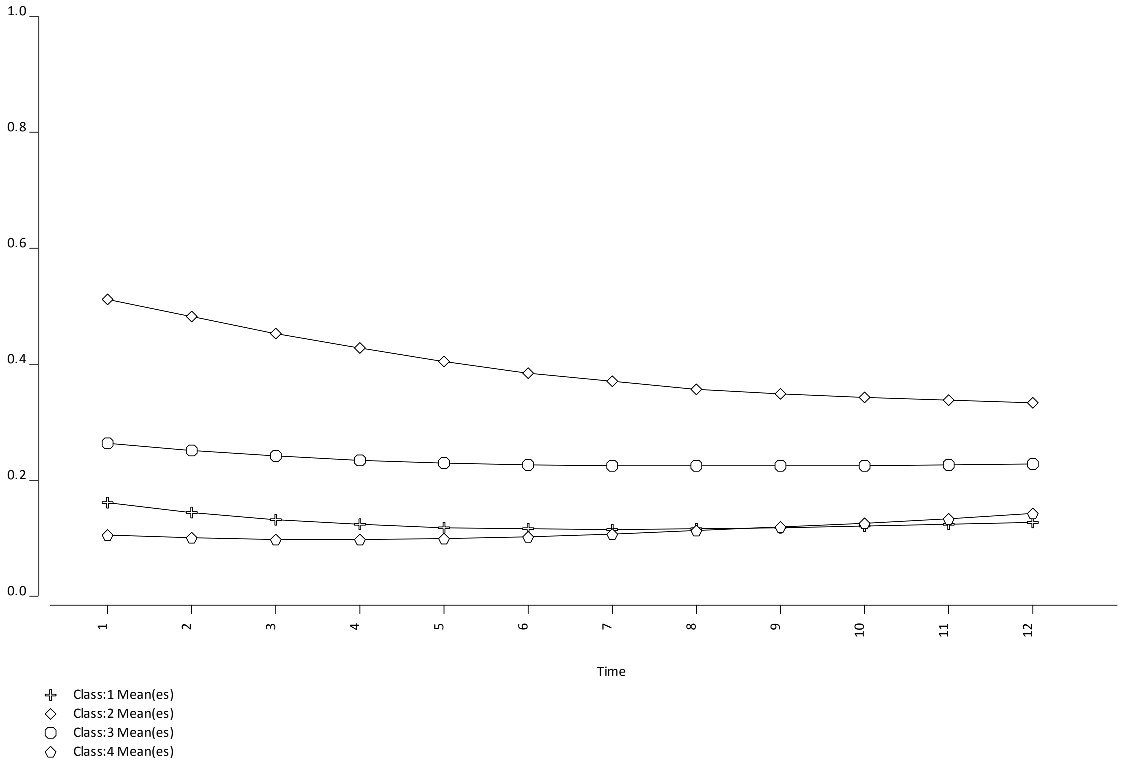
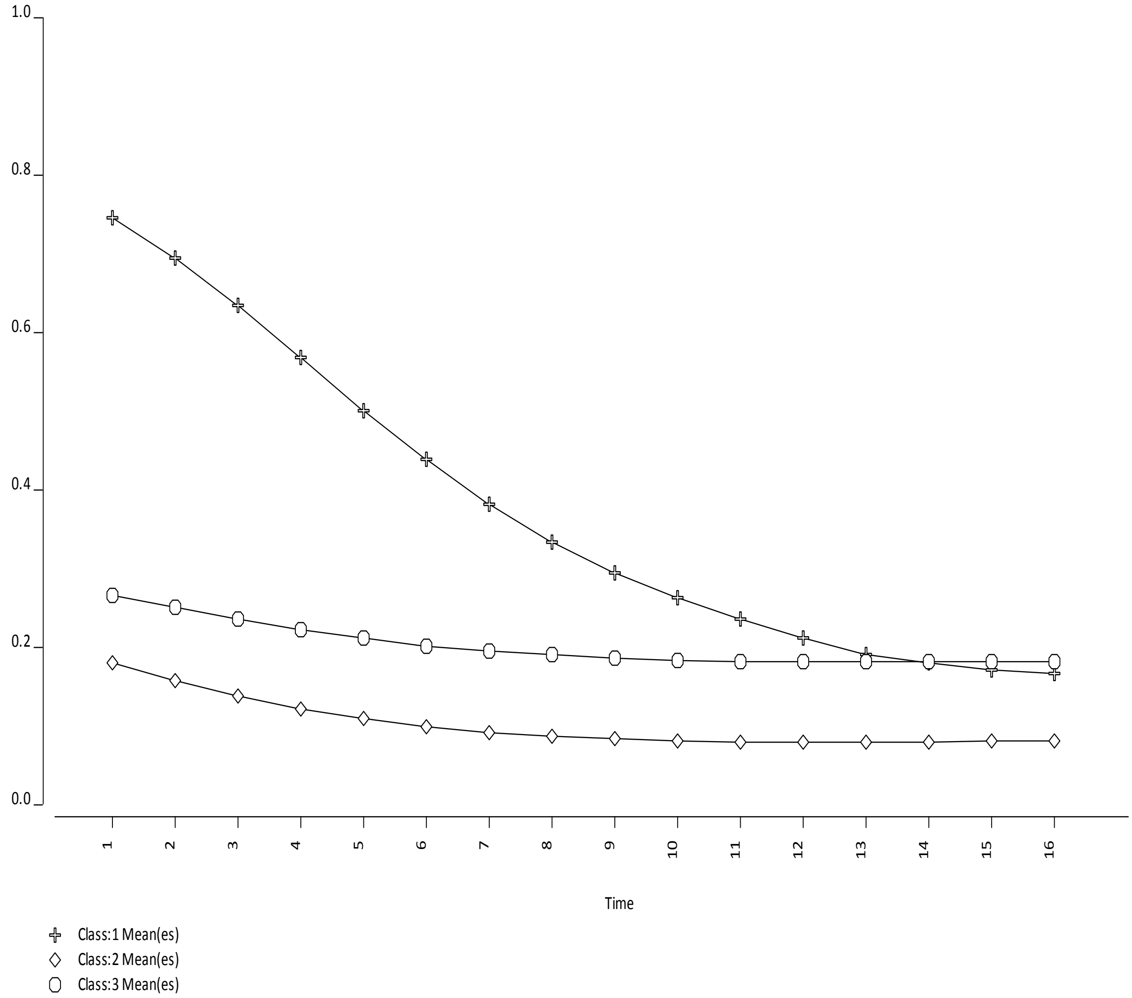


Fig C2 unconditional profile for competing 4 class solution



FigC3 unconditional profile for 3 class solution

Appendix D

Table D1: Conditional Markov Regression Parameters

The given table shows opposite and highly significant gender effects for both time zones of 79 cohort and 97 cohort (https://www.nlsinfo.org/content/cohorts/nlsy79).

term		coef	s.e.	z-value	p-value	Wald(0)	df	p-value
Class(1)	1	0.4805	0.136	3.5337	0.00041	12.487	1	0.00041
Class(2)	1	-0.4805	0.136	-3.5337	0.00041			
Class(1)	gender	-1.1779	0.0352	-33.4676	1.40E-245	1120.081	1	1.20E-245
Class(2)	gender	1.1779	0.0352	33.4676	1.70E-235			
Class(1)	cohort	0.0079	0.0014	5.533	3.10E-08	30.6137	1	3.10E-08
Class(2)	cohort	-0.0079	0.0014	-5.533	3.10E-08			
Class(1)	ethnicity	0.1008	0.0153	6.574	4.90E-11	43.2177	1	4.90E-11
Class(2)	ethnicity	-0.1008	0.0153	-6.574	4.90E-11			

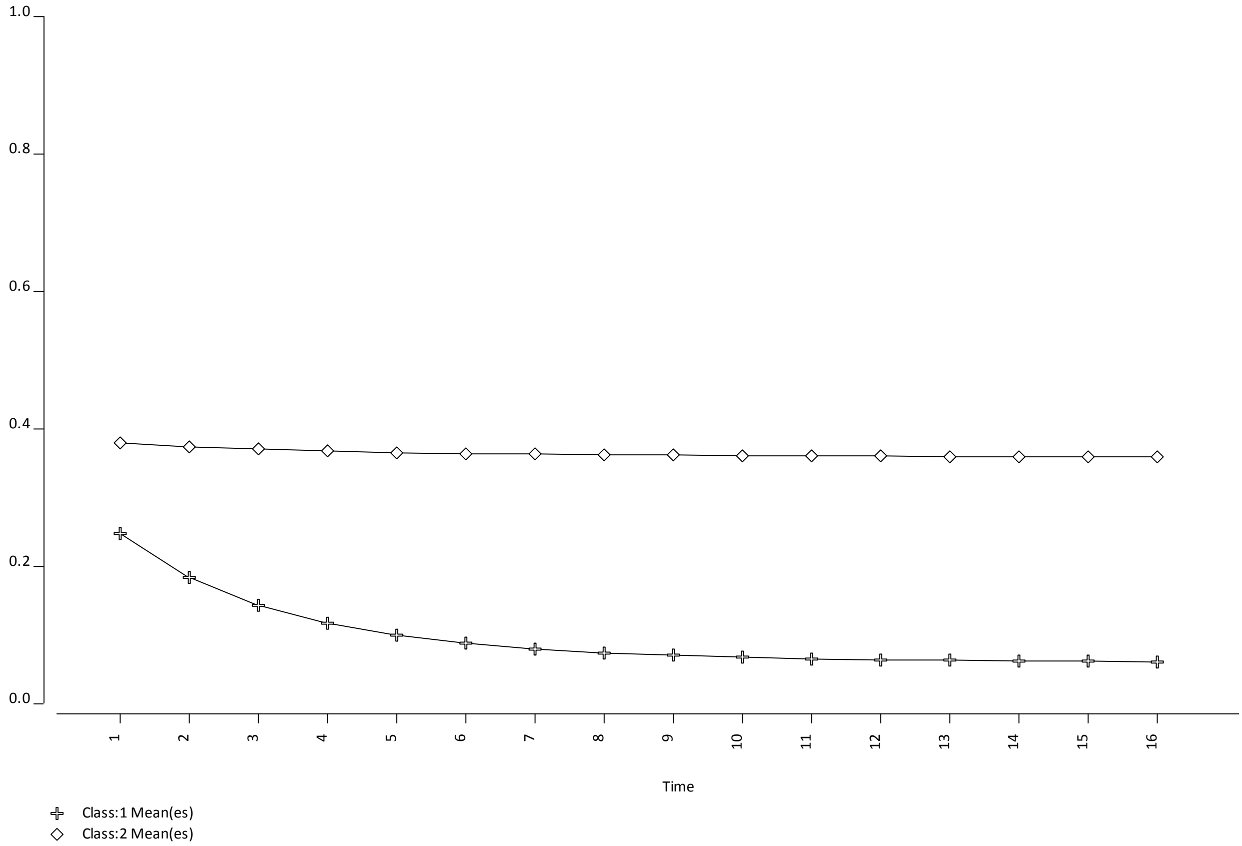


Fig D1. Markov mixtures

Software employed is latent gold 5.5 syntax version and R for basic exploratory data analysis and growth modelling.

References

" Institute for Social and Economic Research". Retrieved from https://www.understandingsociety.ac.uk/

"European Foundation ". Retrieved from https://www.eurofound.europa.eu/

"National Longitudinal Surveys (NLS)". Retrieved from https://www.nlsinfo.org/

"Understanding Society". (2021). Waves 1-13, 2008-2020., from University of Essex, Institute for Social and Economic Research.

Acconcia, A., Carannante, M., Misuraca, M., & Scepi, G. (2020). Measuring vulnerability to poverty with Latent Transition Analysis. *Social Indicators Research*

151, 1-31.

Agresti, A. (2003). *Categorical data analysis* (Vol. 482): John Wiley & Sons.

Agresti, A., Booth, J. G., Hobert, J. P., & Caffo, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology*

30(1), 27-80.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics, 55*(1), 117-128.

Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of public economics*

95(7-8), 476-487.

Andrews, R. L., & Currim, I. S. (2003a). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research, 40*(2), 235-243.

Andrews, R. L., & Currim, I. S. (2003b). Retention of latent segments in regression-based marketing models. *International Journal of Research in Marketing, 20*(4), 315-321.

Andruff, H., Carraro, N., Thompson, A., Gaudreau, P., & Louvet, B. (2009). Latent class growth modelling: a tutorial. *Tutorials in quantitative methods for psychology, 5*(1), 11-24.

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using M plus. *Structural equation modeling: A multidisciplinary Journal*

21(3), 329-341.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*

11, 685-725.

Atkinson, A. B., Marlier, E., & Nolan, B. J. J. o. C. M. S. (2004). Indicators and targets for social inclusion in the European Union. *JCMS:Journal of Common Market Studies*

42(1), 47-75.

Auerbach, K. J., & Collins, L. M. (2006a). A multidimensional developmental model of alcohol use during emerging adulthood. *Journal of studies on alcohol, 67*(6), 917-925.

Auerbach, K. J., & Collins, L. M. (2006b). A multidimensional developmental model of alcohol use during emerging adulthood. *Journal of studies on alcohol, 67*(6), 917-925.

Bacci, S., Pandolfi, S., Pennoni, F., & Classification. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis*

8(2), 125-145.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013a). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological methodology, 43*(1), 272-311.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013b). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological methodology*

43(1), 272-311.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*

803-821.

Bartelheimer, P., Büttner, R., & Schmidt, T. (2011). Dynamic capabilities–A capability approach to life courses and the case of young adults. *Closing the capabilities gap–Renegotiating social justice for the young*

147-164.

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods & Research*

27(4), 525-546.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2014). Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test, 23*(3), 433-465.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2019). *Latent Markov models for longitudinal data*: Chapman and Hall/CRC.

Bartolucci, F., Montanari, G. E., & Pandolfi, S. (2015). Three-step estimation of latent Markov models with covariates. *Computational Statistics & Data Analysis*

83, 287-301.

Basu, K. (1987). Achievements, capabilities and the concept of well-being. *Social Choice And Welfare, 4*(1), 69-76.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*

41(1), 164-171.

Bazillier, R., Boboc, C., & Calavrezo, O. (2014). Employment vulnerability in Europe: Is there a migration effect?

Ben-Ishai, L. (2014). *Access to Paid Leave: An Overlooked Aspect of Economic and Social Inequality*: Center for Law and Social Policy (CLASP).

Benach, J., Vanroelen, C., Vives, A., & De Witte, H. (2013). Quality of employment conditions and employment relations in Europe.

Bescond, D., Chataignier, A., & Mehran, F. (2003). Seven indicators to measure decent work: An international comparison. *Int'l Lab. Rev., 142*, 179.

Bierbrauer, M., Trück, S., & Weron, R. (2004). *Modeling electricity prices with regime switching models.* Paper presented at the International Conference on Computational Science.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence, 22*(7), 719-725.

Bocquier, P., Nordman, C. J., & Vescovo, A. (2010). Employment vulnerability and earnings in urban West Africa. *World Development*

38(9), 1297-1314.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political analysis*

12(1), 3-27.

Broughton, A., Green, M., Rickard, C., Swift, S., Eichhorst, W., Tobsch, V., & Tros, F. (2016). Precarious employment in Europe: Part 1, patterns, trends and policy strategy. *Retrieved February*

26, 2016.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*: Sage Publications, Inc.

Buck, N., & McFall, S. (2011). Understanding Society: design overview. *Longitudinal Life Course Studies*

3(1), 5-17.

Burchell, B., Sehnbruch, K., Piasna, A., & Agloni, N. (2014). The quality of employment and decent work: definitions, methodologies, and ongoing debates. *Cambridge journal of economics*

38(2), 459-477.

Catalina Rubianes, A., & Annoni, P. (2016). Tree-based approaches for understanding growth patterns in the European regions. *Region: the journal of ERSA, 3*(2), 23-45.

Clark, A. E. (2005). Your money or your life: Changing job quality in OECD countries. *British Journal of Industrial Relations, 43*(3), 377-400.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research, 28*(3), 375-389.

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718): John Wiley & Sons.

Cooke, G. B., Donaghey, J., & Zeytinoglu, I. (2013). The nuanced nature of work quality: Evidence from rural Newfoundland and Ireland. *Human Relations, 66*(4), 503-527.

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., & Vermunt, J. K. (2012). Exploring dynamics in mood regulation—mixture latent Markov modeling of ambulatory assessment data. *Psychosomatic medicine, 74*(4), 366-376.

Davoine, L., Erhel, C., & Guergoat-Larivière, M. (2008). A taxonomy of European labour markets using quality indicators.

Dayton, C. M. J. M., Counseling, E. i., & Development. (1991). Educational applications of latent class analysis. *24*(3), 131-141.

De Bustillo, R. M., Fernández-Macías, E., Antón, J.-I., & Esteve, F. (2011). *Measuring more than money: The social economics of job quality*: Edward Elgar Publishing.

Delgado, M., Gómez-Skarmeta, A. F., & Martín, F. (1997). A fuzzy clustering-based rapid prototyping for fuzzy rule-based modeling. *IEEE Transactions on Fuzzy Systems*

5(2), 223-233.

Di Mari, R., & Bakk, Z. (2018). Mostly harmless direct effects: A comparison of different latent Markov modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*

25(3), 467-483.

Di Mari, R., Oberski, D. L., & Vermunt, J. K. (2016). Bias-adjusted three-step latent Markov modeling with covariates. *Structural Equation Modeling: A Multidisciplinary Journal*

23(5), 649-660.

Dias, J. G., Vermunt, J. K., & Ramos, S. (2015). Clustering financial time series: New insights from an extended hidden Markov model. *European Journal of Operational Research, 243*(3), 852-864.

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. (2002). *Analysis of longitudinal data*: Oxford university press.
Duran, B. S., & Odell, P. L. (2013). *Cluster analysis: a survey* (Vol. 100): Springer Science & Business Media.
Erhel, C., & Guergoat-Larivière, M. (2010a). Job quality and labour market performance. *Centre for European Policy Studies Working Document*

(3).
Erhel, C., & Guergoat-Larivière, M. (2010b). Job quality and labour market performance. *Centre for European Policy Studies Working Document*

(3).
Everitt, B. S. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*

6(5), 305-309.
Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011a). Cluster analysis In (5th ed

ed.): John Wiley.
Everitt, B. S., Landau, S., Leese, M., & Stahl, D. C. a. (2011b). Miscellaneous clustering methods. 215-255.
Farrelly, C. M., Schwartz, S. J., Amodeo, A. L., Feaster, D. J., Steinley, D. L., Meca, A., & Picariello, S. (2017). The analysis of bridging constructs with hierarchical clustering methods: An application to identity. *Journal of Research in Personality, 70*, 93-106.
Feldman, B. J., Masyn, K. E., & Conger, R. D. (2009). New approaches to studying problem behaviors: a comparison of methods for modeling longitudinal, categorical adolescent drinking data. *Developmental psychology, 45*(3), 652.
Findlay, P., Kalleberg, A. L., & Warhurst, C. (2013). The challenge of job quality. *Human relations, 66*(4), 441-451.
Fonseca, J. R. (2013). Clustering in the field of social sciences: that is your choice. *International Journal of Social Research Methodology*

16(5), 403-428.
Fraley, C., & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of classification*

24(2), 155-181.
García-Pérez, C., Prieto-Alaiz, M., & Simón, H. J. S. I. R. (2017). A new multidimensional approach to measuring precarious employment. *134*(2), 437-454.
Garzón-Duque, M. O., Cardona-Arango, M. D., Rodríguez-Ospina, F. L., & Segura-Cardona, A. M. (2017). Informality and employment vulnerability: application in sellers with subsistence work. *Revista de saude publica*

51, 89.
Gindling, T. H., & Newhouse, D. (2012). *Self-employment in the developing world*: The World Bank.
Green, F., & Mostafa. (2012). Quality of work and Employment. *European Foundation for the Improvement of Living*

Working Conditions, Dublin http://cje. oxfordjournals. org.
Greenan, N., & Seghir, M. (2017). Measuring vulnerability to adverse working conditions: evidence from European countries.
Gudicha, D. W., & Vermunt, J. K. (2013). Mixture model clustering with covariates using adjusted three-step approaches. In *Algorithms from and for nature and life* (pp. 87-94): Springer.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*: Cambridge University Press.
Hand, D., & Crowder, M. (2017). *Practical longitudinal data analysis*: Routledge.
Hartzel, J., Agresti, A., & Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling, 1*(2), 81-102.
Hastie, T. J., & Pregibon, D. (2017). Generalized linear models. In *Statistical models in S* (pp. 195-247): Routledge.
Hedeker, D., & Mermelstein, R. J. (1998). A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*

33(4), 427-455.
Hedeker, D., & Mermelstein, R. J. J. M. B. R. (1998). A multilevel thresholds of change model for analysis of stages of change data. *33*(4), 427-455.
Helbling, L., & Kanji, S. (2018). Job insecurity: differential effects of subjective and objective measures on life satisfaction trajectories of workers aged 27–30 in Germany. *Social Indicators Research*

137(3), 1145-1162.
Hennig, C., & Liao, T. F. J. J. o. t. R. S. S. S. C. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *62*(3), 309-369.
Hofmans, J., De Gieter, S., & Pepermans, R. (2013). Individual differences in the relationship between satisfaction with job rewards and job satisfaction. *Journal of vocational behavior, 82*(1), 1-9.
Horowitz, J. (2016). *Dimensions of job quality, mechanisms, and subjective well-being in the United States.* Paper presented at the Sociological Forum.
Horowitz, J. L. (1997). Bootstrap methods in econometrics: theory and numerical performance. *Econometric Society Monographs, 28*, 188-222.
Huneeus, F., Landerretche, O., Puentes, E., & Selman, J. (2015). A multidimensional employment quality index for Brazil, 2002–11. *International Labour Review, 154*(2), 195-226.
Ilsøe, A., Larsen, T. P., & Felbo-Kolding, J. (2017). Living hours under pressure: flexibility loopholes in the Danish IR-model. *Employee Relations*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.
Janssen, P., Walther, C., & Lüdeke, M. K. (2012). Cluster analysis to understand socio-ecological systems: a guideline.
Jencks, C., Perman, L., & Rainwater, L. (1988). What is a good job? A new measure of labor-market success. *American journal of sociology*

93(6), 1322-1357.
Jesnes, K. (2018). Approaches to atypical and precarious work. *Fafo: Oslo, Norway*.
Jung, T., & Wickrama, K. A. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and personality psychology compass, 2*(1), 302-317.
Jung, T., & Wickrama, K. A. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social personality psychology compass*

2(1), 302-317.
Kalleberg, A. L. (2011). *Good jobs, bad jobs: The rise of polarized and precarious employment systems in the United States, 1970s-2000s*: Russell Sage Foundation.
Kalleberg, A. L., Reskin, B. F., & Hudson, K. (2000). Bad jobs in America: Standard and nonstandard employment relations and job quality in the United States. *American sociological review*, 256-278.

Kamakura, W. A., Wedel, M., & Agrawal, J. (1994). Concomitant variable latent class models for conjoint analysis. *International Journal of Research in Marketing, 11*(5), 451-464.
Kaplan, D. (2008). An overview of Markov chain methods for the study of stage-sequential developmental processes. *Developmental psychology*

44(2), 457.
Khalili, A., & Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the american Statistical association, 102*(479), 1025-1038.
Kim, M., Vermunt, J., Bakk, Z., Jaki, T., & Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*

23(4), 601-614.
Körner, T., Puch, K., & Wingerter, C. (2009). Quality of employment. *Wiesbaden: Federal Statistical Office of Germany*.
Langeheine, R., Pannekoek, J., & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research, 24*(4), 492-516.
Langeheine, R., Stern, E., & Van de Pol, F. (1994). State mastery learning: dynamic models for longitudinal data. *Applied Psychological Measurement*

18(3), 277-291.
Lanza, S. T., Bray, B. C., & Collins, L. M. (2013). An introduction to latent class and latent transition analysis. *Handbook of psychology, 2*, 691-716.
Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural equation modeling: a multidisciplinary journal*

14(4), 671-694.
Lanza, S. T., Flaherty, B. P., & Collins, L. M. (2003). Latent class and latent transition analysis.
Lawrence, C. J., & Krzanowski, W. J. (1996). Mixture separation for mixed-mode data. *Statistics & Computing*

6(1), 85-92.
Leschke, J., & Watt, A. (2014). Challenges in constructing a multi-dimensional European job quality index. *Social indicators research*

118(1), 1-31.
Leβmann, O. (2012). Applying the Capability Approach Empirically: Overview with Special Attention to Labour. *Management Review*

Living, E. F. f. t. I. o. (2012). Young people and NEETs in Europe: First findings. *Working Conditions %J Dublin. Available at: http://www. eurofound. europa. eu/pubdocs//72/en/2/EFEN. pdf*

Lower-Basch, E. (2007). *Opportunity at work: improving job quality*: CLASP.
Lugo, A. (2007). A proposal for internationally comparable indicators of employment. *Oxford Journal of Development Studies, 35*(4), 361-378.
Lystig, T. C., & Hughes, J. P. (2002). Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics*

11(3), 678-689.
MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual review of psychology, 51*(1), 201-226.
MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series* (Vol. 110): CRC Press.

Magidson, J., & Vermunt, J. K. (2002). A nontechnical introduction to latent class models. *Statistical Innovations white paper*

1, 15.

Magidson, J., Vermunt, J. K., & Tran, B. (2009a). Using a mixture latent Markov model to analyze longitudinal US employment data involving measurement error. *New trends in psychometrics*, 235-242.

Magidson, J., Vermunt, J. K., & Tran, B. (2009b). Using a mixture latent Markov model to analyze longitudinal US employment data involving measurement error. *New trends in psychometrics*

235-242.

Magidson, J., & Vermunt, J. K. J. T. S. h. o. q. m. f. t. s. s. (2004). Latent class models. 175-198.

Malik, M. E., Danish, R. Q., & Munir, Y. J. A. J. o. e. (2012). The impact of pay and promotion on job satisfaction: Evidence from higher education institutes of Pakistan. *2*(4), 6-9.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*

6, 355-378.

McParland, D., Gormley, I. C., McCormick, T. H., Clark, S. J., Kabudula, C., & Collinson, M. A. (2014). Clustering South African households based on their asset status using latent variable models. *The annals of applied statistics, 8*(2), 747.

Mindrila, D. (2020a). Latent Class Analysis.

Mindrila, D. (2020b). Latent Class Analysis. *International Journal for Cross-Disciplinary Subjects in Education (IJCDSE),Volume 11, Issue 2, 2020*.

Moore, W., Pedlow, S., Krishnamurty, P., Wolter, K., & Chicago, I. (2000). National longitudinal survey of youth 1997 (NLSY97). *Technical sampling report: National Opinion Research Center*

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British journal of mathematical*

and statistical psychology

49(2), 313-334.

Muthèn, B. (2004). Latent variable analysis: Growth mixture modeling and related.

Muthén, B., & Asparouhov, T. (2002a). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes*

4(5), 1-22.

Muthén, B., & Asparouhov, T. (2002b). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes, 4*(5), 1-22.

Muthén, B., & Muthén, B. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *24*(6), 882-891.

Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and experimental research*

24(6), 882-891.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81-117.

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2017). Power and type I error of local fit statistics in multilevel latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*

24(2), 216-229.
Nagin, D. (2009). *Group-based modeling of development*: Harvard University Press.
Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods, 4*(2), 139.
Nagin, D. S., Jones, B. L., Passos, V. L., & Tremblay, R. E. (2018). Group-based multi-trajectory modeling. *Group-based multi-trajectory modeling, 27*(7), 2015-2023.
Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology, 31*(3), 327-362.
Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: a group-based method. *Psychological methods, 6*(1), 18.
Nagin, D. S., & Tremblay, R. E. (2005). Developmental trajectory groups: Fact or a useful statistical fiction? *Criminology, 43*(4), 873-904.
Nations., U. (2015). *Handbook on Measuring Quality of Employment: A Statistical Framework*: UN.
Nergaard, K., Alsos, K., Bråten, M., & Jensen, R. S. (2015). Temporary staff in Norwegian working life. *Oslo: Fafo-rapport, 10*.
Nesterenko, V. (2011). QUALITY OF EMPLOYMENT MEASUREMENT IN UKRAINE. *Contemporary Problems of Transformation Process*, 150.
Nylund-Gibson, K., Grimm, R., Quirk, M., & Furlong, M. (2014). A latent transition mixture model using the three-step specification. *Structural Equation Modeling: A Multidisciplinary Journal*

21(3), 439-454.
Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019a). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural equation modeling: A multidisciplinary Journal*

26(6), 967-985.
Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019b). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural equation modeling: A multidisciplinary Journal, 26*(6), 967-985.
Nylund, K., Bellmore, A., Nishina, A., & Graham, S. (2007). Subtypes, severity, and structural stability of peer victimization: What does latent class analysis say? *Child development, 78*(6), 1706-1722.
Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal, 14*(4), 535-569.
Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal*

14(4), 535-569.
Oberski, D. L., van Kollenburg, G. H., & Vermunt, J. K. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification, 7*(3), 267-279.
Olsen, K. M., Kalleberg, A. L., & Nesheim, T. (2010). Perceived job quality in the United States, Great Britain, Norway and West Germany, 1989-2005. *European Journal of Industrial Relations*

16(3), 221-240.
Olsthoorn, M. J. S. I. R. (2014). Measuring precarious employment: A proposal for two indicators of precarious employment based on set-theory and tested with Dutch labor market-data. *119*(1), 421-441.

Paas, L. J., Vermunt, J. K., & Bijmolt, T. H. (2007a). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society*

170(4), 955-974.
Paas, L. J., Vermunt, J. K., & Bijmolt, T. H. (2007b). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society: Series A*

170(4), 955-974.
Parent-Thirion, A., Biletta, I., Cabrita, J., Llave Vargas, O., Vermeylen, G., Wilczynska, A., & Wilkens, M. 6th European Working Conditions Survey: Overview Report; 2017 update; Publications Office of the European Union: Luxembourg, 2017.
Pugliesi, K. (1999). The consequences of emotional labor: Effects on work stress, job satisfaction, and well-being. *Motivation and emotion*

23(2), 125-154.
Qu, Y., Tan, M., & Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 797-810.
Reinecke, J., & Seddig, D. (2011). Growth mixture models in longitudinal research. *Advances in Statistical Analysis, 95*(4), 415-434.
Reinecke, J., & Seddig, D. J. A. A. i. S. A. (2011). Growth mixture models in longitudinal research. *95*(4), 415-434.
Reiser, M., & Lin, Y. (1999). A Goodness-of-Fit Test for the Latent Class Model When Expected Frequencies are Small. *Sociological methodology, 29*(1), 81-111.
Rothwell, J., & Crabtree, S. (2019). Not just a job: New evidence on the quality of work in the United States. *Lumina Foundation, Bill*

Melinda Gates Foundation, Omidyar Network

Gallup

Ryoo, J. H., Wang, C., Swearer, S. M., Hull, M., & Shi, D. (2018). Longitudinal model building using latent transition analysis: an example using school bullying data. *Frontiers in psychology, 9*, 675.
Sánchez, N. S., & Puente, A. C. F. (2015a). Educational and skill mismatches: differential effects on job satisfaction. A study applied to the Spanish job market. *Estudios de Economía, 41*(2), pp. 261-281.
Sánchez, N. S., & Puente, A. C. F. (2015b). Educational and skill mismatches: differential effects on job satisfaction. A study applied to the Spanish job market. *Estudios de Economía, 41*(2), pp. 261-281.
Sehnbruch, K. (2004). From the quantity to the quality of employment: An application of the Capability Approach to the Chilean labour market.
Sehnbruch, K. (2006). *The Chilean labor market: A key to understanding Latin American labor markets*: Springer.
Sehnbruch, K., González, P., Apablaza, M., Méndez, R., & Arriagada, V. (2020). The Quality of Employment (QoE) in nine Latin American countries: A multidimensional perspective. *World Development, 127*, 104738.
Shockey, J. W. (1988). Latent class analysis: an introduction to discrete data models with unobserved variables. *Common Problems/Proper Solutions: Avoiding Error in Quantitative Research. Beverly Hills: Sage*, 288-315.
Simonoff, J. S. (2003). *Analyzing categorical data* (Vol. 496): Springer.

Sinha, P., Calfee, C. S., & Delucchi, K. L. (2021). Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls. *Critical care medicine, 49*(1), e63.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*: Chapman and Hall/CRC.
Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research 5*(4), 31-43.
Stewart, F. (1985). A Basic Needs Approach to Development. In *Planning to Meet Basic Needs* (pp. 1-13). London: Palgrave Macmillan UK.
Strotmann, H., & Volkert, J. (2008). *Lack of Instrumental Freedoms: Social Exclusion from and Unfavourable Inclusion into the Labour Market-An Empirical Analysis for Germany.* Paper presented at the HDCA-conference in Delhi.
Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. *Advances in latent variable mixture models, 2007*(1), 317.
Tuma, M., & Decker, R. (2013). Finite mixture models in market segmentation: a review and suggestions for best practices. *Journal of Business Research Methods*

11(1), pp215-pp215.
United States Department of Labor, B. o. L. S. B. National Longitudinal Surveys. .
Van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological methodology*, 213-247.
Van de Pol, F., & Langeheine, R. J. S. m. (1990). Mixed Markov latent class models. 213-247.
Van De Schoot, R., Sijbrandij, M., Winter, S. D., Depaoli, S., & Vermunt, J. K. (2017). The GRoLTS-checklist: guidelines for reporting on latent trajectory studies. *Structural equation modeling: A multidisciplinary Journal, 24*(3), 451-467.
Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3-28.
Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: a review. *Statistical methods in medical research, 23*(1), 42-59.
Vermeylen, G., Wilkens, M., Biletta, I., & Fromm, A. (2017). *Exploring self-employment in the European Union*: Publications Office of the European Union.
Vermunt, J. D., & Vermetten, Y. J. (2004). Patterns in student learning: Relationships between learning strategies, conceptions of learning, and learning orientations. *Educational psychology review, 16*(4), 359-384.
Vermunt, J. K. (2005). Mixed-effects logistic regression models for indirectly observed discrete outcome variables. *Multivariate Behavioral Research, 40*(3), 281-301.
Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis, 18*(4), 450-469.
Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *J Political analysis, 18*(4), 450-469.
Vermunt, J. K. (2017). *Growth models for categorical response variables: standard, latent-class, and hybrid approaches*: Routledge.
Vermunt, J. K., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational & Behavioral Statistics*

24(2), 179-207.
Vermunt, J. K., & Magidson, J. (2013). Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax. *Statistical Innovations Inc*.
Vermunt, J. K., & Magidson, J. (2021). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural equation modeling: A multidisciplinary Journal, 28*(3), 356-364.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. *Handbook of longitudinal research: Design, measurement and analysis*

373-385.
Vermunt, J. K., & Van Dijk, L. (2001). A nonparametric random-coefficients approach: The latent class regression model. *Multilevel Modelling Newsletter, 13*(2), 6-13.
Vermunt, J. K., & Van Dijk, L. (2001). A nonparametric random-coefficients approach: The latent class regression model. *Multilevel Modelling Newsletter, 13*(2), 6-13.
Von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*(6), 389-406.
Wanger, S. (2017). *What makes employees satisfied with their working time? The role of working hours, time-sovereignty and working conditions for working time and job satisfaction*. Retrieved from
Wedel, M., & DeSarbo, W. S. (1994). A review of recent developments in latent class regression models. *Advanced methods of marketing research*, 352-388.
Wedel, M., & Kamakura, W. A. (2000). Mixture regression models. In *Market segmentation* (pp. 101-124): Springer.
Yamaguchi, K. (2000). Multinomial logit latent-class regression models: An analysis of the predictors of gender-role attitudes among Japanese women. *American Journal of Sociology, 105*(6), 1702-1740.
Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika, 62*(3), 297-330.