The Potential of Cohort Analysis for Vintage Analysis

An Exploration

Merijn Bosman- s1023039
University of Twente, Enschede, The Netherlands
1/23/2012

UNIVERSITY OF TWENTE.

Supervisor: Dr B. Roorda
Co-reader: Ir. H. Kroon

Abstract

This thesis explores whether and how cohort analysis can improve the vintage analysis techniques that are used for the analysis of loans by structured credit market participants.

First, a literature study is conducted to see how vintage analysis is currently applied by credit rating agencies and structured credit investors – two structured credit market participants. The goal of credit rating agencies is to predict the cash flows and risk so that a structured credit tranche can be rated. Structured credit investors use the same information, but their goal is to value the tranche. For credit rating agencies vintage analysis plays a role in the analysis of the historical default data that is used to determine default rates for a particular structured credit transaction. Structured credit investors use vintage analysis in the translation of the results of the collateral analysis into an asset model were assumptions have to be made regarding the default rate, prepayment rate, and the loss severity. It is shown that the current application of vintage analysis is partly based on expert judgment.

Second, a literature study on what cohort analysis entails, shows that it is an analysis technique used in various areas of science (e.g., demography, epidemiology, sociology, and biostatistics) in which statistical attempts are made to partition (variance in) the outcome on an independent variable into the unique components attributable to age, period, and cohort effects.

When vintage analysis is viewed from a cohort analysis perspective it can aid the vintage analysis process by unraveling maturation, extrinsic, and origination effects that have an influence on the structured credit vintage performance trajectories. The quantification of these effects can help structured credit market participants to better understand the historic performance of structured credit and allows them to forecast future trends in performance rates.

However, it is shown that the model identification problem in cohort analysis, which is the result of the perfect linear relationship between age, period, and cohort – or age, period, and vintage in case of structured credit – does not allow for a direct estimation of the three effects by generalized linear cohort models without assigning additional identifying constraints to the model. The model identification problem affects all generalized linear cohort analysis models. A classification of a plethora of cohort analysis models identified in the academic literature that deal with the model identification problem is presented.

Next, an application of cohort analysis to mortgage data shows how to use the tools that are available to conduct cohort analysis of structured credit data and what the differences are between cohort analysis in epidemiology and sociology on the one hand, and cohort analysis conducted on mortgage data on the other. It is shown that it is not difficult to come up with estimates for the maturation, extrinsic, and origination effects. However, the challenge is to produce sensible estimates of the effects.

Finally, the thesis concludes that a hurdle that needs to be taken before the potential of cohort analysis can be used, is fundamental research into the specific causal mechanisms that underlie the performance of structured credit that can be measured and analyzed. Future research should focus on the substantive importance of origination effects. Once this is defined, research can compare the input for the structured credit model that is generated by cohort analysis based vintage analysis methods to the expert judgment based methods that are currently applied by structured credit market participants.

Table of Contents

List of Notations

Where appropriate boldface type is used in this thesis to emphasize that reference is made to a vector or a matrix.

^{T}	Transpose of a matrix
$^{-1}$	Inverse of a matrix
$\boldsymbol{A}^{*}$	Conjugate transpose of matrix $\boldsymbol{A}$
a	Number of age groups
α_i	Fixed effect of the i^{th} age category
$\tilde{\alpha}_i$	Curvature parameter of the i^{th} age category
α_C	Curvature trend (deviations from linearity) of age effect
α_L	Linear trend of age effect
$\boldsymbol{A}_C$	Vector of curvature age variables of the design matrix
$\boldsymbol{A}_L$	Vector of linear age variables of the design matrix
$\boldsymbol{b}$	Vector of coefficients of the design matrix
$\widehat{\boldsymbol{b}}$	Vector of fitted coefficients of the design matrix
$\widehat{\boldsymbol{b}}_{\boldsymbol{c}}$	Vector of fitted coefficients of the design matrix given the constraint c
$\boldsymbol{B}_{\boldsymbol{0}}$	Null subspace of the eigenvector
$\boldsymbol{B}$	Non-null subspace (the complement subspace to the null subspace $\boldsymbol{B}_{\boldsymbol{0}}$)
b_i	i^{th} partial regression (slope) coefficient
$\boldsymbol{c}$	Vector for the constraint(s)
$\boldsymbol{C}_C$	Vector of curvature cohort variables of the design matrix
$\boldsymbol{C}_L$	Vector of linear cohort variables of the design matrix
$Cov(\cdot)$	Covariance of a random variable
η_j	Fixed effect of the j^{th} period category
η_L	Linear trend of period effect
$E(\cdot)$	Expectation of a random variable
ε	Residual error term
$\boldsymbol{\varepsilon}$	Vector of random errors
ε_{ij}	Residual error term
ε_{ijk}	Residual error term after linear regression of cohort on Median Polish residual error
e_k	Residual error term after Median Polish process
$f(\hat{R}_{ij})$	Some function of the observed rate
γ_{a-i+j}	Fixed cohort effect associated with the $(a-i+j)^{\text{th}}$ cohort category
G_c	$G_c = (X^T X)_c^-$, the generalized inverse associated with a particular linear constraint in a generalized linear cohort model
γ_k	Cohort effect of the k^{th} cohort category
γ_L	Linear trend of cohort effect
I	Identity matrix
k	Number of cohorts
$\ln(\cdot)$	Natural logarithm
m	$2(a+p)-3$
μ	Overall mean (intercept)
μ_{ij}	Expected number of arrears cases for the i^{th} age group in the j^{th} period
μ_k	Mean cohort effect

N_{ij}	Amount of risk time of the i^{th} age group in the j^{th} period
O_{ij}	Observed number of cases of the i^{th} age group in the j^{th} period
p	Number of periods
$\boldsymbol{P}_C$	Vector of curvature period variables of the design matrix
$\boldsymbol{P}_L$	Vector of linear period variables of the design matrix
σ^2	Variance
$\hat{R}_{ij}$	Observed rate of the i^{th} age group in the j^{th} period
t	Scalar
$\boldsymbol{v}$	$\boldsymbol{b}$ vector that results in a nontrivial solution to the homogeneous equation $\boldsymbol{Xb} = 0$
$V(\cdot)$	Variance of a random variable
x_i	i^{th} row of the design matrix
$\boldsymbol{X}$	Design matrix of independent (explanatory) variables
Y	Dependent (response) variable
$\boldsymbol{Y}$	Vector of observations on the dependent (response) variable
Y_{ij}	Age–period specific value on the outcome variable for the i^{th} age group at the j^{th} time period

List of Abbreviations

A	Age
ABS	Asset-Backed Security
ALBA	Aggregator of Loans Backed by Assets
APC	Age, Period, Cohort
APCC	Age-Period-Cohort Characteristic
C	Cohort
CCJ	Country Court Judgment
CDO	Collateralized Debt Obligation
CDR	Constant Default Rate
CLO	Collateralized Loan Obligation
CMBS	Commercial Mortgage Backed Security
CPR	Constant Prepayment Rate
CRA	Credit Rating Agency
DBRS	Dominion Bond Rating Service
EL	Expected Loss
FDIC	Federal Deposit Insurance Corporation
FHLMC	Federal Home Loan Mortgage Corporation (Freddie Mac)
FNMA	Federal National Mortgage Association (Fannie Mae)
GLM	Generalized Linear Model
GmbH	Gesellschaft mit beschränkter Haftung (Limited Liability Company)
IE	Intrinsic Estimator
IVA	Individual Voluntary Agreement
LTV	Loan-to-value
MBS	Mortgage-Backed Security
Moody's	Moody's Investor Services
MPT	Median Polish Technique
NPV	Net Present Value
PD	Probability of Default
P	Period
OHL	Oakwood Homeloans Limited
OTC	Over-The-Counter
RR	Rate Ratio
REIT	Real Estate Investment Trust
RMBS	Residential Mortgage Backed Security
SCI	Structured Credit Investor
SPV	Special Purpose Vehicle
S&P	Standard & Poor's Rating Service

“Science, like all creative activity, is exploration, gambling, and adventure. It does not lend itself very well to neat blueprints, detailed road maps, and central planning. Perhaps that’s why it’s fun.”

Herbert Simon (1962, p. 85)

1 Introduction

1.1 Problem Statement

Structured credit is a collective name for financial products comprising tranches of portfolios of credit instruments or exposures (Alexander, Eatwell, Persaud, & Reoch, 2007). The credit crisis that started in 2007-2008 showed that many of the models and tools that investors and credit rating agencies used and still use to value and determine the risk of structured credit products were wrong or, at least, misunderstood (Benmelech & Dlugosz, 2009; Lang & Jagtiani, 2010). Academic research regarding the methods and models used in structured credit markets is largely absent. A lack of publicly available data is the main reason for this (Crouhy, Jarrow, & Turnbull, 2008). Academic research that is present mainly focuses on credit ratings, due to the extensive, publicly available, documentation of the rating methodologies (see for example Ashcraft, Goldsmith-Pinkham, Hull and Vickery (2011); Ashcraft, Goldsmith-Pinkham and Vickery (2010); Chambers, Kelly and Lu (2010); Hull and White (2010); and Pagano and Volpin (2010)). The results of these studies confirm that credit rating agencies underestimated the credit risk associated with structured credit products and failed to adjust their ratings quickly enough to deteriorating market conditions.

This thesis looks at a widely used technique that is applied in the structured credit markets: vintage analysis. A vintage can be defined as a group of loans that all originated within a specific time period (see Figure 1 for an example of yearly vintages).

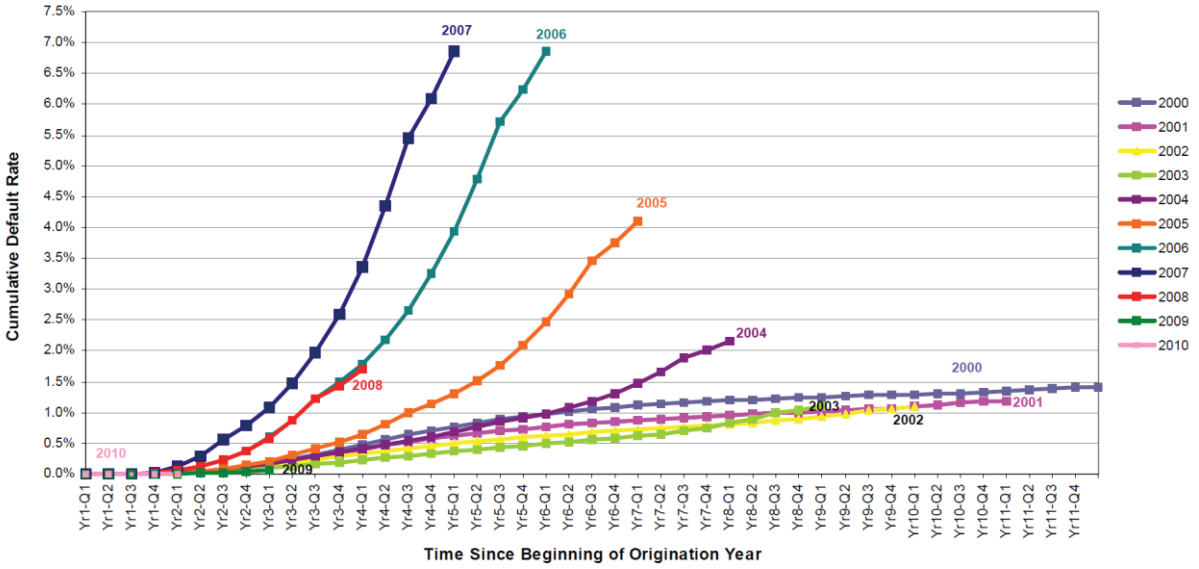


Figure 1: Fannie Mae Single-Family Cumulative Default Rates Grouped by Yearly Vintage (Fannie Mae, May 6, 2011).

Vintage analysis is used to recognize portfolio dynamics and behavior patterns based on pools of loans with common characteristics (Burns & Stanley, 2001; Raynes & Rutledge, 2003). Besides being used in the rating processes of structured products by credit rating agencies, vintage analysis is used by structured credit investors in their analysis and valuation of structured products. These market participants recognize that changing underwriting standards over time, the effect of extrinsic variables – such as changing macro-economic conditions – and the effect of aging, influence the performance of structured credit products. However, as is shown in chapters 2 and 3, the current application of vintage

analysis in practice is partly based on the expert judgment of the credit rating agencies' or structured credit investment managers' analysts.

A body of research that deals with unraveling age, period, and cohort effects is cohort analysis. Cohort analysis is a common analysis technique used in various areas of science (e.g., demography, epidemiology, sociology, and biostatistics) and has been widely discussed in academics (see for example: W.M. Mason & N.H. Wolfinger (2001), de Vaus (2001), and Yang (2007)). Yang and Land (2008) describe cohort methods as follows:

> "For the past 80 years or so, demographers, epidemiologists, and social scientists have attempted to analyze data using age (A) and time period (P) as explanatory variables to study phenomena that are time specific. An analytic focus on cohort (C) membership, as defined by the period and age at which an individual observation can first enter an age-by-period data array, is also important for substantive understanding. Accordingly, investigators have developed models for situations in which all three—age, period, and cohort (APC)—are potentially of importance to studying a substantive phenomenon. One common goal of APC analysis is to assess the effects of one of the three factors on some outcomes of interest net of the influences of the other two time-related dimensions" (pp. 297-298).

When it is possible to replace individual by loan and cohort by vintage in the above description, cohort analysis could aid the vintage analysis process by quantifying the maturation, extrinsic and origination effects on structured credit vintage performance. As a specific technique that aids in the vintage analysis of credit instruments or exposures, it is, as far as the author is aware, absent in the academic literature.

In sum, despite its widespread use in practice, academic research that focuses on loan vintage analysis is scarce. Currently, vintage analysis is a technique that combined with the experience of the analyst aids the structured credit analysis process. Cohort analysis' focus on quantifying age, period, and cohort effects makes it a suitable candidate to aid vintage analysis. This leads to the objective of this thesis which is discussed in the next section.

1.2 Research Objective

The objective of this thesis is to explore whether and how cohort analysis can improve the vintage analysis techniques that are used for the analysis of loans by structured credit market participants.

To study whether and how cohort analysis can improve vintage analysis a three step approach is followed. First, the role, and application of, vintage analysis within structured credit markets has to be defined. Second, the potential of cohort analysis, as technique to aid vintage analysis, has to be explored. Finally, if it is likely that cohort analysis can benefit vintage analysis, a proposal of how to utilize the benefits has to be developed.

1.3 Research Questions

The central research question of this thesis is:

How can cohort analysis improve the vintage analysis of loans?

In order to be able to answer this question, the following sub-questions are formulated:

1. *How is vintage analysis currently applied by structured credit market participants?*

2. *What is cohort analysis?*
3. *What developments in cohort analysis are relevant for vintage analysis?*

1.4 Research Design

This research is explorative in nature. Explorative research can be distinguished from descriptive and explanatory research. Explorative research is typically conducted "in the interest of getting to know or to increase understanding of a new or little researched setting, group or phenomenon" (Ruane, 2005, p. 12). The main academic value of explorative studies is generating research questions and hypotheses for additional investigation (Royse, 2010). De Groot (as cited in Reymen, 2001, p. 9) describes explorative research as:

> "Explorative research is empirical research that is appropriate when the researcher is, on a relatively broad domain with little useful theory, confronted with an amount of observations or variables for which relatively few relevant facts are known. The researcher is, however, aiming at a certain type of relations, with corresponding ideas and relatively vague expectations. This aim determines which facts will be taken into account, what will be measured, and which kinds of relations will be studied. The goal of this kind of research is mainly not the ordering of facts or the creation of an overview of 'the existing', but it aims at establishing relations that are considered to be relevant for a certain theoretical or practical goal. The researcher starts from certain expectations, from a more or less theoretical frame: He is trying to find relations in the material, but these are not defined by him in advance in the form of sharp hypotheses that can be tested; these hypotheses can thus also not yet be tested as such. Exact theory and/or hypothesis forming and testing must follow explorative research."

This description fits the nature of this research. At the moment the academic body of knowledge regarding secondary credit markets is scarce. As noted, the majority of the research that is existent focuses on the credit ratings in relation to structured credit. The result is a lack of well-defined models, empirical and quantitative support, and academic insight in the structured credit phenomenon. This broadly corresponds to the being on a "relatively broad domain with little useful theory" part of de Groot's description of explorative research. Second, an interdisciplinary study is conducted. The usefulness of a statistical method, cohort analysis, which is used in various non-financial social research disciplines for vintage analysis, a technique used in structured credit markets, is assessed. This corresponds to the "establishing relations that are considered to be relevant for a certain theoretical or practical goal" part of de Groot's description of explorative research. The explorative nature of the research is also expressed in the intended contribution of this thesis which is discussed next.

1.5 Research Contribution

This thesis aims for both an academic and practical contribution. First, there is an intended academic contribution. In section 1.1 it was noted that academic research regarding the methods and models used in structured credit markets is largely absent. Section 1.4 explained that the main academic value of exploratory studies is generating research questions and hypotheses for additional investigation. This thesis will not only map how vintage analysis is used in practice, but also show how different fields of science could come together in solving structured credit related problems. In doing so, this thesis:

1) aims to minimize the identified gap that is present in the current academic literature;

2) aims to show how insight from different academic disciplines might help to improve existing methods; and
3) aims to inspire other researchers to pursue research in directions where this thesis has left off.

Second, there is an intended practical contribution. By contributing to the understanding of loan vintages this thesis aims to make an improvement to vintage analysis. Society has criticized financial institutions for their role in the credit crisis and is looking for an answer to the question of how we can prevent this from happening again. By mapping the methods that are currently used, by providing suggestions for improvement, and by inspiring other researchers and market participants to develop and test improved models, ultimately the information asymmetry between market participants in structured credit markets will be reduced. By reducing the information asymmetry between market participants, transparency, and price discovery can be enhanced. In the end, this will result in an improved secondary loan market. This thesis is a first step in that direction.

1.6 Thesis Outline

This thesis is divided into six chapters. The outline is elaborated upon and visualized in Figure 2 below.

Chapter 2 will provide an answer to the first sub-question of this thesis: How is vintage analysis currently applied by structured credit market participants? An analysis is made of the role of vintage analysis in structured credit analysis. The main method used to gain insight into how vintage analysis is applied in practice is a literature study. The chapter can be divided into two parts. First, a short introduction into structured credit markets is provided. Second, an assessment of the use of vintage analysis by two main parties active in structured credit markets – credit rating agencies and structured credit investors – is made.

Next, chapter 3 provides an answer to the second sub-question of this thesis: What is cohort analysis? The chapter starts with the introduction of a definition of cohort analysis. Subsequently, two ways of defining what constitutes a cohort effect are introduced. After that, the model identification problem that is associated with cohort analysis is discussed. Then, a categorization of the plethora of models for conducting cohort analysis that are present in the academic literature is presented. The chapter concludes with a selection of cohort models that are discussed in more detail.

In chapter 4 the third sub-question – What developments in cohort analysis are relevant for vintage analysis? – is answered. An assessment of the relevance of cohort analysis for improving vintage analysis is made. In that sense chapter 4 provides a synthesis of chapters 2 and 3.

Chapter 5 illustrates how a cohort analysis can be applied to a dataset containing mortgage information. The goal is to get additional insight into the process of applying cohort analysis techniques to loan data.

Finally, chapter 6 will conclude this thesis. The chapter starts by answering the main research question of this thesis: How can cohort analysis improve the vintage analysis of loans? Subsequently, recommendations for further research are formulated. Finally, the limitations of the thesis are discussed.

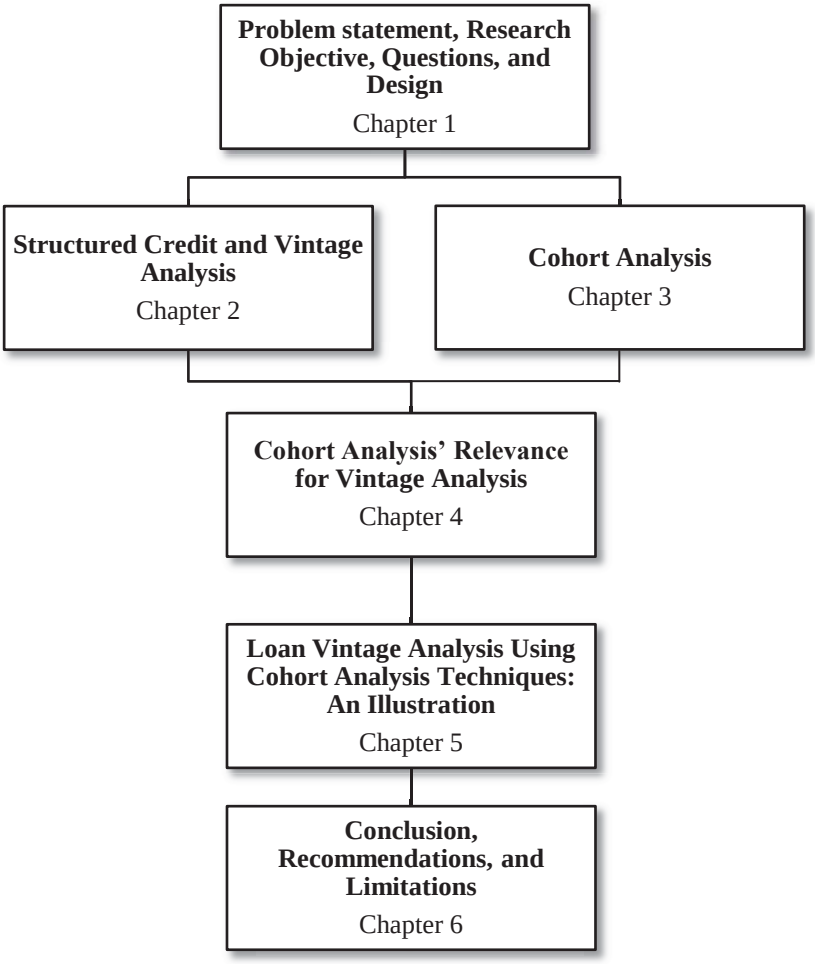


Figure 2: Thesis Outline.

2 Structured Credit and Vintage Analysis

In this chapter the first sub-question will be answered: How is vintage analysis currently applied by structured credit market participants?

As was mentioned in chapter 1, in credit markets, a *vintage* can be defined as a group of loans that all originated within a specific time period. The time period considered is problem-specific and is called the *vintage date*. The vintages can be viewed as "a set of overlapping time series with different starting times" (Breeden, 2007, p. 4761). *Vintage analysis* refers to the process of monitoring groups of loans and comparing performance across past groups.

Structured credit products can be analyzed at different levels. Vintage analysis is one of these levels. The relation between the vintage level of analysis and other levels of analysis is visualized in Figure 3. There are multiple structured credit types. Multiple issuers are active in each credit type and each issuer has one or more programs they use to issue their structured products. Each program consists of multiple vintages that have different origination dates and each vintage consists of varying numbers of accounts.

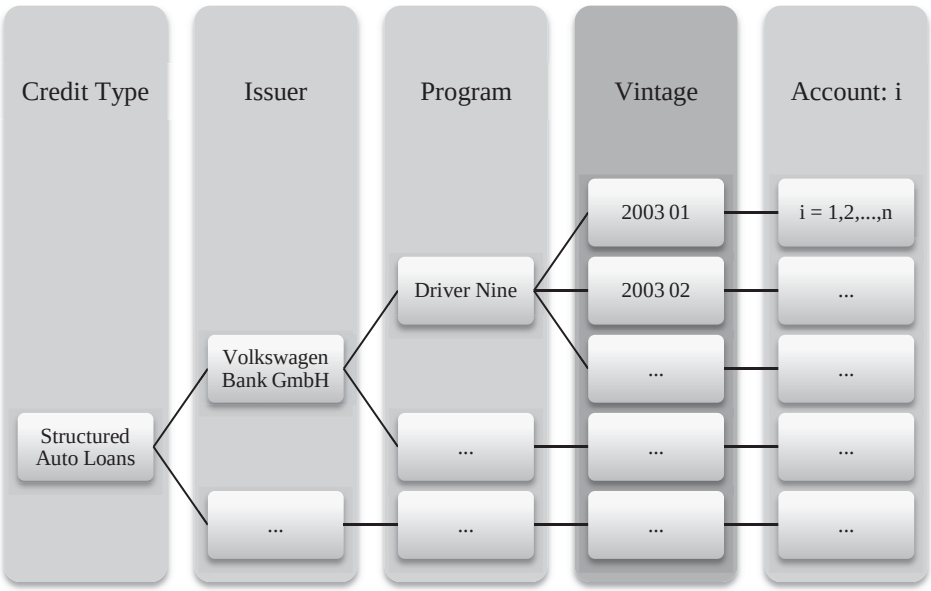


Figure 3: Vintage Level of Analysis Combined with Other Levels of Analysis.

This chapter looks at the use of vintage analysis by structured credit market participants. Therefore, this chapter starts with a short introduction on structured credit and securitization. Section 2.1 elaborates upon the definition of structured credit that was presented in chapter 1. The goal of this section is to get an idea of the financial products that are represented by the structured credit umbrella and what the securitization process entails. Next, section 2.2 elaborates upon the role of two of the main parties that are active in structured credit markets: the credit rating agencies and structured credit investors. Finally, building on the knowledge provided in section 2.1 and section 2.2, the use of vintage analysis by structured credit market participants is elaborated upon in section 2.3.

2.1 Structured Credit and Securitization

In chapter 1 structured credit was defined as a collective name for financial products comprising tranches of portfolios of credit instruments or exposures. The tranches of portfolios of credit instruments or exposures are formed by securitized assets or loans. Securitization can be broadly defined as the process by which assets or loans backed by assets with common features are packaged into (interest bearing) securities with marketable investment characteristics (Bhattacharya & Fabozzi, 1996). Investors generally only bear the risk arising from these receivables and are generally independent from the credit risk of the (former) owner of such assets (which is called the originator or seller). The assets that are backing the structured credit products are divers and can vary from mortgages to royalties from David Bowie's song catalogue (Megginson & Smart, 2008) to entire businesses (e.g., the 2006, $1.6 billion, whole business securitization of Dunkin' Brands (Bryer, Lebson, & Asbell, 2011)). Figure 4 lists some of the financial products that fall under the structured credit umbrella.

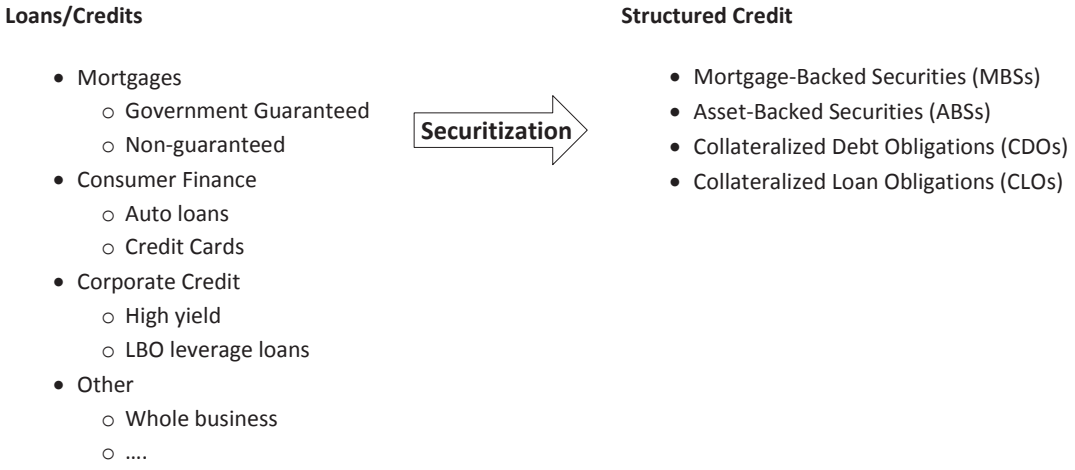


Figure 4: The Structured Credit Umbrella.

The general legal and cash flow dynamics of a typical securitization are shown in Figure 5. In its most basic form the securitization process involves two steps. First, investors pay cash up-front to purchase the securities and the right to receive the cash flow of the assets of the trust. The assets can be any (combination) of the assets listed in Figure 4. Next, as shown in step 1, the seller originates a pool of cash flow generating assets that it wants, for example, to remove from its balance sheet, pools them into what is called the reference portfolio, and sells them into a bankruptcy remote trust or special purpose vehicle (SPV). In step 2, the trust or SPV issues tradable interest-bearing notes or certificates to investors and pledges the cash flows from the receivables to the trust. A security interest is perfected by the trust in the receivables to the trust.

The financial products are structured with credit enhancement features to protect investors from credit losses. The security is structured into several slices: senior tranche(s), mezzanine tranche(s), and a junior tranche in order of seniority, which offer a sliding scale of coupon rates based on the level of credit protection afforded to the note or security. Generally, cash flow will be paid in order of priority, first to the senior tranche, then to the mezzanine tranche, and finally to the junior tranche. Usually, interest is paid first, then principal. The mezzanine tranche could not receive any interest payments unless the senior tranche is current. The same would apply to principal payments.

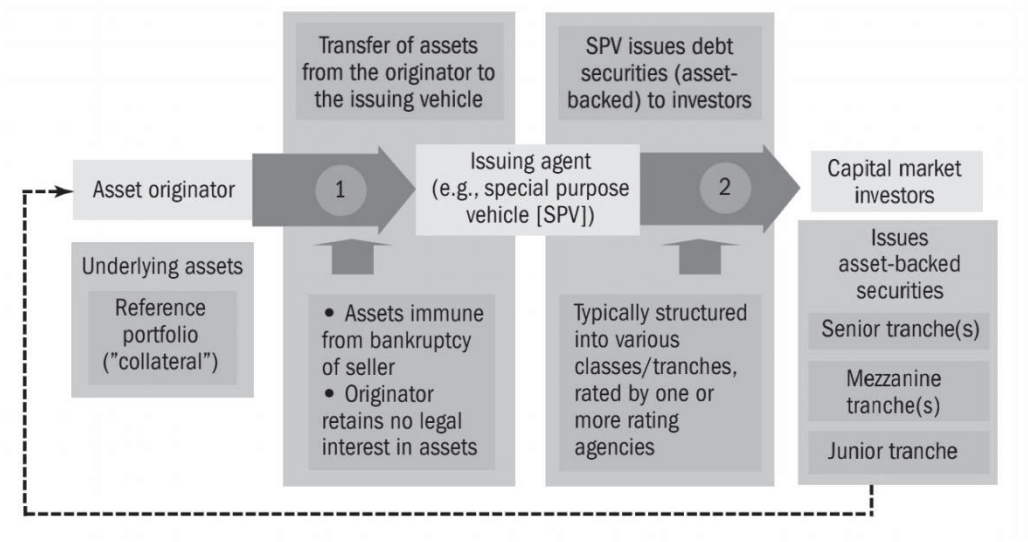


Figure 5: Simplified Standard Structured Credit Security Structure (Jobst, 2008, p. 48).

In the early days (70-80's) the securitization structures were simple: pass-through securities with mortgages as collateral in which each investor got a pro-rata share of the cash flows generated (Hayre, 2001). This implied that each investor received income from the investment as if they owned a small piece of each loan in the pool of mortgages. Because these securities came with a government guarantee against credit risk and had an expected return that was above those of the more traditional fixed income classes, they were an attractive investment to a number of different institutional investors, especially when they were required to hold high-quality assets[1]. However, the structure also came with several drawbacks. For example, due to the call option inherent in the mortgage that allows the borrower to prepay his mortgage and the put option that represents the default option that a borrower has, investors could not specify ahead of time when they would be paid[2]. The resulting uncertain return and investment horizon reduced the attractiveness of the asset for many investors. This is a drawback, because different investors have varying risk profiles. Some choose aggressive risk-return investment strategies, while others follow more risk-averse strategies.

When the private sector adopted the securitization model, it tailored securities to investor demands. First, tranching was introduced. The pass-through structure was replaced by a system in which the income created by loans in the pool was divided into different income streams suited to the investment horizon and risk preferences of investors. Second, to replace the government guarantee, privately-sponsored securitization relied on three approaches (Making Securitization Work for Financial Stability and Economic Growth, 2009). First, sponsors of private securitizations developed statistical models to estimate how much excess servicing, over-collateralization, subordination, or residual tranching they would need in order to make the majority of the securities at least investment grade[3]. Second, a sponsor could arrange for a monoline insurer to guarantee against credit risk[4]. Finally, and

[1] See Hayre (2001, p. 12) for a comparison of the historical performance of U.S. mortgage, corporate, and treasury securities between 1982 and 1999.

[2] The value of this put option depends on whether or not the mortgagor has recourse to assets of the mortgagee and the pace at which this recourse can be executed. The value of this put option depends on the form of recourse available. For more information see Ghent and Kudlyak (2011).

[3] A bond is considered investment grade if its credit rating is BBB- or higher by Standard & Poor's or Baa3 or higher by Moody's or BBB(low) or higher by DBRS and Fitch.

[4] A monoline covers only a single line of insurance. In exchange for an enhancement fee they provide an unconditional guarantee of payment to holders of bonds in the event of a verifiable default event. They are called

most importantly, sponsors depended on the rating agencies to certify the adequacy of these taken precautions to enable most of the securities to be rated investment grade or higher.

2.2 Structured Credit, Credit Rating Agencies, and Structured Credit Investors

Two major parties in the securitization chain are the credit rating agencies (CRAs) and the structured credit investors (SCIs) (in addition the issuer of securitizations). The main reason why these parties are considered in this thesis is the availability of information. First, as was noted in chapter 1, in comparison to other parties, the methods of CRAs are well documented in both the academic literature as well as in documents issued by the rating agencies themselves. Second, the role of SCIs is documented in numerous primers issued by investments banks and in handbooks that are available in print.

2.2.1 Credit Rating Agencies

CRAs summarize the quality of a debtor and inform market participants about repayment prospects. The credit rating industry is highly concentrated, with two companies, Standard & Poor's Ratings Services (S&P) and Moody's Investors Service (Moody's), dominating the market in most countries, and Fitch and Dominion Bond Rating Service (DBRS) following behind. This high concentration can be attributed to high barriers to entry (Hill, 2004). The major barriers are caused by the reputational capital and the scope of coverage built by the major CRAs over time. The market shares in structured credit markets provide a similar picture (see Appendix A for more information).

A credit rating is a precondition for a debt offering in virtually every country with a debt market, since credit ratings are put into the requirements of most regulators (White, 2010). The credit rating is a measurement of relative credit risk; it is an opinion on the creditworthiness of a debt issue or issuer. The credit rating is summarized as a discrete alphanumeric mark, which is periodically reviewed over the bond's or note's life (Ashcraft, et al., 2010).

In structured credit market CRAs assign credit ratings to the various issues of notes or tranches backed by the assets in the structure of the structured credit product. The note or tranche rating in structured credit markets reflects an opinion about both the credit quality of the reference portfolio and the extent of credit support that must be provided through the transaction's structure in order for the tranche to receive the rating targeted by the deal's arrangers (Ashcraft, et al., 2011; Ashcraft, et al., 2010). Fitch, Moody's, and S&P state that a given rating should in principle have a consistent interpretation through time and across different security types (Ashcraft, et al., 2010; Cornaggia, Cornaggia, & Hund, 2011; Raynes & Rutledge, 2003). Against this goal, CRAs also emphasize their belief that investors desire a degree of rating stability in response to macroeconomic shocks. Therefore, ratings are revised only gradually in response to changes in economic conditions, a practice known as 'rating through the cycle'[5].

It is important to note that ratings from various agencies do not convey the same information. DBRS, Fitch, and S&P perceive its ratings primarily as an opinion on the probability of default of a tranche (i.e., a certain securities' rating expresses a certain probability of default for such security, with a

monoline because state regulators imposed requirements on the capital structure of these insurers and restricted the type of risk they could take to the liability for third-party debt (Kregel, 2008).

[5] Through-the-cycle ratings are intended to measure default risk over long investment horizons and respond only to changes in the permanent component of credit quality (Altman & Rijken, 2004).

default being defined as a "first-dollar-loss" of a tranche). Moody's ratings on the other hand, tends to reflect the agencies' opinion on the expected loss on a tranche (i.e., a certain securities' rating expresses a certain expected loss for such a note, with an expected loss generally being defined as probability of default times loss severity in case of default). In this thesis these two approaches are termed, respectively, the probability of default (PD) approach and the expected loss (EL) approach.

Tranche Rating Methodology

In this sub-section the tranche rating methodology for a Residential Mortgage Backed Security (RMBS) is reviewed. This is done to illustrate the quantitative and qualitative assessments that rating agencies make. The tranche rating methodology for RMBS is reviewed, since RMBS form the largest structured credit class and the general methodology used in the remaining securitization classes is similar.

The arranger of the RMBS initiates the rating process by sending the credit rating agency a range of data on each of the loans to be held by the SPV, the proposed capital structure of the trust and the proposed levels of credit enhancement to be provided to each RMBS tranche issued by the SPV (SEC, 2008). The RMBS rating process itself involves a combination of quantitative measures and qualitative assessments.

In terms of quantitative measures, CRAs maintain prepayment, default, and loss models, which use as inputs macroeconomic variables, as well as loan characteristics and estimated asset correlations. Asset correlations within the portfolio determine default correlations and thus the likelihood of occurrence of joint defaults in a given period. The CRAs simulate paths of the macroeconomic variables, which, together with the loan characteristics and asset correlations, are used as input for their models to calculate a curve of prepayments, defaults, and losses (Ashcraft, et al., 2010). The individual curves are then aggregated across paths to produce a single curve for each of the three variables. This distribution is used to set subordination levels below each rating class, after taking into account credit enhancement features such as excess spread and insurance (Ashcraft, et al., 2010). The subordination level determines how tranches are protected from default losses (in general, realized losses are first absorbed by subordinated tranches) [6]. The assumptions behind the ratings are revised on a regular basis.

Besides the quantitative measures, each rating methodology involves a number of key areas where qualitative assessments must be applied. These include a review of the legal documentation, the structure of the models used, and decisions about forward looking measures such as the distribution of changes in the aforementioned macroeconomic variables (Ashcraft, et al., 2010). Specific ratings for each RMBS deal also incorporate further subjective assessments of the quality of mortgage originators and servicers, representations and warranties, and other judgmental adjustments (Ashcraft, et al., 2010).

This can be translated to the PD and EL approaches (Münkel, 2006). Under the PD approach quantitative and qualitative analysis is combined in the following way. DBRS, Fitch, and S&P typically calculate within (stress-) scenarios whether the SPV is able to pay interest in full and on time on a note and principal on time (typically upon maturity). Failure of the above leads to a securities' default. Alternatively, under the EL approach Moody's derives a (stress-) scenario based loss that arises from the reference portfolio and has to be distributed to the respective noteholders (i.e., what is

[6] For each tranche subordination is defined as (Ashcraft, et al., 2010):

$$1 - \frac{(face\ value\ of\ securities\ at\ that\ class\ or\ a\ more\ senior\ class)}{(total\ face\ value\ of\ mortgages\ in\ the\ deal)}$$

the size of losses a noteholder has to bear in a given rating/stress scenario). Moody's determines the probability of such a loss to occur while taking into account the average life of the respective note. The sum product of these scenarios will produce an expected loss and a weighted average life for a certain note. Thus, in other words: Expected Loss = (present value scheduled interest + principal payments note) – (present value interest + principal payments note in stress-/rating scenario), with the discount factor being equal to the coupon rate.

2.2.2 Structured Credit Investors

Besides CRAs, Structured Credit Investors (SCIs) are another important category of structured credit market participants. Structured credit securities are often designed to appeal to particular kinds of institutional investors: structured credit investments of less than $1 million are not common. A network of structured credit dealers sell, trade, and make markets in structured products. These transactions are executed over-the-counter (OTC), directly from dealer to dealer, rather than through an exchange.

Table 1 shows the holdings by investor type for U.S. mortgage related securities. This gives an idea of the type of investors that invest in these securities[7].

Table 1
U.S. Mortgage-Related Security Holdings by Investor Type

	$ billions				% of total			
Investor Type	**2006****	**2005**	**2004**	**2003**	**2006****	**2005**	**2004**	**2003**
FDIC Commercial Banks	969.8	897.1	876.4	775.6	14.4%	14.1%	15.6%	15.5%
All Thrifts	242.9	242.6	234.3	206.5	3.6%	3.8%	4.2%	4.1%
Federal Credit Unions	70.5	54.5	27.5	28.5	1.0%	0.9%	0.5%	0.6%
Depository	**1283.2**	**1194.2**	**1138.2**	**1010.6**	**19.1%**	**18.7%**	**20.3%**	**20.2%**
FNMA/FHLMC Portfolio	1150.0	1123.2	1260.9	1232.5	17.1%	17.6%	22.5%	24.6%
Foreign Investors	850.0	802.0	490.0	285.0	12.6%	12.6%	8.7%	5.7%
Mutual Funds	400.0	405.0	375.0	387.0	5.9%	6.3%	6.7%	7.7%
All other investors*	387.5	360.0	201.2	261.9	5.8%	5.6%	3.6%	5.2%
Personal Sector	360.0	355.0	235.0	200.0	5.3%	5.6%	4.2%	4.0%
Life Insurance Companies	300.0	285.0	265.0	240.0	4.5%	4.5%	4.7%	4.8%
Public Pension Funds	190.0	180.0	152.0	120.0	2.8%	2.8%	2.7%	2.4%
Private Pension Funds	175.0	160.0	115.0	105.0	2.6%	2.5%	2.0%	2.1%
FHL Banks	127.8	122.3	113.1	97.9	1.9%	1.9%	2.0%	2.0%
Securities Brokers & Dealers	115.0	95.0	50.0	35.0	1.7%	1.5%	0.9%	0.7%
REITS	112.1	107.4	79.0	28.6	1.7%	1.7%	1.4%	0.6%
Major Investors	**4167.4**	**3994.9**	**3336.2**	**2992.9**	**61.9%**	**62.6%**	**59.4%**	**59.7%**
Total Outstanding	**6733.8**	**6383.3**	**5612.6**	**5014.1**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Note. From "The Changing Face of the Mortgage Market," by L. Goodman (2007), Retrieved August 8, 2011 from http://www.mortgagebankers.org/files/CREF/docs/2007/ViewfromWallStret-WallStreetAnalystsUpdate-LaurieGoodman.pdf
***** Other investors include hedge funds, nonprofits, property/casualty insurers, state/local governments, and other groups
****** Midyear

[7] Interesting to note:
1. The split between holdings of securities backed by mortgages that are in depository and securities that are held by end-investors.
2. The total outstanding amount of mortgage related securities in the U.S. increased by almost 35% in the 2003-2006 period.

Structured Credit Product Valuation in a Nutshell

A structured products' intrinsic value is closely related to the interest and principal cash-flows due on the notes and the likelihood and timing of those being made in part or full (Servigny & Jobst, 2007). For RMBS there are two key risks impacting the likelihood of these payments being made: credit and prepayment risk. Credit risk manifests itself in the form of defaults, delinquencies, and losses. These variables interact with each other to reduce the total amount of principal and interest available to bondholders. Prepayment risk manifests itself in the form of (partial) prepayments being made faster or slower than anticipated. In the case of prepayments, investors receive their proceeds more quickly than originally anticipated. This forces investors to reinvest the notional amount at levels that may be suboptimal. Also, prepayments tend to limit the interest payments available to investors and hence the value of the asset.

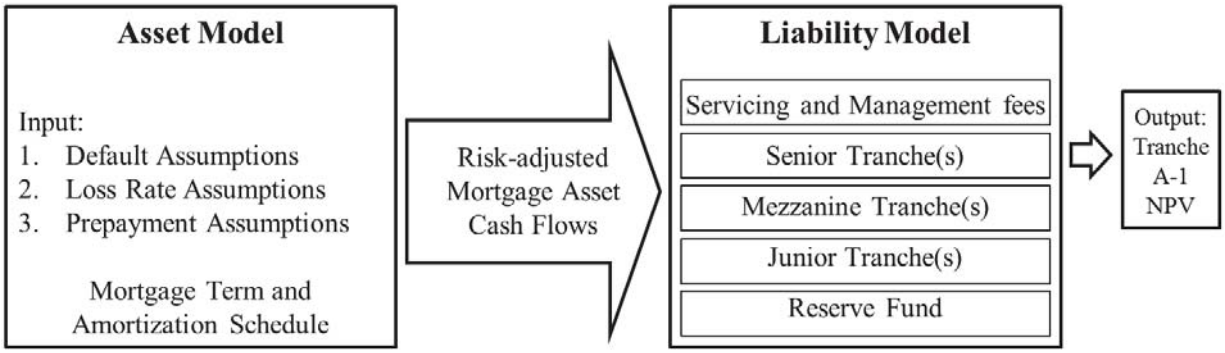


Figure 6: Modeling of Structured Credit Products, Based on Servigny & Jobst (2007, p. 577).

Figure 6 shows the process of valuing a RMBS tranche. In general, a valuation of a structured credit product can be broken down into three components (Kothari, 2006; Melennec, 2000). First an analysis of the collateral is made. On the basis of this analysis the inputs for the asset model shown in Figure 6 are generated. Next, the credit enhancement and, finally, cash flow mechanics are analyzed, which taken together translate into the liability model of Figure 6. These three steps are shortly discussed below.

First, there are the underlying assets. The underlying assets are the most important aspect of the evaluation, because whether or not the issued securities are repaid depends in the first instance on the performance of its assets. The most significant aspects to examine are (Kothari, 2006; Melennec, 2000):

- Quality final debtors: Final debtors can be individuals (e.g., a RMBS) or companies (e.g., a commercial mortgage backed security (CMBS)). The credit risk of a residential mortgage portfolio containing thousands of mortgages can be analyzed using actuarial approaches and is strongly linked to the economic cycle. This approach is less suited to a product based on, for example, five commercial mortgages, which are generally analyzed on a debtor by debtor basis.
- Maturity monitoring: The maturity of the collateral can be long-term or short-term. Long-term bonds issued on collateral consisting of 30-day commercial debts (short-term) are generally revolving. Revolving means that the SPV uses collateral cash flows to buy new collateral over a reinvestment period. This reinvestment period is defined in the deal documentation. In a revolving pool the deal documentation also specifies the reinvestment criteria that protect the collateral quality of the pool. Alternatively, mortgage loan portfolios with a long maturity

(e.g., 15 or 30 years) are referred to as static: the portfolio constituency does not change during the course of the transaction and is amortized with each repayment[8].

- Loan origination: Origination covers the issuing process of a new loan. Inspection of the issuers' origination procedures provides an indication of the riskiness of the collateral. Origination is more important when the transaction includes a reinvestment phase in new assets, due to be generated at a later date. This is especially the case for future flow ABSs, securitizations of future receivables. The securitized portfolio is only described by the asset creation procedure.

Second, there is the credit enhancement. As was discussed in section 2.1, credit enhancements are cushions against defaults and prepayments build into the deal structure. This is done in such a way that they do not affect the quality of issued securities. Credit enhancements are based on assumptions regarding the timing and size of the defaults and prepayments. It is important to evaluate both these assumptions and the sensitivity of the assumptions to changes.

Third, are the cash flow mechanics. The SPVs main sources of income are the final debtors' principal and interest payments. The deal documentation specifies how this is allotted to the bond holders. This is important to understand, since the order of subordination of the various tranches depends on the principles of cash flow allocation.

2.3 Vintage Analysis in Practice

As was shown in the last section, the data requirements for CRAs and SCIs are relatively straightforward. CRAs want a prediction of the cash flows and risk so that a tranche can be rated; SCIs use the same information, but their goal is to value the tranche. In the case of a RMBS these goals can usually be simplified to three assumptions: the default rate, prepayment rate, and the loss severity.

2.3.1 Rating Methodology and the Role of Vintage Analysis

CRAs use vintage analysis to generate input parameters for their structured credit rating models. As the future behavior of the underlying asset pool is uncertain, the CRAs' models are based on a probabilistic approach using historical vintage data. In this section the role of vintage analysis in the structured credit rating process is illustrated by discussing Moody's rating approach in more detail (Moody's, 2005, 2009). The role of vintage analysis in the structured credit rating process of DRBS, S&P, and Moody's is similar and will therefore not be discussed further.

According to Moody's, "the analysis of historical information based on static vintages is one of the most effective approaches to infer accurate parameters for the determination of the default probability distribution of future pools" (Moody's, 2005, p. 9). The backbone of Moody's approach is the default probability distribution. Typically, Moody's assumes these losses to be log-normally distributed. The distribution describes various cumulative default scenarios that can be experienced by the underlying collateral pool and assigns a probability of occurrence to each of the scenarios (Moody's, 2005). Historical default data of the originator, or a similar portfolio, is analyzed in order to determine the mean cumulative default rate and standard deviation for a particular transaction.

It is this historical analysis where vintage analysis plays a role. In line with the definition presented in chapter 1 and the current chapter, vintages are created with each vintage representing all loans originated within a given time period. Default rates are tracked separately for each vintage for each

[8] Note that sometimes mortgage backed securities are also structured as revolving pools.

month after origination. On the basis of this information cumulative default rates are produced that reflect all loans that have defaulted since origination up to any given time period (an example that is using quarterly updated data was shown in Figure 1 in chapter 1).

For those vintages that have been recently originated, and therefore have less default data (e.g., the 2010 vintage shown in Figure 1 in chapter 1), Moody's extrapolates the default rates to the remaining up to 4 years after origination following the historical pattern observed on older vintages (Moody's, 2005, 2009). The extrapolation process may introduce biased and distortive effects on the calculation of the standard deviation. For this reason, Moody's also considers the calculation of the standard deviation based on raw data.

Qualitative assessments also must be applied. For example, Moody's makes further adjustments to the raw mean and the standard deviation of the default rate probability distribution to take account of the seasoning of the pool being securitized and the economic environment during the vintage history (Moody's, 2005, 2009). Moody's reasoning is that if a pool is seasoned, part of the cumulative defaults of the pool have already been realized and these defaults will not ultimately affect the cash flows to noteholders. In addition Moody's notes that through its rating committee process, the analysis of macro-economic variables, originator-specific features (e.g., tightening of collection procedures), and other qualitative aspects can lead to an adjustment of the historical data examined for each transaction.

2.3.2 Structured Credit Investments and the Role of Vintage Analysis

As was explained in chapter 2.2.2, and visualized in Figure 6 of the same section, a valuation of a structured credit product, in general, can be broken down into three components: analysis of the collateral, credit enhancement, and cash flow mechanics. Step 2, the credit enhancement and step 3, the cash flow mechanics, basically involve an interpretation of the deal prospectus and translating this into a liability model[9]. These models are usually obtained via a subscription to a provider of structured fixed-income cashflow models.

The translation of the results of step 1, the analysis of the collateral, into an asset model is the part where most assumptions have to be made. It is this step where vintage analysis plays a role. As was explained in sections 2.2.1 and 2.2.2, assumptions have to be made regarding the default rate, prepayment rate, and the loss severity. These three assumptions are used to produce the input variables of the asset model as displayed in Figure 6; the default, loss, and prepayment vectors. The default and prepayment rates are, in general, assumed to be a constant default rate (CDR) and a constant prepayment rate (CPR); fixed annual percentage rates of defaults and prepayments that are applied to the collateral (Choudhry, Joannas, Landuyt, Pereira, & Pienaar, 2010; Fabozzi, 2000). The CDR and CPR thus are average level of defaults and prepayments in the portfolio over the life of the projections. Loss severity is the amount of losses, including both missed interest and principal write-downs, incurred by a defaulted security or loan, as a share of its principal balance. The underlying mortgages of a RMBS deal, for example, have collateral backing them; in case of a default the collateral is used to recover part of the outstanding balance. So the default amount times the loss severity gives the ultimate loss amount. The development of assumptions is largely based on expert judgment of the analyst(s). This will be explained below.

[9] The prospectus is a description of the structured credit deal that is drafted by attorneys and details all the agreements, duties and responsibilities of all parties involved, expenses and payments, and ultimately the interest and principal to be returned to the investors (Hu, 2011).

Figure 7 shows an example of a vintage graph that is created from data representing the whole German auto loan business of Volkswagen Bank GmbH, which includes both securitized and not securitized loans issued between January 2004 and March 2011. The loans are grouped by the month in which they were issued, thereby creating 77 vintages. Each vintage is represented in the graph by an individual line. The horizontal axis indicates the age of the loan vintage in months. The vertical axis indicates the cumulative net losses as a percentage of the outstanding discounted principal balance.

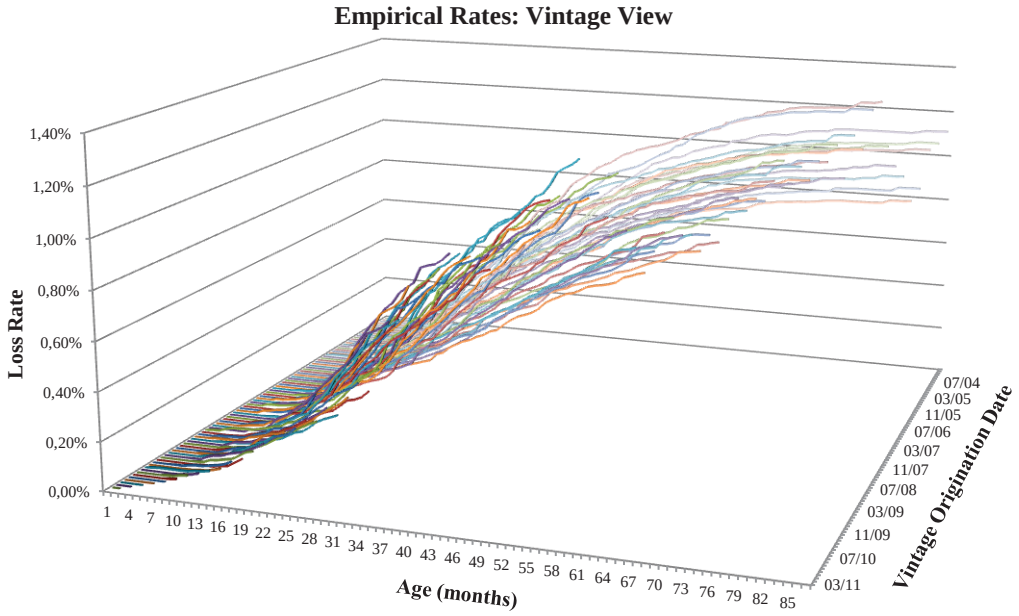


Figure 7: Loss-Rates of All German Auto Loans Issued by Volkswagen Bank GmbH Between January 2004 and March 2011. Data source: RBS/UniCredit (2011) and author's calculations.

Assume an investor is considering investing in a pool of German auto loans. When no prior static pool history is available for the deal, loss curves are constructed using the static pool history of other lenders judged to have similar underwriting characteristics (peers) (Raynes & Rutledge, 2003). In that case, the vintage diagram shown in Figure 7 and the accompanying data aid investors in determining the right input parameter for the loss rate of the portfolio (the loss rate is the default rate times the loss severity rate). Eyeballing the graph indicates that there seems to be homogeneity among vintages: each vintage follows more or less the same trajectory. When it is perceived that the pool on which the graph is based accurately reflects the pool under investigation, then the procedure used by most private-sector analysts to arrive at their estimates is to extrapolate the performance – defaults, loss severities, and total loss rates – of each vintage, based on its own history as well as the typical progression pattern through time (Greenlaw, Hatzius, Kashyap, & Shin, 2008). Next, these parameters are adjusted for both the perceived effect of future extrinsic effects, such as the economic cycle and the seasoning of the loans.

Alternatively, when the data graphed in Figure 7 is considered to be similar to the population, but the pool underlying the investment is considered to be somewhat different than the population, then the data underlying the graph could be considered a base-case. An assessment of the riskiness of the deal under investigation compared to this base-case is made and subsequently performance of the vintages are adjusted for the perceived riskiness compared to the base-case. The perceived riskiness includes the perceived effect of extrinsic effects, such as the economic cycle and the seasoning of the loans.

2.4 Summary

The goal of this chapter was to show how vintage analysis is currently applied by structured credit market participants, thereby answering the first sub-question of this thesis. The chapter started with a definition of the vintage level of analysis and compared it to other levels of analysis. Next, a short primer on structured credit and securitization was presented and two important market participants were introduced: the CRAs and the SCIs. Subsequently, an introduction to how CRAs and SCIs conduct structured credit analysis was given. Both participants create cashflow models that take into account the cashflow mechanics and credit enhancement of the deal. It was explained that the creation of these models involved an interpretation of the deal documentation. The analysis of the collateral is the part were vintage analysis plays a role for both parties. This role is slightly different for each of the parties, since CRAs want a prediction of the cash flows and risk so that a tranche can be rated; investors use the same information, but their goal is to value the tranche.

To see how CRAs predict risk, an introduction to the structured credit rating methodology was discussed and a distinction made between the PD and EL approaches. The role that vintage analysis plays in the rating process was discussed by analyzing Moody's structured credit rating approach in more detail. Vintage analysis plays a role in the analysis of the historical default data that is used to determine default rates for a particular transaction. Moody's makes further adjustments to take account of the seasoning of the pool being securitized and extrinsic effects such as the economic environment during the vintage history.

For SCIs, vintage analysis is used in the translation of the analysis of the collateral into an asset model were assumptions have to be made regarding the default rate, prepayment rate, and the loss severity. When enough data is available the performance – defaults, loss severities, and total loss rates – of each vintage is extrapolated, based on its own history as well as the typical progression pattern through time and adjusted for the perceived effect of future extrinsic effects such as the economic cycle and the seasoning of the loan pool. Alternatively, a base-case scenario for these variables can be constructed on the basis of data which is considered to accurately reflect the population and an adjustment is made to reflect the perceived riskiness of the portfolio under investigation compared to the base-case. The output, three assumptions regarding the performance, are used to produce the input variables of the asset model: the default, loss, and prepayment vectors.

The next chapter, chapter 3, provides an answer to the second sub-question of this thesis: What is cohort analysis?

3 Cohort Analysis

> *"The fundamental question in cohort analysis is that of determining whether the phenomenon under examination is cohort-based, or whether some other conceptualization – age-based for example – is more appropriate .... Even a superficial examination of the 'either-or' question leads to the conclusion that 'both' might also be acceptable. That is, there is, in general, no logical reason for ruling out the possibility that both cohort and age may be relevant to the study of some phenomenon. Furthermore, it is but a short step to conclude that not only might aging- and origin-related processes (i.e., age and cohort) be relevant to the matter at hand, but also that instantaneous processes (i.e., period) might also be pertinent. Once this point is accepted, however, the problem of distinguishing the effects of age from those of period and cohort can become difficult."*
>
> W.M. Mason and Fienberg (1985a, pp. 1-2)

This chapter answers the second sub-question of this thesis: What is cohort analysis? This is done by introducing cohort analysis and the developments in this field. This is covered in the following manner. Section 3.1 defines cohort analysis and the age, period, and cohort effects that cohort analysis aims to unravel. Before different cohort analysis techniques are introduced, two concepts in cohort analysis are presented. First, section 3.2 makes a distinction between different conceptualizations of cohorts and categorizes these into a sociologic and an epidemiologic conceptualization of cohorts. This distinction is made because they result in a fundamentally different execution of cohort analysis. Second, section 3.3 introduces a standard cohort table that serves to discuss the logic behind cohort analysis. Once the logic is presented, a bridge can be made to the identification problem in cohort analysis. Since the identification problem lies at the basis of the plethora of cohort models available in the academic literature an elaborate discussion of this problem is presented in the second part of section 3.3. Furthermore, section 3.4. starts with a categorization of a selection of cohort analysis models available in the literature. Finally, four of these approaches are discussed in more detail.

3.1 Cohort Analysis and Age, Period, and Cohort Effects

The identification and estimation of distinct effects for age, time period, and cohort on outcome variables or event rates, has long been a goal of analysis in demography, medicine (epidemiology and biostatistics), sociology, political science, and other social sciences. Age refers to the time since a subject or entity entered a study; period refers to the calendar date at which the outcome was determined; and a cohort consists of "people, entities or objects who share a common experience during a specified period of time" (Glenn, 2005, p. V). Thus, a cohort identifies the calendar time when an individual or entity entered the study.

Age-period-cohort analysis or cohort analysis in short, is one of the methods used in an effort to separate the effects of age, period, and cohort. The opening quotation of this chapter by Mason and Fienberg provides guidance on why it is important to look at the age, period, and cohort effects. Cohort analysis is a form of longitudinal study. The goal in longitudinal studies is to measure change over time by collecting data concerning at least two time points (de Vaus, 2001). Several definitions of cohort analysis can be found in the literature:

1. Liao defines cohort analysis as "quantitative research using a measure of the concept of cohort and studying its effect on some outcome variable(s)" (as cited in Glenn, 2005, p. V).

2. Glenn (2005) proposes a more specific definition of cohort analysis by describing it as "studies in which two or more cohorts are compared with regard to at least one dependent variable" (p. 3).
3. W.M. Mason & N.H. Wolfinger (2001) define cohort analysis as "studies that seek to explain an outcome through exploitation of differences between cohorts, age and period" (p. 2189).
4. Keyes et al. (2010) provide a somewhat similar definition and define cohort modeling strategies as "statistical attempts to partition variance into the unique components attributable to age, period, and cohort effects" (p. 1102).

These definitions were placed in specific order from abstract to concrete. Liao's definition provides no guidance on how to measure the effect of cohort on the outcome variable. In this respect, the definition provided by W.M. Mason and N.H. Wolfinger provides more direction. Taking into account the four definitions above, in this thesis *cohort analysis* is defined as: 'an analysis technique in which statistical attempts are made to partition (variance in) the outcome on an independent variable into the unique components attributable to age, period, and cohort effects'.

3.1.1 Age, Period, and Cohort Effects

As explained above, cohort analysis distinguishes between three types of time-related variation in the phenomena of interest: age effects, period effects, and cohort effects.

Age effects are the variations associated with different age groups. These effects represent the common developmental processes that are associated with particular ages or stages in the life course (Yang, 2008). Age often influences risk of diseases and socio-economic outcomes (Holford, 2005). In epidemiologic studies, for example, age effects represent aging-related physiological or developmental changes and offer clues to etiology[10] (Yang, 2010).

Period effects are the variations over time periods that affect all age groups under observation simultaneously. In demographics, for example, period effects reflect different formative experiences resulting from the intersection of individual biographies and macro-social influences (Yang, 2010).

Cohort effects reflect the changes across groups of individuals, entities or objects that experience an initial event, such as birth, during the same time period. A birth cohort in demographics, for example, shares the same birth years and ages together. Birth cohorts born in different time periods conceivably develop different life course careers due to different historical and social conditions that are encountered (Yang, 2010).

3.2 Cohorts: Sociologic and Epidemiologic Conceptualizations

As was noted at the start of this chapter, cohort analysis is used in various disciplines of science. Cohort analysis in different disciplines served different purposes and, therefore, developed in different directions.

One of the assumptions that underlies many of the cohort methods that have been developed over the last decades is the assumption that age, period, and cohort effects are additive (Glenn, 1976). The additivity assumption implies that age effects are the same for all cohorts and periods, that cohort effects are the same for all ages and periods, and that period effects are the same for all ages and cohorts, i.e., there is no interaction effect (Glenn, 1976). However, it might be possible that a period-specific event only has an impact on the behavior of a specific cluster of cohorts in a certain age phase.

[10] Etiology is a branch of medical science concerned with the causes and origins of diseases.

For that reason, in a more recent work, Glenn (2005) notes that the additivity assumption can be easily tested and rarely holds in practice.

Several decades after this observation was first made by Glenn, Keyes et al. (2010) observed that analysts often use different conceptualizations of cohorts, and therefore, different research questions, statistical methods, analyses, and interpretations which lead to different empirical results. On the basis of this notion, Keyes et al. distinguish between two different conceptualizations: an epidemiologic definition of cohorts and a sociologic definition of cohorts. These definitions can be related to the additivity assumption since they involve a distinction between a view where age and period are confounders of cohort effects and the view where cohort effects are a result of the interaction of age and period effects. This notion and the definitions of the different conceptualizations of cohorts will be elaborated upon below.

3.2.1 The Sociologic Oriented Conceptualization of Cohorts

Sociological theories place focus on cohorts and on determining the way in which cohort membership affects the lives of persons across the life course. This view was popularized by Ryder (1965). He made an extended argument for the conceptual relevance of cohorts to a range of substantive issues in social research. Ryder argued that cohort membership could be as important in determining behavior as other social features such as socioeconomic status (Yang, 2010). Also, Ryder posited that a cohort can be conceived as a structural category, whereby the unique circumstances and conditions through which cohorts emerge, come of age, and die, provide a record of social and structural change (Keyes, et al., 2010). As a result, the conditions and the resources that each cohort is born into, and in which they live their collective lives, may uniquely shape the patterns and experiences of health and mortality for that cohort (Keyes, et al., 2010).

Studies adopting this conceptualization of cohorts often posit cohort effects as representing the totality of environmental influences for a specific birth group that are unique to the cohort itself. The effects of period and age make it difficult to identify a cohort effect, because all three variables are linked with time. Separating the effects of historical influences (cohort effects), contemporaneous influences (period effects), and exposure accumulation (age effects) becomes necessary to obtain a unique estimate of cohort effects under the assumptions of the sociological definition (Keyes, et al., 2010).

In other words, the sociological definition conceives of age and period as confounders of the cohort effect. Confounding exists if meaningfully different interpretations of the relationship of interest result when an extraneous variable is ignored or included in the data analysis (Kleinbaum, Kupper, & Muller, 2007). This variable poses a problem when it is unequally distributed between the cohorts. The most common concern about confounding is that it may create the appearance of a cause-effect that does not actually exist (Kleinbaum, et al., 2007). The sociological definition assumes that cohorts have unique characteristics confounded by age and period effects.

3.2.2 The Epidemiologic Oriented Conceptualization of Cohorts

The epidemiologic definition of a cohort effect, on the other hand, suggests that a cohort effect occurs when different distributions of outcome variables or an event rate arises from a changing or new environmental cause affecting age groups differently (Keyes, et al., 2010). A cohort effect, therefore, is conceptualized as a period effect that is differentially experienced through age-specific exposure or susceptibility to that event or cause (i.e., interaction or effect modification). A cohort effect can affect a population in two different ways (Keyes, et al., 2010):

1. A population-level environmental cause is unequally distributed in the population.
2. A population-level exposure differentially affects age groups who are in the midst of a critical developmental period, during which exposure has long-lasting effects on lifetime outcome variables or an event rate.

In contrast to the sociological definition which conceives of age and period as confounders of the cohort effect, the epidemiological definition conceives of cohort effects as the interaction or effect modification of period and age effects (Keyes, et al., 2010). Interaction is the condition in which the relationship of interest is changing at different levels or values of the extraneous variable (Kleinbaum, et al., 2007). The epidemiologic definition assumes that period and age effects interact to produce cohort effects. Since the additivity assumption implies that there is no interaction effect, this assumption is only related to the sociological conceptualization of cohorts.

3.3 The Logic Behind Cohort Analysis and the Identification Problem in Cohort Analysis

Before different cohort analysis models are presented, it is useful to first introduce the logic behind cohort analysis and to get insight into the identification problem. This allows for a better understanding of the purpose and (consequences of) the assumptions behind the many models that were developed over the last decades.

3.3.1 Cohort Tables

A first step in cohort analysis is the tabulation of the data. The standard cohort table is constructed by cross-sectional data sets juxtaposing the relationship between age and some dependent variable, with the age intervals equal to the intervals between periods for which there are data (de Vaus, 2001; Glenn, 2005; O'Brien, 2010). An example of a cohort table, under the assumption of the sociologic conceptualization of cohorts, is shown in Table 2.

Table 2
Cohort table for the case of 4 periods and 4 age groups.

		Period (j)			
		$j = 1$	$j = 2$	$j = 3$	$j = 4$
Age Group (i)	$i = 1$	$\mu + \alpha_1 + \eta_1 + \gamma_4$	$\mu + \alpha_1 + \eta_2 + \gamma_5$	$\mu + \alpha_1 + \eta_3 + \gamma_6$	$\mu + \alpha_1 + \eta_4 + \gamma_7$
	$i = 2$	$\mu + \alpha_2 + \eta_1 + \gamma_3$	$\mu + \alpha_2 + \eta_2 + \gamma_4$	$\mu + \alpha_2 + \eta_3 + \gamma_5$	$\mu + \alpha_2 + \eta_4 + \gamma_6$
	$i = 3$	$\mu + \alpha_3 + \eta_1 + \gamma_2$	$\mu + \alpha_3 + \eta_2 + \gamma_3$	$\mu + \alpha_3 + \eta_3 + \gamma_4$	$\mu + \alpha_3 + \eta_4 + \gamma_5$
	$i = 4$	$\mu + \alpha_4 + \eta_1 + \gamma_1$	$\mu + \alpha_4 + \eta_2 + \gamma_2$	$\mu + \alpha_4 + \eta_3 + \gamma_3$	$\mu + \alpha_4 + \eta_4 + \gamma_4$

Note. μ denotes the overall mean; α_i denotes the fixed effect of the i[th] age category, η_j denotes the fixed effect of the j[th] period category, and γ_{a-i+j} denotes the fixed cohort effect associated with the i[th] age category and the j[th] period category. From "The Age–Period–Cohort Conundrum as Two Fundamental Problems," by R. M. O'Brien, 2010, Quality & Quantity, p. 3.

Each cell contains an expected value of the dependent variable on a particular measurement date j. From a cohort analysis perspective this value is a combination of an overall mean, age effect, period effect, and cohort effect. Each column (j) in the table is a set of cross-sectional data in which, consistent with the sociological definition, age and cohort are confounded. Similarly, in each row (i) in which there are data on four different cohorts when they were at the same age level, period and cohort effects are confounded. Finally, each cohort represented in the table, except the one that was in age group 4 in period 1 and the one that was in age group 1 in period 4, can be traced for at least two age groups as it grew older by starting in the leftmost cell in which it is represented and reading diagonally

down and to the right. In the data, in each of these cohort diagonals, age and period effects are confounded.

Visual inspection of a table filled with data is sometimes misleading, since nonappearance of an observable relationship does not confirm that such a relationship is absent. The data in each column of the standard cohort table suffer from the same confounding or interaction of effects as the data from a cross-sectional study, and the data in each cohort diagonal confound age and period effects or are the result of interaction of age and period effects in the same way as do the data from a panel study (Glenn, 2005).

However, in the cohort table, there are multiple columns and multiple cohort diagonals which raised the hopes of many researchers adhering to the sociological conceptualization of cohorts for a way to use statistical procedures to separate age, period, and cohort effects (Glenn, 2005). Nevertheless, several authors warned that it is impossible to statistically separate age, period, and cohorts effects, except when all effects are non-linear (Glenn, 1976, 2005; Holford, 2005; Rodgers, 1982a). This is the result of the 'model identification problem' in cohort analysis. In the following section this problem is elaborated upon and the implications for cohort analysis are discussed.

3.3.2 The Model Identification Problem in Cohort Analysis

In section 3.1 cohort analysis was defined as an analysis technique in which statistical attempts are made to partition (variance in) the outcome on an independent variable into the unique components attributable to age, period, and cohort effects. The individual effects of age, period, and cohort on a variable of interest are usually estimated by using generalized linear models.

However, as mentioned in section 3.3.1, no statistical model can simultaneously estimate age, period, and cohort effects because of the perfect linear relationship between age, period, and cohort. The perfect linear relationship between age, period, and cohort gives rise to what the cohort analysis literature termed the model identification problem of cohort analysis (Glenn, 1976, 2005; Kupper, Janis, Karmous, & Greenberg, 1985; K. O. Mason, Mason, Winsborough, & Poole, 1973; W.M. Mason & N.H. Wolfinger, 2001; Ryder, 1965; Von Furstenberg & Green, 1974; Yang, 2007; Yang & Land, 2008). This model identification problem, or identification problem in short, exists whenever three or more independent variables need to be included in an analysis and each one is a perfect (linear) function of the others, or, in other words, knowledge of the value of two of the variables on an observation provides knowledge of the third. In the case of cohort analysis the relationship is Cohort = Period – Age ($C = P - A$). Because of the equality $C = P - A$ it is not possible to estimate to separate the effects of cohorts, ages, and periods in a generalized linear model of the form

$$Y = f(C, P, A) \quad (1)$$

Where:

- Y denotes the dependent variable.

without some kind of restriction on the function f.

To further explain the identification problem some statistical terminology has to be introduced. Collinearity exists whenever an independent variable is highly correlated with another independent variable in a regression equation (Allen, 2004). Unity is the case where perfect collinearity exists; one independent variable is a perfect linear function of the other independent variables in a regression equation (Allen, 2004). This occurs, for example, when a variable is constructed as a linear function of

other variables, as is the case with cohort analysis. If we include this variable in, for example, a standard regression equation, ordinary least-squares (OLS) estimation procedures will fail, because all of the variance in the constructed variable can be explained by the variables used to construct it. In other words, generalized linear models such as multiple regression analysis are trying to separate out the effects of two or more variables, even though they are correlated with each other. To separate these effects, however, there must be some remaining variation on a variable when the other variables are held constant. When two variables are perfectly correlated and you hold one of the variables constant, then the other must be constant as well. Hence, it is impossible to separate their effects on the dependent variable.

This can also be explained more formally. In cohort analysis the multiple correlation of each independent variable with the other ones is unity, i.e., perfect collinearity exists. Perfect collinearity in multiple regression results in that it will not be possible to calculate the inverse of the matrix of the covariances among the independent variables, since it is a singular matrix (i.e., its determinant is zero). For instance, given a dataset in the form of the standard cohort table, the relationship between an outcome variable Y and the three variables age, period, and cohort in a normal regression could be written as:

$$Y = \mu + b_1 A + b_2 P + b_3 C + \varepsilon \qquad (2)$$

Where:

- Y denotes the dependent variable;
- μ denotes the intercept;
- b_i denotes the partial slopes (regression coefficients) associated with age, A, with period, P, and with cohort, C; and
- ε denotes the residual error term (the effect of unmeasured variables, measurement errors, and so on).

When it is assumed that all the four variables are standardized (i.e., their means have been subtracted from initial values for each variable), then equation (2) can be written in matrix form as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon} \qquad (3)$$

Where:

- $\boldsymbol{Y}$ denotes a $n * 1$ vector of observations on the response variable;
- $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ denotes a $n * p$ matrix of n observations on p predictor variables. $\boldsymbol{X}$ is also called the design matrix;
- $\boldsymbol{b}$ denotes a $p * 1$ vector of regression coefficients; and
- $\boldsymbol{\varepsilon}$ denotes a $n * 1$ vector of random errors.

Usually it is assumed that the vector of random errors has mean zero, i.e., $E(\boldsymbol{\varepsilon}) = 0$ and a constant diagonal covariance matrix, i.e., $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\boldsymbol{T}}) = Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix of order n (Lazaridis, 1986).

When $\boldsymbol{X}$ is of full column rank, the OLS estimator is the solution $\boldsymbol{b}$ of the normal equation:

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \qquad (4)$$

Where:

- $\boldsymbol{X^T}$ is the transposed matrix of $\boldsymbol{X}$; and
- $(\boldsymbol{X^T X})^{-1}$ is the inverse of $\boldsymbol{X^T X}$.

The linear relationship between the age, period, and cohort variables translates to a design matrix, $\boldsymbol{X}$, that is one less than full column rank (one column can be written as a linear combination of the others). This implies that $\boldsymbol{X^T X}$ is singular, in other words, the inverse of $\boldsymbol{X^T X}$, $(\boldsymbol{X^T X})^{-1}$ does not exist[11]. It follows that the solution to the normal equations is not unique. There will be an infinite number of OLS estimators; one for each possible linear combination of column vectors that result in a vector identical to one of the columns of the design matrix, $\boldsymbol{X}$.

Glenn (2005) notes that if all effects are nonlinear, it may be statistically possible to estimate age, period, and cohort effects with reasonable accuracy; if any major component of the variation is linear, however, a statistical separation of the effects is impossible. Thus far, most researchers posit the latter is the case, which makes the standard generalized models unsuitable (Glenn, 2005).

3.4 Cohort Analysis Models

W.M. Mason and N.H. Wolfinger (2001, p. 2190) state that "the cohort analysis identification problem is the point of departure for all modern discussions of techniques of cohort analysis". Attempts to solve or mitigate the identification problem that was introduced in section 3.3 resulted in a variety of methodological approaches. This section discusses several of the cohort modeling strategies that have been developed over the past decades.

Several authors have made attempts to categorize the plethora of cohort models available in the literature. Robertson and Boyle (1998), for example, distinguished between four classes of cohort models in epidemiology[12]:

1. Models based on (arbitrary) linear constraints;
2. Models based on the use of a penalty function;
3. Models using individual records of cases; and
4. Models based on estimable functions (i.e., functions that do not depend on the constraints adopted to find a particular set of parameter estimates).

In a later article Robertson, Boyle, and Gandini (1999) add a fifth class of cohort models in epidemiology:

5. Models which impose a time-series structure on the time effects (i.e., autoregressive models).

Yang (2005, 2010) distinguishes between cohort models in demography and cohort models in biostatistics and epidemiology. For biostatistics and epidemiologic cohort models, she is referring to the above mentioned classification of Robertson and Boyle (1998), and Robertson et al. (1999). For demographic cohort models Yang distinguishes between three classes:

1. Models based on constraints;
2. Models using proxy variables (i.e., age-period-cohort-characteristic models); and
3. Models using nonlinear transformations.

[11] A matrix has an inverse only if it is square, and even then only if it is nonsingular, i.e., when its columns or rows are linearly independent (Ben-Israel & Greville, 2003).
[12] Robertson and Boyle (1998) note that they only consider cohort models that fall into the class of generalized linear models.

Keyes et al. (2010) distinguish between three different kind of models, based on their notion of the sociological and epidemiological conceptualizations of cohorts (see section 3.2), and the accompanying distinction between 1st order and 2nd order effects:

1. Models in which 1st order effects are estimated and interpreted;
2. Models in which 2nd order effects are estimated and interpreted; and
3. Hybrid models in which 1st order effects are estimated but 2nd order effects interpreted.

1st order effects are determined based on the assumption that age, period, and cohort can exist independently of each other and have a linear relationship with the outcome of interest. Each linear slope is estimated by controlling for the additive effect of the other two effects. These linear relationships are what Keyes et al. term 1st order effects. 2nd order effects are those which have a non-linear relationship with the outcome of interest.

On the basis of what has been discussed so far, several observations can be made regarding the categorization of cohort models. First, the categorizations are, with exception of Yang (2005, 2010), specific to the background of the authors. Robertson and Boyle (1998), Robertson et al. (1999), and Keyes et al. (2010) provide literature reviews of cohort models in epidemiology. Yang (2005, 2010) provides a hybrid review of models in sociology and refers to Robertson and Boyle for an overview of the epidemiologic literature. Second, although the overviews provide a first indication of the models available in the literature, a thorough interdisciplinary literature review is missing. Yang is the only one that makes a distinction between cohort analysis models in sociology and epidemiology. Finally, note that the categorization of 1st and 2nd order effects of Keyes et al. could be combined with the categorizations of the other authors mentioned above to create an additional categorization layer.

Using these observations one possible interdisciplinary categorization of cohort models available in the academic literature is provided in Table 3. First, the models were divided into three broad categories, 1st order, 2nd order, and hybrid models based on the notion of Keyes et al. Second, the example studies referenced in the above mentioned articles were investigated to see whether the categorizations provided any overlap. Finally, additional literature that was studied for this thesis was included in the table.

In the next subsections several models are discussed that, according to the author, provide a decent representation of the categorization of cohort models that was proposed in this section. Section 3.4.1 discusses the Mason, Mason, Winsborough, and Poole method, a 1st order constraint-based method that is widely used in sociology and can be regarded as the 'father' of all 1st order constraint-based models. Section 3.4.2 discusses the Median polish technique, a 2nd order model. Section 3.4.3 discusses the Holford approach, this model and its offspring are widely used in epidemiology. The Holford approach is a hybrid approach between 1st and 2nd order models and deals with several of the issues that were raised with the Mason, Mason, Winsborough, and Poole based methods. Finally, section 3.4.4 discusses the intrinsic estimator approach, a recent development in 1st order constraint based methods which was developed by researchers with backgrounds in biostatistics and sociology.

Table 3

Categorization of Cohort Models

1st Order Models	(Arbitrary) Constraint-Based Models	Equate two coefficient within one of the three dimensions	Barret (1973; 1978); **W.M. Mason, Mason, Winsborough & Poole (1973)**; Knoke & Hout (1974); Fienberg & Mason (1979); Hardings & Jencks (2003)
			Nakamura (1986) - Bayesian approach
		Drop an effect	Firebaugh & Davis (1988); Glenn (1994); Myers & Lee (1998)
		Constrain the effect of a variable to be proportional to (an)other substantive variable(s) (a.k.a. proxy variable approach or APCC model)	Farkas (1977) - Period effect proportional to unemployment rate
			Heckmann & Robb (1985) - Proxy variable approach
			W.M. Mason & Fienberg (1985b); Kahn & Mason (1987) - Cohort effect proportional to cohort size
			Rodgers (1982a); O'Brien , Stockard & Isaacson (1999) – APCC
		Use a penalty function	Osmond & Gardner (1982); Decarli & La Vecchia (1987) - Minimizing penalty function
			López-Abente, Pollán & Jiménez (1993); Chie, Chen, Lee, Chen & Lin (1995)
		Set a slope to zero	Roush, Schymura, Holford, White & Flannery (1985); Roush, Holford, Schymura & White (1987)
		Restrict the range of the slopes	Wickramaratne, Weismann, Leaf & Holford (1989)
		Constraint associated with a generalized inverse	**Yang, Fu & Land (2004) - Intrinsic Estimator (Moore-Penrose generalized inverse)**
	Models based on Individual Records		Robertson & Boyle (1986); Tango (1988); Lee & Lin (1994); McNally, Alexander, Stains & Cartaright (1997)
	Time-series Models		Lee & Lin (1996) – Autoregressive model
			Berzuini & Clayton (1994) - Bayesian approach
2nd Order Models			**Shaphar & Li (1999); Selvin (2004) – Mean/median polish technique**
Hybrid Models			**Holford (1983, 1991, 1992, 2005) - Deviations from linearity/Polynomial Trends**
			Holford, Zhang & McKey (1994) - Logarithmic age effect
			Tang & Kurashina (1987); Clayton & Schifflers (1987); Tarone & Chu (1992, 1996) - Linear/Drift models

Note. Models that are in boldface type are discussed in sections 3.4.1 to 3.4.4

3.4.1 Mason, Mason, Winsborough, and Poole Approach

One of the first approaches to mitigating the identification problem was the 1st order constraint-based regression of Mason, Mason, Winsborough and Poole (1973) in which at least one category of age, period, and cohort is constrained in some manner. This method is also known as the 'cohort accounting' or 'multiple classification' model (Yang, et al., 2004; Yang, Schulhofer-Wohl, Fu, & Land, 2008).

Mason et al. recognize that the model identification problem can be solved by fitting a model in which the relationship of at least one of the three variables to the outcome variable, Y, in equation 2 is constrained to be nonlinear. However, since in most cohort analyses the analyst starts with little prior information about the relationship of either age, period, or cohort to the dependent variable, they consider this method not very useful because it requires assumptions that are too restrictive (K. O. Mason, et al., 1973). Alternatively, Mason et al. propose to use a relatively functional free model: the multiple classification model.

The multiple classification model specifies the dependent variable to be the result of effect parameters associated with particular levels of each independent variable (K. O. Mason, et al., 1973):

$$Y_{ij} = \mu + \alpha_i + \eta_j + \gamma_k + \varepsilon_{ij} \qquad (5)$$

Where:

- Y_{ij} denotes the age–period-specific value on the outcome variable for the i[th] age group for $i = 1, \dots, a$ age groups at the j[th] time period for $j = 1, \dots, p$ time periods;
- μ denotes the intercept;
- α_i denotes the i[th] row age effect or the coefficient for the i[th] age group;
- η_j denotes the j[th] column period effect or the coefficient for the j[th] time period;
- γ_k denotes the k[th] diagonal cohort effect or the coefficient for the k[th] cohort for $k = 1, \dots, (a + p - 1)$, with $k = a - i + j$; and
- ε_{ij} denotes the residual error term (the effect of unmeasured variables, measurement errors, and so on). The vector of errors $\boldsymbol{\varepsilon}$ is such that $E(\boldsymbol{\varepsilon}) = 0$ and $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$; that is, the errors are uncorrelated, with means 0 and variances σ^2.

The model postulates unique effects for each category within each dimension and each dimension is represented exhaustively by its categories. Graybill (1961, p. 227) proves that, given the assumption that errors are uncorrelated, with means 0 and variances σ^2, there exist no linear functions of the observations that yield unbiased estimates for the coefficients of models such as (5) which postulate unique effects for each category within each dimension and where each dimension is represented exhaustively by its categories. However, under the assumption that several age groups, cohorts, or time periods have identical effects on the dependent variable, Mason et al. show that it is possible to estimate differences of the form $(\alpha_{i'} - \alpha_i)$, $(\eta_{j'} - \eta_j)$, and $(\gamma_{k'} - \gamma_k)$ in equation 5 for $i \neq i'; j \neq j'$; and $k \neq k'$. That is, the differences between the effects of any two categories within a dimension become estimable.

Mason et al. demonstrate that the minimal assumption needed to achieve estimability is to assume that two age groups, two time periods, or two birth cohorts have identical effects on the dependent variable. Moreover, they show that estimates of the form $(\widehat{\alpha_{i'} - \alpha_i})$ will vary according to the pair of coefficients

assumed to be equal, although $\hat{Y}_{ij}$ will be the same regardless of which pair this is. This means that three-way cohort analyses are sensible if the researcher has strong a priori conception that allows him to assume equality among various effect parameters.

However, the theoretical basis for such an assumption is often lacking. Mason et al. (1973) propose to employ more restrictions than necessary to overcome the dilemma of choosing what restrictions to apply. The reason for employing more restrictions is that when more than the minimal assumptions needed to achieve estimability are applied (e.g. assume that two pairs of coefficients are equal), not only all estimates for effects will differ, but estimates of Y_{ij} will differ as well. Therefore, distinct models will lead to distinct fits of the data. For this reason Mason et al. maintain that a clearer picture of the "true" effects in a given set of cohort data might be obtained by comparing the results from several distinct models making more than the minimal assumptions needed for estimability (e.g., choosing the model with the largest coefficient of multiple determination, R^2). Mason et al. extend this approach by proposing to perform a stepwise incremental model in which whole dimensions are added or excluded from the model to provide additional information about the ability of the three dimensions to explain variance in the dependent variable.

The approach marked the start of a fierce debate over the methodological merits and flaws of cohort analysis (Glenn, 1976; Knoke & Hout, 1976; W. M. Mason, Mason, & Winsborough, 1976; Rodgers, 1982a, 1982b; Smith, Mason, & Fienberg, 1982). Nonetheless, the Mason, Mason, Winsborough, and Poole approach has been widely used and, as Table 3 shows, knows many variants that have been developed over the last decades. Nakamura (1986), for example, uses a Bayesian approach to specify restrictions.

The advantage of the Mason, Mason, Winsborough and Poole approach is that it is quite simple to understand and to apply from a statistical perspective. Also, it does not require dropping a factor completely from the model. Nevertheless, although the model is easy to apply and mathematically correct in the sense that the linear dependency between age, period, and cohort is broken, it has received some criticism in the statistical literature. The main arguments against the method and its application can be categorized as follows (Glenn, 1976, 2005; Keyes, et al., 2010; Winship & Harding, 2008; Yang, et al., 2004):

1. It is difficult to find restrictions that can be theoretically justified;
2. If the constraints are even slightly misspecified, this can have major consequences for the parameter estimates; and
3. Restrictions are rarely tested.

In other words, the linear dependence is broken in the statistical model only and not in the real world (Glenn, 1976, 2005). Therefore, the obtained results might be meaningless.

Another comment by Glenn (1976) is based on the assumed additivity of the method. Part of his criticism is based on the argument that modeling the effects as additively separable already imposed too many constraints on the model and do not allow for interactions between, for example, cohort and changes over time. As was noted in section 3.2, this additivity assumption rarely holds in practice. W.M. Mason et al. (1976) replied to Glenn's comments by noting that Glenn ignores the purpose of models; they are by

definition a simplification of reality and the focus should be on the insight gained by the model instead of on the flaws in the model.

In the following sections, several models that (partially) deal with the criticism on the Mason, Mason, and Poole based methods are discussed.

3.4.2 Median Polish Technique

The Median Polish Technique (MPT) was introduced to cohort analysis in 1996 by Selvin (Li & Baker, 2012; Selvin, 2004). An application of the MPT can be found in Keyes and Li (2010). MPT estimates cohort effects as partial interaction (2nd order effect) of age and period effects. Thereby, the MPT technique adheres to the epidemiologic definition of cohorts. Interaction effects are 2nd order effects by definition, since they represent deviations from linearity.

The 2nd order effects produced by the MPT model non-linearities in the age and period effect These non-linearities are subsequently partitioned into a systematic component (cohort effect) and an unsystematic component (random error). The MPT explicitly tests whether the effects of age and period interact to produce an effect that is more than what would be expected given their additional influences. As opposed to the Mason, Mason, Winsborough and Poole approach, the MPT estimates a two-factor model (age and period), and therefore, no constraints are necessary (as is the case with the three-factor model with collinear slopes for age, period, and cohort).

Conceptually, MPT can be explained by picturing a standard cohort table, as the one that was introduced in Table 2 in section 3.1.1. The MPT removes the additive effect of age (row) and period (column) by iteratively subtracting the median value of each row and column. After several iterations, the residual values stabilize (i.e., the median residual of each row or column approximates zero). The residuals are then regressed on indicator variables for cohort membership using standard linear regression; the extent to which the cohort variable predicts the residual is the cohort effect. The remaining residual unaccounted for by cohort is considered to be nonsystematic random error.

The relative magnitude of cohort effects by regressing the residuals e_k on cohort category (entered as a collection of indicator variables for the $a * p - 1$ cohorts, $k = 1,2, \dots, a + p - 1$) using linear regression is assessed as follows:

$$e_k = \mu_k + \gamma_k + \varepsilon_{ijk} \qquad (6)$$

Where the residuals, e_k are a function of an intercept, μ_k, with $E(\mu_k) = 0$, a vector of cohort effects, γ_k, and a vector of error terms, ε_{ijk} (the error terms representing the random error unaccounted for by the cohort effect across i age, j period, and k cohort categories).

The MPT tests whether these deviations from additive age and period influences follow a systematic pattern that can be predicted by cohort membership; if so, the deviations are attributed to cohort effects. The γ estimates (one for each cohort category) reflect the log rate that reflects a ratio of cohort effects (i.e., the ratio of the non-additive effect for one cohort to that of the non-additive effect for a reference cohort). The exponentiation of each γ estimate derived from equation 6 indicates the excess rate attributable to each cohort category. Each cohort category can then be compared to the reference cohort to

obtain a relative estimate of the size of the cohort effect. Finally, the residuals from this model can be examined for violations of parametric assumptions.

Note that the MPT is non-unique. The outcome of the MPT may depend on whether rows or columns are tried first and is very resistant to outliers. Starting by operating on columns rather than rows may lead to a different (but qualitatively similar) answer.

3.4.3 Holford Approach

An alternative approach was developed by Holford (1983, 1991, 1992, 2005). The Holford approach can be categorized as a hybrid of the sociological and epidemiologic definition, because "while conceptually the Holford approach acknowledges the interpretive utility of linear effects for age, period, and cohort (i.e., the sociologically-oriented approach), it accepts the reality that these linear effects cannot be estimated validly simultaneously, and thus, focuses on the estimation and interpretation of the non-linear effects (i.e., the epidemiologically-oriented approach)" (Keyes, et al., 2010, p. 1102). The Holford approach estimates the cohort effect as 2nd order function in a model in which 1st and 2nd order age and period effects are considered confounders of the 1st and 2nd order cohort effects.

The Holford approach focuses on linear deviations known as curvatures, a measure which can be interpreted as reflecting changes in the direction or steepness of the slope of the underlying age, period, and cohort effects (2nd order effects) without estimating the magnitude of the actual slope (1st order estimate) (Keyes, et al., 2010). So the curvature is summarizing the overall direction of the non-linear trends over time. Thus, a perfectly linear slope as measured by a 1st order estimate would evidence no significant linear contrast (2nd order effect).

Curvatures are specific to each factor in the cohort analysis. The Holford approach produces estimates of curvatures of the age-group, period, and cohort coefficients by controlling for the linear trends of age, period, and cohorts and using orthogonal polynomials to estimate the deviations of the individual age, period, and cohort effects from linearity (O'Brien, 2010). In the Holford approach, the 1st order estimates used to derive the 2nd order functions are not interpreted. The development of the Holford approach was sparked by the recognition that the same curvature estimates will emerge regardless of the particular constraint chosen for model identification (i.e., the 2nd order results are constraint-invariant).

This approach can be explained more formally as follows. The Holford originated in epidemiology. In epidemiology, usually Poisson regression modeling is used to estimate the age, period, and cohort effects with the assumptions that the outcome variable follows a Poisson distribution and is a multiplicative function of the included model parameters, making the logarithm of the rates an additive function of the parameters (Tabeau, 2001). The model then becomes:

$$\ln(Y_{ij}) = \mu + \alpha_i + \eta_j + \gamma_k \qquad (7)$$

Where:

- Y_{ij} denotes the age–period-specific value on the outcome variable for the i[th] age group for $i = 1, \dots, a$ age groups at the j[th] time period for $j = 1, \dots, p$ time periods;
- μ denotes the intercept;
- α_i denotes the i[th] row age effect or the coefficient for the i[th] age group;

- η_j denotes the j[th] column period effect or the coefficient for the j[th] time period; and
- γ_k denotes the k[th] diagonal cohort effect or the coefficient for the k[th] cohort for $k = 1, \dots, (a + p - 1)$, with $k = a - i + j$;

This model can be reparametrized by re-expressing each effect in the model as a deviation from the mean of all effects of that type so that (also remember the step of going from equation 2 to equation 3):

$$\sum_i a_i = \sum_j \eta_j = \sum_k \gamma_k = 0 \qquad (8)$$

To get around the model identification problem Holford (1983) proposes to use a generalized inverse to solve the set of normal equations that provide maximum likelihood estimators. However, Holford notes that although the particular generalized inverse does not influence the significance test for parameters, the arbitrary selection of an inverse does have an effect on the parameters themselves. This means that different researchers using the same set of data can come up with different estimates of the age, period, and cohort effects. This is the reason that Holford considers estimable functions of the effects, which are invariant to the particular generalized inverse selected and hence do not depend on the particular constraint used.

The method proposed is to describe the trend of the parameters in two components: linear trend and curvature (or deviations from linearity). When the factor 'age' is represented by the effects α_i, the linear trend can be described by the contrast:

$$\alpha_L = C \sum_i c_i \alpha_i \quad (9)$$

Where:

- $c_i = i - \frac{a+1}{2} = i - \frac{1}{2}a - \frac{1}{2}$ for $i = 1, \dots, a$ age groups; and
- $C = (\sum_i c_i^2)^{-1}$

Then the curvature component is given by the age effects with the linear trend removed:

$$\tilde{\alpha}_i = \alpha_i - c_i \alpha_L \quad (10)$$

Following the notation of Holford (1983) these two components are parameterized in the columns of the design matrix, $\boldsymbol{X}$, as follows. Linear age is represented by $A_L(i) = c_i$, curvature by $A_{Cl}(i)$ ($l = 1, \dots, a - 2$) and the A_{Cl} are orthogonal to the A_L (i.e. $\sum_i A_L(i)\, A_{Cl}(i) = 0$). The curvature component are found by using second- and higher-order orthogonal polynomials (however, alternative methods exist and are listed at the end of this section).

The curvature parameters are given by:

$$\tilde{\alpha}_i = \sum_l A_{Cl}(i) \alpha_{Cl} \qquad (11)$$

Where:

- α_{Cl} represents the parameter associated with the column $A_{Cl}(\cdot)$ of the design matrix.

In a similar manner the columns for period and cohort effects are portioned by using $\boldsymbol{P_L}$,$\boldsymbol{P_C}$, and $\boldsymbol{C_L}$, $\boldsymbol{C_C}$ respectively; which yield the parameters η_L, η_C, γ_L, γ_C, respectively.

The design matrix can then be written as follows:

$$\boldsymbol{X} = (\mathbf{1}\ \boldsymbol{A_L P_L C_L A_C P_C C_C}) \qquad (12)$$

Parameters corresponding to this design matrix are $b' = (\mu, \alpha_C', \eta_C', \gamma_C', \alpha_L, \eta_L, \gamma_L)$.

This matrix is not of full column rank, because

$$\boldsymbol{C_L} = \boldsymbol{P_L} - \boldsymbol{A_L} \quad (13)$$

This is the point where the generalized inverse is used to solve the identification problem.

The main disadvantage of models based on the Holford approach is the inability to determine the overall linear component of trend, because this component determines the overall direction of the particular temporal trend. Yet, several interesting uses remain. As noted, curvature means any departure from a linear trend. This departure has been effectively used in several ways (Holford, 2004):

1. Second differences: Differences in the trends for two adjacent time points;
2. Change in slope: Trend changes over longer periods;
3. Polynomial trends: Temporal effects represented by including integer powers (the coefficient for the 1st order term is not estimable, whereas all higher order terms are); and
4. Splines: Sometimes preferred for representing curves.

3.4.4 Intrinsic Estimator Approach

The Intrinsic Estimator (IE) approach, a 1st order constraint based method, is based on recent developments in cohort methodology in biostatistics that have emphasized the utility of estimable functions, which are invariant to the selection of constraints on the parameters (Kupper, et al., 1985; Yang, 2005, 2008; Yang, et al., 2004). The IE approach considers an orthogonal decomposition of the parameter space into a null space for the singular design matrix and a non-null space, where the intrinsic estimator is obtained by the Moore-Penrose generalized inverse[13] (Yang, 2008).

The IE approach is similar to the constrained-based methods, such as the Mason, Mason, Winsborough, and Poole method, in the sense that it is designed to estimate each of the age effect, period effect, and cohort effect coefficients, not just their deviations from linearity or the unique variance accounted for by each of these sets of dummy variables (O'Brien, 2010). Moreover, the IE approach adopts a similar method as the 1st order effects based methods in the sense that it is based on placing a linear restriction on the column of the rank deficient design-matrix $\boldsymbol{X}$. However, this method is different in the sense that the constraint is associated with a generalized inverse. As opposed to the Holford approach, the generalized inverse provides a unique solution to the age, period, and cohort coefficient given that constraint.

[13] A generalized inverse of a given matrix $\boldsymbol{A}$ is a matrix $\boldsymbol{X}$ associated in some way with $\boldsymbol{A}$ that (Ben-Israel & Greville, 2003):

1. Exists for a class of matrices larger than the class of nonsingular matrices;
2. Has some of the properties of the usual inverse; and
3. Reduces to the usual inverse when $\boldsymbol{A}$ is nonsingular.

The O'Brien explanation of the IE approach

The rationale behind the IE approach can be explained as follows. If we label the number of columns of X in equation (3) as $m = [2(a + p) - 3]$, then it was shown that an identification problem arises because only $m - 1$ of these columns are linearly independent. In matrix notation this can be expressed as follows (O'Brien, 2010):

$$b_1\begin{bmatrix}1\\1\\\vdots\\1\\1\end{bmatrix} + b_2\begin{bmatrix}x\\x\\\vdots\\x\\x\end{bmatrix} + b_3\begin{bmatrix}x\\x\\\vdots\\x\\x\end{bmatrix} + \cdots + b_{m-1}\begin{bmatrix}x\\x\\\vdots\\x\\x\end{bmatrix} + b_m\begin{bmatrix}x\\x\\\vdots\\x\\x\end{bmatrix} = \begin{bmatrix}0\\0\\\vdots\\0\\0\end{bmatrix} \quad (14)$$

Where:

- b_i denotes the partial regression coefficients that, when they are multiplied with the column vectors, produce the zero-vector on the right hand side of the equal sign for $i = 1,2, \dots, m$.

O'Brien (2010) notes that the fact that such a vector of b's exists (amongst which are non-zero b's) shows that the columns of $\boldsymbol{X}$ are linearly dependent. In general, one and only one such vector of b's exists for the cohort model when no special constraints are placed on the model, which is unique up to multiplication by a scalar (O'Brien, 2010). In linear algebra this means that there is a nontrivial solution to the $a * p$ homogeneous equations, and consequently:

$$\boldsymbol{Xv} = 0 \qquad (15)$$

Where:

- $\boldsymbol{X}$ is the $ap * 2(a + p) - 3$ design matrix; and
- $\boldsymbol{v}$ is a $2(a + p) - 3 * 1$ vector containing the b's of equation 14.

$\boldsymbol{v}$ is said to be in the null space of $\boldsymbol{X}$ and is labeled as the null vector. It is the linear combination of columns of $\boldsymbol{X}$ that result in the zero-vector. That there is only one such vector indicates that the rank of the $\boldsymbol{X}$ matrix is just one less than full column rank and that a single linear constraint should allow for a solution to the identification problem (O'Brien, 2010). Thus, to find a unique solution to the individual dummy variables, a linear constraint must be chosen. Each such constraint is associated with a generalized inverse that allows for a solution given the specified linear constraint.

Following the notation of O'Brien the symbol $\boldsymbol{G_c}$ is used to represent the generalized inverse associated with a particular constraint, where $\boldsymbol{G_c} = (\boldsymbol{X^T X})_c^-$. If $\boldsymbol{X^T Y}$ in equation 4 is multiplied by the generalized inverse, the following solution is obtained: $\widehat{\boldsymbol{b}}_c = \boldsymbol{G_c X^T Y}$ where $\boldsymbol{G_c}$ is the generalized inverse associated with the constraint and $\widehat{\boldsymbol{b}}_c$ is the vector of parameter estimates for the equation given the constraint.

Under Mason, Mason, Winsborough, and Poole based methods, the constraint typically involves setting two of the coefficients associated with age, period, or cohort to be equal (or to zero) and the choice of this constraint determines a unique generalized inverse that is used to solve for the age-group, period, and cohort coefficients (O'Brien, 2010). The assumption is that $\boldsymbol{c^T b} = 0$, where $\boldsymbol{c}$ is the $m * 1$ vector for the constraint and $\boldsymbol{b}$ is the $m * 1$ vector of population effect coefficients. To the extent that $\boldsymbol{c^T b} = 0$ is not true, the estimates will be biased (O'Brien, 2010).

In the case of the IE approach the constraint involves $\boldsymbol{v}$ (an $m * 1$ vector), with the assumption that $\boldsymbol{v^T b} = 0$. Specifically, $\boldsymbol{v}$ is the null vector (the vector of coefficients that when multiplied with the columns of $\boldsymbol{X}$ results in the zero vector). To the extent that this assumption is not true, the estimates associated with this constraint will be biased. The generalized inverse that is associated with this is the constraint used by Yang, Schulhofer-Wohl, Fu and Land (2008): the Moore-Penrose generalized inverse. In the next paragraph the IE approach is explained algebraically.

Algebraic and Geometric Representation of the IE Approach

The general idea of the IE approach is explained algebraically by Yang et al. (2008)[14]. It has been shown that each of the infinite estimators of the parameter vector of equation 4 denoted as $\hat{b}$, can be decomposed into the direct sum of two linear subspaces that are orthogonal/independent to each other in the parameter space and written as (Yang, 2005; Yang, et al., 2004):

$$\widehat{\boldsymbol{b}} = \boldsymbol{B} + t\boldsymbol{B_0} \quad (16)$$

Where:

- t denotes a scalar corresponding to a specific solution;
- $\boldsymbol{B_0}$ denotes the null subspace of the eigenvector and is corresponding to the (unique) zero eigenvalue of the matrix $\boldsymbol{X^T X}$ of equation 4; and
- $\boldsymbol{B}$ denotes the non-null subspace that is the complement subspace orthogonal to the null space, i.e,:

$$\boldsymbol{B} = \left(\boldsymbol{I} - \boldsymbol{B_0 B_0^T}\right)\widehat{\boldsymbol{b}} \quad (17)$$

$t\boldsymbol{B_0}$ is in the null space of the design matrix $\boldsymbol{X}$ and represents trends of linear constraints – different equality constraints used by estimators, such as b_1 and b_2, yield different values of t. This can be represented geometrically as in Figure 8. The special parameter vector $\boldsymbol{B_0}$ corresponding to $t = 0$ satisfies the geometric projection. It is this special parameter vector that the IE estimates.

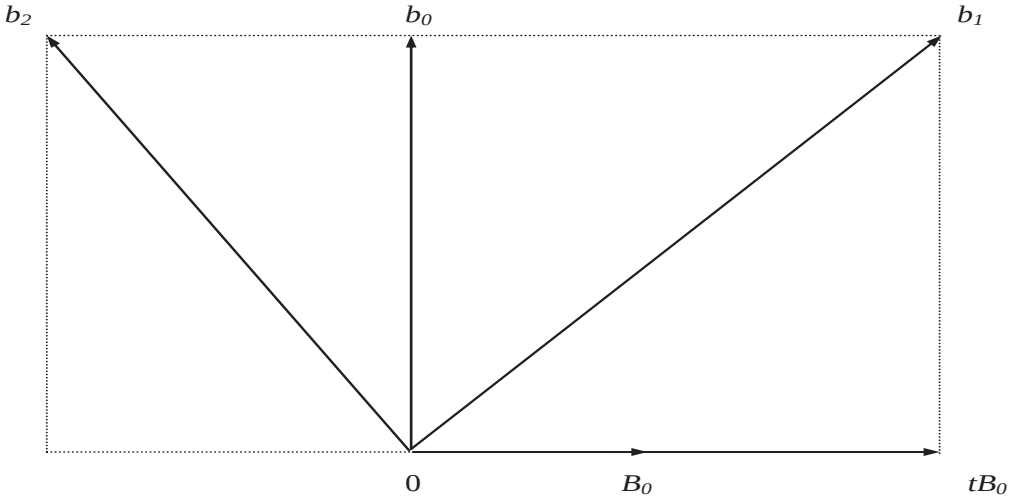


Figure 8: Geometric Representation of the Parameter Vector Orthogonal Decomposition (Land, 2011).

[14] For a geometric representation of the IE approach see Yang et al. (2008).

This single null vector reflects the fact that the design-matrix has a rank that is just one less than full column rank.

$\boldsymbol{B_0}$ does not depend on the observed outcomes $\boldsymbol{Y}$, but is fixed by the design matrix $\boldsymbol{X}$. Subsequently with equation 16, any cohort estimator obtained by placing any identifying constraint(s) on the design matrix can be written as the linear combination $\boldsymbol{B} + t\boldsymbol{B_0}$, where $\boldsymbol{B}$ is the special estimator called the intrinsic estimator that lies in the parameter subspace that is orthogonal to the null space and is determined by the Moore-Penrose generalized inverse (a.k.a. the pseudo-inverse)[15].

In vector space terminology, the identifying constraint imposed by the IE approach to estimate the model amounts to a constraint on the orientation of the parameter vector (in equation 4) in the parameter space. That is, it constrains the design matrix to have zero influence on the estimated coefficient (i.e., by setting $t = 0$) (Yang, 2005).

The generalized inverse associated with the IE estimator, the Moore-Penrose generalized inverse, has some properties that makes it a suitable choice for a generalized inverse in the absence of other information (estimability, finite time period unbiasedness, relative efficiency, asymptotic consistency[16]) (Yang, et al., 2004; Yang, et al., 2008). The Moore-Penrose generalized inverse is seen as a good choice of a generalized inverse when there are no theoretically or empirically compelling reasons to use a different generalized inverse.

Since this model is a recent development in the cohort analysis literature, it has not yet received much feedback. However, O'Brien (2010) points out that, if for some reason a researcher is able to determine the constraints beforehand (i.e., on the basis of theory); then the choice of a Mason, Mason, Winsborough and Poole based model might well be preferable, since the properties do not protect the solution it produces from biased estimates of the data generating parameters. This is so, because whether a Mason,

---

[15] Penrose showed that for every finite matrix $\boldsymbol{A}$ (square or rectangular) of real or complex elements, there is a unique matrix $\boldsymbol{X}$ satisfying the four equations (Ben-Israel & Greville, 2003; Seber, 2008):

1. $\boldsymbol{AXA} = \boldsymbol{A}$ (18);
2. $\boldsymbol{XAX} = \boldsymbol{X}$ (19);
3. $(\boldsymbol{AX})^T = \boldsymbol{AX}$ (20), i.e., the matrix $\boldsymbol{AX}$ is symmetric; and
4. $(\boldsymbol{XA})^* = \boldsymbol{XA}$ (21), i.e., the matrix $\boldsymbol{XA}$ is symmetric.

Where $\boldsymbol{A}^*$ denotes the conjugate transpose of $\boldsymbol{A}$. If $\boldsymbol{A}$ is nonsingular, then $\boldsymbol{X} = \boldsymbol{A}^{-1}$ trivially satisfies the four equations.

[16] These properties can be described as follows (Land, 2011):

1. **Estimability:** Yang et al. (2004) established that the IE satisfies the Kupper et al. (1985) condition for estimability, namely $l^T B_0 = 0$ where l^T is a constraint vector (of appropriate dimension) that defines a linear function $l^T b$ of b. Estimable functions are desirable as statistical estimators because they are linear functions of the unidentified parameter vector that can be estimated without bias, i.e., they have unbiased estimators.
2. **Unbiasedness*:*** For a fixed number of time periods of data, the IE is an unbiased estimator of the special parameterization (or linear function) b_0 of b.
3. **Relative efficiency*:*** For a fixed number of time periods of data, it has a smaller variance than any other generalized linear estimator.
4. **Asymptotic consistency:** Derived from the fact that the length of the eigenvector B_0 decreases with increasing numbers of time periods of data, and, converges to zero as the number of periods of data increases without bound. Which means that estimators converge toward the IE B as the number of periods increase

Mason, Winsborough and Poole based model provides estimates that are correct in terms of the "generating process," depends on whether the constraint is consistent or inconsistent with that process.

3.4 Summary

This chapter provided an answer to the second sub-question of this thesis: What is cohort analysis? Cohort analysis was introduced and developments in this field discussed.

Cohort analysis was defined as an analysis technique in which statistical attempts are made to partition (variance in) the outcome on an independent variable into the unique components attributable to age, period, and cohort effects.

A distinction between the sociological and epidemiologic conceptualizations of cohorts can be made, since they result in a fundamentally different execution of cohort analysis. The sociological definition assumes that cohorts have unique characteristics confounded by age and period effects and, therefore, can exist independently of age and period effects, while the epidemiologic definition assumes that period and age effects interact to produce cohort effects.

A plethora of models to conduct cohort analysis were identified in the literature. That there are this many models can be attributed to the model identification problem. In cohort analysis the multiple correlation of each independent variable with the other ones is unity, i.e., perfect collinearity exists. Therefore, standard generalized linear models such as multiple regression are unable to identify the regression coefficients.

The models available in the literature were categorized in Table 3 in section 3.4 according to how they dealt with the age, period, and cohort parameters. To that end the distinction between models that estimate 1^{st} order effects, 2^{nd} order effects, and hybrid models were made. 1^{st} order effects have a linear relationship with the outcome of interest. 2^{nd} order effects are those which have a non-linear relationship with the outcome of interest.

Four models were discussed in more detail. First, the Mason, Mason, Winsborough, and Poole method, a 1^{st} order constraint-based method rooted in sociology. The approach makes use of a multiple classification model. The model specifies the dependent variable to be the result of effect parameters associated with particular levels of each independent variable. At least one category of age, period, and cohort is constrained in by setting it equal to another category of the same dimension. The main criticism to this approach is that the linear dependence is broken in the statistical model only and not in the real world. Therefore, the obtained results might be meaningless.

Second, the MPT, a 2^{nd} order model was discussed. MPT estimates cohort effects as partial interaction (2^{nd} order effect) of age and period effects. Thereby, it adheres to the epidemiologic definition of cohorts. The MPT explicitly tests whether the effects of age and period interact to produce an effect that is more than what would be expected given their additional influences. Since the MPT estimates an age and period model no constraints are necessary. The residuals are then regressed on indicator variables for cohort membership; the extent to which the cohort variable predicts the residual is the cohort effect. However, results from the MPT are non-unique.

Third, the Holford approach is discussed, a hybrid approach between 1^{st} and 2^{nd} order models, which is widely used in epidemiology. Holford proposes to use a generalized inverse to get around the model

identification problem. However, Holford notes that the particular generalized inverse used has an effect on the estimates of the age, period, and cohort effects. Therefore, Holford focuses on linear deviations known as curvatures, a measure which can be interpreted as reflecting changes in the direction or steepness of the slope of the underlying age, period, and cohort effects (2^{nd} order effects) without estimating the magnitude of the actual slope (1^{st} order estimate). The advantage of this approach is that the 2^{nd} order results are constraint-invariant and, therefore, a unique solution is obtained. The main disadvantage is the inability to determine the overall linear component of the trend.

Finally, the intrinsic estimator approach is discussed. The intrinsic estimator approach is a recent development in 1^{st} order constraint based methods which was developed by researchers with backgrounds in biostatistics and sociology. Based on estimable functions, which are invariant to the selection of constraints on the parameters, and on the singular value decomposition of matrices via the Moore-Penrose generalized inverse, the intrinsic estimator yields robust estimates of trends by age, period, and cohort and uniquely determines the coefficient estimates. The Moore-Penrose generalized inverse has some properties that result in robust estimates. The main disadvantage of this approach is that, since it is a recently developed approach it has not yet received much feedback in the academic literature as the other models have.

The next chapter, chapter 4, will discuss which developments in cohort analysis are relevant for vintage analysis.

4 Cohort Analysis' Relevance for Vintage Analysis

> *"The age-period-cohort effect identification problem arises because analysts want something for nothing: a general statistical decomposition of data without specific subject matter motivation underlying the decomposition. In a sense it is a blessing for social science that a purely statistical approach to the problem is bound to fail."*
>
> Heckman and Robb (1985, pp. 144-145)

Chapter 2 discussed the role of vintage analysis in structured credit analysis. Chapter 3 explained what cohort analysis entails. This chapter forms a synthesis of chapters 2 and 3 and, thereby, provides an answer to the third sub-question of this thesis: What developments in cohort analysis are relevant for vintage analysis?

The current chapter is structured as follows. In section 4.1 vintage analysis is viewed from a cohort analysis perspective. Section 4.2 builds on this view and explains the usefulness of a cohort analysis based vintage analysis. Finally, section 4.3 ends the chapter with a discussion regarding the developments in cohort analysis that are relevant for vintage analysis.

4.1 Vintage analysis from a Cohort Analysis Perspective

Two major parties in the securitization chain that were identified in chapter 2 are the credit rating agencies (CRAs) and the structured credit investors (SCIs). CRAs predict the cash flows and associated risk of a structured credit product so that a tranche can be rated; SCIs use the same information, but their goal is to value the tranche. The quantitative side of the rating and valuation of a structured credit tranche is conducted via a model that takes into account three components of a structured credit tranche: the collateral, the credit enhancement, and the cash flow mechanics. Analysis of the latter two components was explained to be a relatively straightforward interpretation of the deal documentation of the structured credit product. Analysis of the first component, the collateral, should result in assumptions about the default rate, prepayment rate, and the loss severity that serve as inputs for the rating and valuation models.

In section 2.3 it was explained that the analysis of the collateral is where vintage analysis plays a role. The process of monitoring groups of loans and comparing performance across past groups was defined to be the task of vintage analysis. In one way or another, structured credit market participants have always concerned themselves with the measurement of maturation, extrinsic, and origination effects. Both CRAs and SCIs recognize that changing underwriting standards, extrinsic effects and aging have an impact on the shape of vintage trajectories and thus the performance of structured credit. This notion can be explained further by considering the data behind the graph of Figure 7 in chapter 2. This graph was based on data of which a small part is shown in Table 4. This table is identical to the standard cohort table introduced in section 3.3.1 constructed by cross-sectional data sets juxtaposing the relationship between age and some dependent variable, in this case the loss-rate on auto loans, with the age intervals equal to the intervals between periods for which there are data, in this case monthly intervals.

Table 4

Loss-Rates of All German Auto Loans Issued by Volkswagen Bank GmbH Between August 2008 and March 2011. Data source: RBS/UniCredit (2011) and author's calculations.

Age/ Period	08-08	09-08	10-08	11-08	12-08	01-09	02-09	03-09	04-09	05-09	06-09	07-09	08-09	09-09	10-09	11-09	12-09	01-10	02-10	03-10	04-10	05-10	06-10	07-10	08-10	09-10	10-10	11-10	12-10	01-11	02-11	03-11
1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%		
4	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,02%	0,00%	0,01%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%			
5	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,02%	0,00%	0,01%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%				
6	0,00%	0,01%	0,01%	0,01%	0,00%	0,00%	0,02%	0,00%	0,01%	0,01%	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%	0,01%	0,00%	0,00%	0,00%	0,01%	0,00%	0,01%	0,00%	0,00%	0,00%	0,00%					
7	0,00%	0,01%	0,02%	0,01%	0,00%	0,02%	0,03%	0,01%	0,01%	0,01%	0,01%	0,01%	0,01%	0,01%	0,00%	0,01%	0,01%	0,02%	0,00%	0,00%	0,01%	0,00%	0,01%	0,01%	0,01%	0,01%						
8	0,00%	0,01%	0,02%	0,02%	0,00%	0,02%	0,03%	0,02%	0,01%	0,01%	0,01%	0,02%	0,01%	0,01%	0,02%	0,01%	0,02%	0,03%	0,02%	0,00%	0,01%	0,01%	0,01%	0,01%	0,01%							
9	0,01%	0,02%	0,03%	0,03%	0,01%	0,02%	0,03%	0,02%	0,03%	0,03%	0,01%	0,03%	0,02%	0,02%	0,03%	0,02%	0,02%	0,03%	0,02%	0,00%	0,02%	0,01%	0,01%	0,02%								
10	0,03%	0,04%	0,03%	0,04%	0,03%	0,05%	0,04%	0,03%	0,04%	0,03%	0,01%	0,03%	0,02%	0,02%	0,04%	0,02%	0,04%	0,04%	0,03%	0,01%	0,02%	0,01%	0,02%									
11	0,05%	0,06%	0,04%	0,06%	0,05%	0,08%	0,05%	0,04%	0,05%	0,04%	0,02%	0,04%	0,03%	0,03%	0,04%	0,04%	0,05%	0,05%	0,03%	0,01%	0,03%	0,02%										
12	0,06%	0,06%	0,06%	0,08%	0,06%	0,09%	0,06%	0,04%	0,07%	0,05%	0,03%	0,05%	0,03%	0,06%	0,05%	0,06%	0,06%	0,05%	0,05%	0,03%	0,05%											
13	0,08%	0,07%	0,07%	0,11%	0,07%	0,11%	0,08%	0,05%	0,07%	0,05%	0,04%	0,05%	0,04%	0,09%	0,06%	0,08%	0,08%	0,07%	0,07%	0,05%												
14	0,10%	0,09%	0,11%	0,13%	0,09%	0,14%	0,08%	0,07%	0,08%	0,06%	0,05%	0,07%	0,06%	0,09%	0,07%	0,11%	0,09%	0,08%	0,08%													
15	0,13%	0,12%	0,16%	0,17%	0,10%	0,15%	0,10%	0,08%	0,10%	0,08%	0,07%	0,10%	0,08%	0,11%	0,10%	0,12%	0,11%	0,10%														
16	0,17%	0,13%	0,20%	0,19%	0,14%	0,19%	0,12%	0,09%	0,12%	0,10%	0,08%	0,11%	0,10%	0,13%	0,11%	0,13%	0,13%															
17	0,21%	0,16%	0,23%	0,22%	0,17%	0,22%	0,14%	0,12%	0,14%	0,12%	0,09%	0,13%	0,11%	0,14%	0,12%	0,14%																
18	0,26%	0,22%	0,26%	0,25%	0,21%	0,25%	0,14%	0,14%	0,17%	0,13%	0,11%	0,15%	0,13%	0,16%	0,15%																	
19	0,30%	0,25%	0,29%	0,28%	0,27%	0,29%	0,16%	0,15%	0,19%	0,16%	0,14%	0,17%	0,15%	0,17%																		
20	0,32%	0,28%	0,32%	0,32%	0,30%	0,32%	0,19%	0,16%	0,22%	0,18%	0,15%	0,19%	0,16%																			
21	0,34%	0,32%	0,36%	0,38%	0,35%	0,34%	0,22%	0,18%	0,24%	0,20%	0,16%	0,22%																				
22	0,36%	0,36%	0,39%	0,41%	0,37%	0,38%	0,25%	0,21%	0,26%	0,22%	0,18%																					
23	0,42%	0,39%	0,41%	0,43%	0,41%	0,43%	0,29%	0,22%	0,28%	0,25%																						
24	0,47%	0,42%	0,43%	0,48%	0,47%	0,48%	0,31%	0,23%	0,30%																							
25	0,49%	0,46%	0,47%	0,55%	0,50%	0,50%	0,33%	0,26%																								
26	0,52%	0,49%	0,53%	0,57%	0,51%	0,51%	0,35%																									
27	0,55%	0,55%	0,57%	0,59%	0,52%	0,54%																										
28	0,62%	0,61%	0,58%	0,62%	0,54%																											
29	0,66%	0,64%	0,60%	0,68%																												
30	0,68%	0,67%	0,63%																													
31	0,70%	0,69%																														
32	0,74%																															

Viewing vintage analysis from this perspective shows that analysts deal with the following problem: How to distinguish between the effect of changing underwriting standards over time, extrinsic effects – such as changing macro-economic conditions – and the effect of aging?

In Table 4 these three effects are captured by two distinguishable time dimensions: the age dimension and the calendar time dimension. As was described in chapter 3, cohort analysis deals with this dual-time nature of vintages. In chapter 3 cohort analysis was defined as an analysis technique in which statistical attempts are made to partition (variance in) the outcome on an independent variable into the unique components attributable to age, period, and cohort effects. These three effects can be translated to structured credit effects. First, there is the effect of aging; this is the equivalent of what is termed the *maturation effect* in social sciences (de Vaus, 2001). Besides maturation, *extrinsic effects* that are a result of changing macro-economic conditions, changing competition, or changes in law, influence the trajectory of vintages. Finally, there also is an *origination effect* in credit markets that is a consequence of changing underwriting standards that result in differences in vintage performance (Saunders & Allen, 2010).

A vintage analysis that takes into account maturation, extrinsic, and origination effects and uses cohort analysis based models to separate these effects is termed *cohort analysis based vintage analysis* in the rest of this thesis.

4.2 Usefulness of Cohort Analysis Based Vintage Analysis

Viewing vintage analysis from a cohort analysis perspective allows analysts to draw on eight decades of cohort analysis research. An extensive body of research in social science that was discussed in chapter 3 investigates how to identify and estimate age, period, and cohort effects. Cohort analysis based vintage analysis provides the tools to model and quantify the maturation, extrinsic, and origination effects on structured credit performance. Two aspects of unraveling maturation, extrinsic, and origination effects on structured credit performance are of interest to structured credit market participants. First the quantification of maturation, extrinsic, and origination effects allows them to better understand the historic performance of structured credit. Second, the quantification allows them to forecast future trends in performance rates more accurately.

Structured credit market participants have a pragmatic and non-theoretical reason for considering maturation and aging effects, which is their need for forecasting future trends in performance rates. If, especially over the short-term, either maturation or aging effects predominate, then one avenue for providing projections is to model the observed trend in these effects and continue it into the future. Extrapolation of maturation effects is relatively more accurate than extrapolation of extrinsic and origination effects due to the self-maturation nature of loans.

In chapter 3 a bottom-up perspective was adopted to identify the various cohort analysis models available in the literature. The models were categorized according to whether they estimated 1^{st} order effects, 2^{nd} order effects, or a combination of these effects. In section 3.4 four cohort analysis models were studied in more detail. Since the goal of this section is to identify the relevant developments in cohort analysis for vintage analysis a top-down perspective is adopted.

With exception of the MPT, the general structure of the cohort analysis models that were studied in more detail in chapter 3 were based on the multiple classification model of Mason et al. (1973) which was represented as:

$$Y_{ij} = \mu + \alpha_i + \eta_j + \gamma_k + \varepsilon_{ij} \qquad (22)$$

Around the time that Mason et al. published their article on the multiple classification model, Nelder and Wedderburn (1972) published their theory on generalized linear models (GLMs). GLMs extend traditional linear models, such as the multiple classification models in their original form, so that a linear predictor is mapped through a link function to the mean of a response characterized by any member of the exponential family of distributions and by allowing the magnitude of the variance of each measurement to be a function of its predicted value (Nelder & Wedderburn, 1972). The generalization admits a model specification allowing for continuous or discrete outcomes and allows for a description of the variance of the mean.

The implication of the GLM theory for the multiple classification model was that the observations of the response variable were not restricted to be characterized by the normal or Gaussian distribution anymore, nor that the distributions for all observations was restricted to have a common variance σ^2.

The response variable of the multiple classification model, Y_{ij}, was generalized in the following sense:

$$Y_{ij} = f(\hat{R}_{ij}) = f\left(\frac{O_{ij}}{N_{ij}}\right) \quad (23)$$

Where:

- Y_{ij} denotes the age–period-specific value on the outcome variable for the i[th] age group for $i = 1, \dots, a$ age groups at the j[th] time period for $j = 1, \dots, p$ time periods;
- $f(\hat{R}_{ij}) = f\left(\frac{O_{ij}}{N_{ij}}\right)$ represents some function of the observed rate $\hat{R}_{ij}$;
- $\hat{R}_{ij}$ denotes the observed rate of the i^{th} age group in the j^{th} period;
- O_{ij} denotes the observed number of cases (deaths, illnesses, etcetera in epidemiology, and arrears, losses, prepayments, etcetera in structured credit) of the i^{th} age group in the j^{th} period; and
- N_{ij} denotes the amount of risk time (number of person- or loan-years at risk) of the i^{th} age group in the j^{th} period.

In GLMs, the only random component of the multiple classification model, the error term, ε_{ij}, is still assumed to have mean (or expected value) $E(\varepsilon_{ij}) = 0$, as in the article of Mason et al. (1973). However, in GLMs the variance and other distributional properties of ε_{ij}, are tied to the assumption made about the stochastic nature of Y_{ij} and hence of $\hat{R}_{ij}$ (Kupper, et al., 1985).

As was noted GLMs allow the response variable to be characterized by any member of the exponential family of distributions. Numerous choices for $f(\hat{R}_{ij})$ have been considered in cohort analysis literature, some of the more popular ones being $f(\hat{R}_{ij}) = \hat{R}_{ij}$, $f(\hat{R}_{ij}) = \ln(\hat{R}_{ij})$ for $\hat{R}_{ij} > 0$, $f(\hat{R}_{ij}) = \ln(1 + \hat{R}_{ij})$ to permit consideration of observed rates equal to zero, and the logit transformation $f(\hat{R}_{ij}) =$

$\ln[\hat{R}_{ij}/(1-\hat{R}_{ij})]$ for proportions, i.e., rates that have been suitably scaled to lie between 0 and 1 in value (Kupper, et al., 1985). Which choice for $f(\hat{R}_{ij})$ is suitable depends on the problem at hand.

In spite of cohort analysis' theoretical merits and conceptual relevance, it is noted in chapter 3 that generalized linear cohort models have their own problem. If calendar time is denoted by P, each vintage denoted by its origination time C, then the age, A (or months-on-book for monthly vintages), is calculated by $A = P - C \; \forall \; P > C$. This perfect linear relationship gives rises to the model identification problem of cohort analysis that was discussed extensively in chapter 3. The consequence of the model identification problem is that the maturation, extrinsic, and origination effects cannot be directly estimated by the generalized linear models that were introduced in chapter 3.

Besides the models based on generalized linear models, which have to deal with the model identification problem, three other approaches were identified in cohort analysis literature in Table 3 of chapter 3.4:

1. The Median Polish Technique (MPT) in which the age and period effects are identified via an iterative procedure (the median polish) and the cohort effects quantified by regression of the median polish residuals on cohort categories was introduced in section 3.4.2;
2. The proxy variable or Age-Period-Cohort Characteristic (APCC) models that use one or more proxy variables to replace age, period, or cohort coefficients; and
3. Nonlinear parametric (algebraic) transformation approaches that define a nonlinear parametric function of one of the age, period, or cohort variables so that its relationship to others is nonlinear.

4.3 Discussion

The purpose of this chapter was to provide an answer to the third sub-question of this thesis: What developments in cohort analysis are relevant for vintage analysis?

Section 4.1 showed that when vintage analysis is viewed as a problem of how to unravel maturation, extrinsic, and origination effects on structured credit performance, cohort analysis can play are role in the quantification of these effects. The problem then becomes what model to unravel these effects.

Debate continues regarding the legitimacy of modeling assumptions of various statistical methods for cohort analysis. Cohort models based on generalized linear models are widely studied and are relatively easy to interpret, but a direct estimate of the three effects is not possible without assigning additional identifying constraints to the generalized linear model. The MPT produces non-unique solutions. The APCC models have the problem of identifying the "correct" proxy variables that adequately represent the effect to be measured. Nonlinear parametric transformations have the difficulty that it may not be evident what nonlinear function should be defined for the effects of age, period, or cohort. This is important, since in vintage analysis, as well as in cohort analysis, the analyst usually starts with either a tabula rasa or conflicting hypotheses.

A starting point that helps in this discussion was introduced in chapter 3. Models can be differentiated by their different conceptualizations of cohorts. This comes from the believe that cohort effects can either exist independently of age and period effects, or that cohort effects are the result of the interaction of age and period, respectively. These conceptualizations are as relevant for cohort analysis as for vintage analysis: before a origination effect can be modeled it has to be properly defined. Defining the origination

effect comes down to having a robust theory available regarding the conceptualizations between maturation, extrinsic, and origination effects.

Also, in chapter 3 cohort based models were categorized based on whether they estimate 1st order effects or 2nd order effects. 1st order models make use of constraints. The dozen of constraints that were identified in the literature (see Table 3 in section 3.4) each affect the results in their own way. Models that estimate 2nd order effects are constraint invariant.

The opening quotation of this chapter by Heckman and Robb summarizes the current problem of vintage based cohort analysis. The model identification problem was explained to be a lack of understanding of the theory behind the problem that is analyzed. The problem of the selection of a suitable model for cohort analysis based vintage analysis is this too. Theory is what structured credit is short of. Once the relationships between vintage, maturation, and extrinsic effects have been properly defined, cohort analysis can aid the vintage analysis process. Because then the analyst has a theoretical basis for why a certain model was chosen, why a certain restriction was applied, and as a result the output can be properly interpreted. In sum, the third sub-question can only be partly answered.

The next chapter adds to this discussion, and thereby to the third sub-question, by applying a cohort analysis model to a dataset of mortgages. Due to the lack of understanding of the relationship between maturation, extrinsic, and origination effects, a full blown vintage based cohort analysis in which input for the structured credit model via cohort based vintage analysis is acquired goes beyond the scope of this thesis. However, the exercise serves as a way to gain additional insight into the application of cohort analysis to structured credit data.

5 Loan Vintage Analysis Using Cohort Analysis Techniques: An Illustration

In this chapter a cohort model is applied to a loan tape[17] consisting of mortgage data. The purpose of this chapter is to gain insight into the application of cohort techniques to structured credit data. As was shown in chapter 4 any cohort analysis based vintage analysis should rely on theory. In the introduction of this thesis it was stated that the body of research regarding structured credit is limited and needs to be further developed. In chapter 4 it was concluded that theory regarding the conceptualization of what constitutes an origination effect is missing in the academic literature. Therefore, a properly conducted cohort analysis based vintage analysis in which input for the structured credit model via cohort based vintage analysis is acquired is beyond the scope of this thesis. However, an application of cohort analysis provides insight into the process of application of cohort models, the difficulties encountered during this process, and the differences between cohort analysis in epidemiology and sociology on the one hand, and vintage analysis in structured credit markets on the other.

This chapter is structured as follows. Section 5.1 describes the dataset that is analyzed. Section 5.2 describes the analysis of the dataset. Finally, section 5.3 concludes with a discussion of the lessons learned from the application of the cohort analysis model to the dataset.

5.1 The Aggregator of Loans Backed by Assets (ALBA) Loan Tapes

The dataset used in this thesis consists of monthly updated loan-level data from three United Kingdom non-conforming residential mortgage backed securities: ALBA 2006-1, ALBA 2006-2, and ALBA 2007-1. The timespan of the historic data is from the period June 2008 to May 2010. Background information regarding ALBA deals can be found in Appendix B.

The historical performance of the ALBA deals can be derived from the loan tapes. The quality of the loan tape data varies from issuer to issuer. Several standard data formats exist which provides some level of standardization, though proprietary formats persist. This poses two challenges for CRAs and SCIs: data merging (due to the different formats), and error checking (due to the quality differences). Besides data scrubbing and error checking, a due diligence process is usually part of the loan tape quality assessment[18].

The data breadth and quality of the ALBA loan tapes is relatively low. Figure 9 shows the structure of the ALBA loan tapes. The upper panel shows the structure of the master loan tape, which contains static data for approximately 16.000 mortgages. It contains loan-level data for few but the most important fields. The middle panel shows the structure of the ALBA 2007-1 history loan tape. This loan tape contains monthly dynamic historic data for each of the loans; more than 175.000 observations. Also here most data fields are empty. The bottom panel shows a summary of the ALBA 2007-1 history loan tape. Here we see another problem: comma's are missing throughout the loan tape. In order to restore the comma's, the loan

[17] The term loan tapes is a holdover from the days when paper loan files were photographed and the images stored on computer tapes. Nowadays, the loan tapes consist of (extracts from) databases that are holding information for each of the loans. Thus, a loan tape offers a précis of the relevant data for each loan that is part of the collateral of a securitization deal.

[18] Professional investors usually assess the quality of the loan tapes by analyzing loan pool samples to compare the information on the loan tape to information contained in the hard-copy documents. Next to that, the due diligence process usually involves an assessment of whether the issuer adhered to the underwriting standards and applicable laws and an assessment of the validity of the appraised value of the collateral.

tapes were compared with the three deal prospectus that helped to determine the correct position of the comma's.

Since the loan tapes contain such limited information, the performance indicator of the ALBA deals that is focused on, is the aggregate arrears rate. The dataset contains all the information necessary to compute the arrears rates for the mortgages. Whenever a borrower misses a payment on a loan, or does not make his required payment in full, he falls into arrears. Arrears on a mortgage loan indicate that the borrower is likely to be suffering some degree of financial stress and, as such, these loans have higher risk of going into default.

Table 5 shows historical data on arrears, from contracts originated between October 2005 and July 2008 , which missed a (partial) payment between June 2008 and May 2010, grouped by month of origination. The data structure of this dataset can defined as a vintage time series. The mortgages originated from the same month constitute a monthly vintage. The vintage time series refer to the longitudinal observations and performance measurements for the cohort of mortgages that share the same origination date, and therefore, the same age. The arrears rate are aligned in calendar time, while they can be also aligned in life time. The arrears rates data displayed, for each portfolio of mortgages originated in a particular month, are expressed as a percentage of the number of mortgages of the total portfolio. A mortgage is either in arrears or not, where arrears is defined as missing a (partial) payment.

```
> alba <- read.table( "Alba All Master.csv", sep=",", header=TRUE )
> str(alba)
'data.frame':   15991 obs. of  59 variables:
 $ M_PoolID                : Factor w/ 6 levels "ALBA 2006-1G",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ M_LoanID                : int  10015906 10017304 10018308 10019301 10020108 10021907 10024102 10024701 10025901 10027103 ...
 $ ClassType               : logi  NA NA NA NA NA NA ...
 $ LoanOriginator          : logi  NA NA NA NA NA NA ...
 $ OriginalPrincipalBalance: logi  NA NA NA NA NA NA ...
 $ OriginalFABalance       : logi  NA NA NA NA NA NA ...
 $ OriginalPropertyValue   : int  170000 345000 170000 230000 160000 75000 165000 180000 315000 105000 ...
 $ PropertyValuationDate   : Factor w/ 743 levels "2005-04-07","2005-04-13",..: 333 413 365 378 85 449 442 389 402 403 ...
 $ ValuationType           : logi  NA NA NA NA NA NA ...
 $ LoanPropertyRegion      : Factor w/ 11 levels "","East Anglia",..: 7 4 5 3 10 6 8 5 8 10 ...
 $ PropertyType            : logi  NA NA NA NA NA NA ...
 $ OriginalLTV             : logi  NA NA NA NA NA NA ...
 $ TermToMaturity          : int  240 276 288 168 192 288 288 204 288 348 ...
 $ LoanOriginationDate     : Factor w/ 328 levels "","2005-06-01",..: 119 34 21 34 31 28 57 33 38 33 ...
 $ MaturityDate            : logi  NA NA NA NA NA NA ...
 $ OriginalRateType        : logi  NA NA NA NA NA NA ...
 $ ReversionaryDate        : logi  NA NA NA NA NA NA ...
 $ ReversionaryRate        : logi  NA NA NA NA NA NA ...
 $ OriginalInterestRate    : logi  NA NA NA NA NA NA ...
 $ ReversionaryMargin      : logi  NA NA NA NA NA NA ...
 $ SelfCertification       : logi  NA NA NA NA NA NA ...
 $ EmploymentStatus        : logi  NA NA NA NA NA NA ...
 $ FirstTimeBuyer          : logi  NA NA NA NA NA NA ...
 $ IncomeSource1           : int  0 70000 30000 55000 40000 18500 45000 40000 75000 18000 ...
 $ IncomeSource2           : int  0 0 0 6000 0 10400 0 0 0 0 ...
 $ IncomeSource3           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ IncomeSource4           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ OtherIncome             : logi  NA NA NA NA NA NA ...
 $ BTLStatus               : logi  NA NA NA NA NA NA ...
 $ RighttoBuy              : logi  NA NA NA NA NA NA ...
 $ LoanPurpose             : logi  NA NA NA NA NA NA ...
 $ RepaymentType           : Factor w/ 3 levels "IO","Other","RE": 1 1 1 1 3 3 1 1 1 3 ...


> alba2 <- read.table( "ALBA 2007 History.csv", sep=",", header=TRUE )
> str(alba2)
'data.frame':   175189 obs. of  23 variables:
 $ PoolID                  : Factor w/ 2 levels "ALBA 2007-1G",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ LoanID                  : int  570400 572604 590303 597701 599502 648210 731702 744306 745909 746902 ...
 $ ReportDate              : Factor w/ 24 levels "2008-06-30","2008-07-31",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ CurrentPrincipalBalance: int  7483856 0 0 6526576 12664841 12213579 7322259 0 6109482 17184133 ...
 $ CurrentLTV              : logi  NA NA NA NA NA NA ...
 $ RemainingTerm           : logi  NA NA NA NA NA NA ...
 $ CurrentRateType         : logi  NA NA NA NA NA NA ...
 $ CurrentBaseRate         : logi  NA NA NA NA NA NA ...
 $ CurrentMargin           : logi  NA NA NA NA NA NA ...
 $ CurrentInterestRate     : logi  NA NA NA NA NA NA ...
 $ MortgageServiceAmount   : int  51738 57654 49475 45329 80151 77676 47666 0 42988 92645 ...
 $ DelinquencyStatus       : logi  NA NA NA NA NA NA ...
 $ ArrearsBalance          : int  0 0 0 0 0 -214 0 0 0 0 ...
 $ Repossession            : logi  NA NA NA NA NA NA ...
 $ CurrentPropertyVal      : logi  NA NA NA NA NA NA ...
 $ CurrentTypeVal          : logi  NA NA NA NA NA NA ...
 $ CurrentDateVal          : logi  NA NA NA NA NA NA ...
 $ StatusAsDetermination   : Factor w/ 1 level "Live": 1 1 1 1 1 1 1 1 1 1 ...
 $ LoanAge                 : num  20.4 20.7 24.4 21.8 19 ...
 $ tempDefaulted           : logi  NA NA NA NA NA NA ...
 $ CurrentDTI              : logi  NA NA NA NA NA NA ...
 $ CurrentAVGValue         : logi  NA NA NA NA NA NA ...
 $ audit_date              : Factor w/ 18685 levels "2010-07-28 23:33:29",..: 7248 7248 7248 7248 7248 7249 7249 7249 7249 7249 ...


> summary(alba2)
          PoolID           LoanID              ReportDate     CurrentPrincipalBalance CurrentLTV     RemainingTerm  CurrentRateType CurrentBaseRate
 ALBA 2007-1G : 62110   Min.   :  570400   2010-04-30:  7303   Min.   : -375596        Mode:logical   Mode:logical   Mode:logical    Mode:logical
 ALBA 2007-1G2:113079   1st Qu.:57295304   2010-05-31:  7303   1st Qu.: 3470345        NA's:175189    NA's:175189    NA's:175189     NA's:175189
                        Median :62058008   2010-02-28:  7302   Median :10157541
                        Mean   :61267352   2010-03-31:  7302   Mean   :10229192
                        3rd Qu.:66690304   2009-12-31:  7301   3rd Qu.:14949773
                        Max.   :99328605   2010-01-31:  7301   Max.   :70922359
                                           (Other)   :131377
 CurrentMargin  CurrentInterestRate MortgageServiceAmount DelinquencyStatus ArrearsBalance     Repossession   CurrentPropertyVal CurrentTypeVal
 Mode:logical   Mode:logical        Min.   :     0         Mode:logical      Min.   :-2436267   Mode:logical   Mode:logical       Mode:logical
 NA's:175189    NA's:175189         1st Qu.: 31550         NA's:175189       1st Qu.:       0   NA's:175189    NA's:175189        NA's:175189
                                    Median : 49672                           Median :       0
                                    Mean   : 56994                           Mean   :   43337
                                    3rd Qu.: 73984                           3rd Qu.:       0
                                    Max.   :430056                           Max.   : 4853048

 CurrentDateVal StatusAsDetermination    LoanAge         tempDefaulted  CurrentDTI     CurrentAVGValue              audit_date
 Mode:logical   Live:175189            Min.   : 0.03333   Mode:logical   Mode:logical   Mode:logical    2010-07-29 01:37:33:    27
 NA's:175189                           1st Qu.:25.76667   NA's:175189    NA's:175189    NA's:175189     2010-07-28 23:38:42:    26
                                       Median :31.83333                                                 2010-07-29 01:36:05:    26
                                       Mean   :31.85864                                                 2010-07-29 00:44:16:    25
                                       3rd Qu.:38.00000                                                 2010-07-29 01:36:07:    25
                                       Max.   :47.93333                                                 2010-07-29 01:37:35:    25
                                       NA's   : 2.00000                                                 (Other)            :175035
```

Figure 9: Structure (Upper Two Panels) and Summary (Lower Panel) of the ALBA Master Loan Tape and ALBA 2007-1 History Loan Tape.

Table 5
Cohort by Period Tabulation of ALBA Arrears Rate

Origination Date	Period																							
	6/08	7/08	8/08	9/08	10/08	11/08	12/08	1/09	2/09	3/09	4/09	5/09	6/09	7/09	8/09	9/09	10/09	11/09	12/09	1/10	2/10	3/10	4/10	5/10
10/05	0,200	0,192	0,200	0,194	0,195	0,205	0,206	0,216	0,217	0,209	0,199	0,201	0,198	0,192	0,195	0,191	0,181	0,182	0,173	0,154	0,156	0,152	0,155	0,161
11/05	0,200	0,200	0,201	0,196	0,197	0,200	0,209	0,211	0,207	0,200	0,196	0,191	0,191	0,186	0,184	0,183	0,179	0,176	0,174	0,166	0,166	0,160	0,158	0,162
12/05	0,228	0,218	0,228	0,222	0,213	0,230	0,217	0,222	0,215	0,206	0,204	0,198	0,190	0,197	0,198	0,203	0,198	0,201	0,195	0,182	0,187	0,173	0,171	0,175
3/06	0,250	0,250	0,250	0,000	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250	0,250
4/06	0,309	0,309	0,318	0,298	0,311	0,305	0,296	0,290	0,286	0,296	0,281	0,296	0,298	0,296	0,288	0,284	0,275	0,275	0,275	0,275	0,277	0,260	0,254	0,267
5/06	0,317	0,307	0,313	0,310	0,309	0,311	0,309	0,310	0,309	0,300	0,297	0,292	0,292	0,282	0,282	0,275	0,269	0,263	0,256	0,253	0,253	0,245	0,243	0,246
6/06	0,340	0,330	0,337	0,324	0,329	0,319	0,319	0,316	0,320	0,314	0,311	0,303	0,282	0,274	0,265	0,272	0,257	0,254	0,246	0,252	0,252	0,243	0,241	0,243
7/06	0,230	0,233	0,252	0,260	0,279	0,275	0,282	0,267	0,271	0,260	0,233	0,218	0,214	0,191	0,168	0,164	0,172	0,164	0,153	0,149	0,149	0,149	0,153	0,145
8/06	0,172	0,178	0,181	0,179	0,211	0,223	0,227	0,220	0,218	0,214	0,217	0,212	0,205	0,197	0,196	0,188	0,179	0,176	0,178	0,176	0,171	0,168	0,167	0,173
9/06	0,571	0,476	0,500	0,500	0,429	0,429	0,429	0,381	0,381	0,381	0,381	0,333	0,333	0,333	0,286	0,286	0,286	0,191	0,191	0,286	0,286	0,286	0,238	0,238
10/06	0,269	0,275	0,278	0,273	0,279	0,286	0,304	0,304	0,306	0,298	0,284	0,287	0,276	0,261	0,265	0,256	0,250	0,243	0,232	0,236	0,231	0,226	0,231	0,226
11/06	0,229	0,230	0,238	0,239	0,247	0,255	0,264	0,270	0,270	0,261	0,258	0,254	0,246	0,244	0,240	0,230	0,224	0,223	0,220	0,229	0,225	0,220	0,227	0,226
12/06	0,262	0,263	0,280	0,278	0,274	0,284	0,288	0,291	0,295	0,285	0,280	0,281	0,276	0,267	0,263	0,260	0,254	0,251	0,251	0,251	0,249	0,242	0,240	0,242
1/07	0,329	0,329	0,343	0,357	0,371	0,371	0,357	0,371	0,357	0,357	0,343	0,343	0,357	0,329	0,314	0,329	0,314	0,286	0,300	0,314	0,343	0,300	0,300	0,329
2/07	0,333	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,333	0,333	0,333	0,333
4/07	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,000	0,200	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
5/07	0,400	0,400	0,400	0,400	0,200	0,400	0,400	0,400	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200	0,200
6/07	0,364	0,364	0,364	0,364	0,364	0,364	0,364	0,364	0,364	0,273	0,273	0,273	0,273	0,273	0,273	0,182	0,182	0,182	0,182	0,182	0,182	0,182	0,182	0,182
7/07	0,375	0,375	0,375	0,375	0,375	0,375	0,375	0,375	0,375	0,375	0,250	0,250	0,250	0,125	0,125	0,125	0,125	0,125	0,125	0,125	0,125	0,125	0,125	0,125
8/07	0,222	0,111	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222	0,222
9/07	0,286	0,286	0,286	0,286	0,143	0,143	0,143	0,143	0,143	0,000	0,000	0,000	0,000	0,143	0,143	0,143	0,143	0,143	0,143	0,000	0,000	0,000	0,000	0,000
10/07	0,300	0,200	0,300	0,300	0,400	0,400	0,400	0,500	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,400	0,300	0,300	0,300	0,400	0,400	0,400	0,300	0,400
11/07	0,143	0,143	0,143	0,143	0,143	0,143	0,286	0,286	0,286	0,143	0,143	0,143	0,143	0,286	0,143	0,143	0,143	0,143	0,143	0,286	0,286	0,143	0,286	0,143
12/07	0,000	0,000	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500
1/08	0,571	0,571	0,714	0,571	0,429	0,429	0,429	0,429	0,429	0,429	0,571	0,571	0,571	0,571	0,571	0,571	0,571	0,571	0,571	0,571	NA	0,429	0,429	0,429
2/08	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600	0,600
4/08	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500
5/08	0,000	0,250	0,250	0,250	0,250	0,500	0,000	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,500	0,250
6/08	NA	0,143	0,286	0,571	0,571	0,571	0,571	0,571	0,571	0,571	0,571	0,429	0,429	0,286	0,286	0,143	0,143	0,143	0,143	0,143	0,143	0,143	0,143	0,143
7/08	NA	NA	0,667	0,000	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333	0,333

5.2 Cohort Analysis of ALBA Data

Section 5.1 provided a description of the data. In this section the tabulated data of Table 5 is analyzed using a cohort analysis based method. The analysis is done in R (R Development Core Team, 2011), a free software environment for statistical computing and graphics. R is chosen because it is open source, free, widely used in academics, and has many packages available to aid research in many areas of science. Tinn-R is used as a code editor to replace the native R graphical user interface (Faria, 2011). Tinn-R allows for syntax highlighting of R code. To conduct the vintage analysis Epi "A Package for Statistical Analysis in Epidemiology" was installed in R (Carstensen, Plummer, Laara, & Hills, 2011). Epi contains functions to conduct epidemiological analysis in R. The code used to fit the models and to create the graphs can be found in Appendix C.

The cohort model fitted to the data is a Holford based approach as described by Carstensen (2007). Since there currently is no theoretical basis for choosing a model, the decision to use this model is an arbitrary one. The aim of the model is to give an overview of (1) the magnitude of the rates, (2) the variation by age and (3) time trends in the rates. In essence, the method proposed by Carstensen uses the age, period, and cohort terms within a generalized linear model framework with a Poisson family error structure, a log link function, and an offset of log(loan risk-time) to account for changes in the total number of loans in the ALBA deal (the offset term separates the loan months from the rate).

The initial model can be represented as follows:

$$Y_{ij} = ln(\hat{R}_{ij}) = ln\left(\frac{O_{ij}}{N_{ij}}\right) = \alpha_i + \eta_j + \gamma_k \quad (24)$$

Where:

- Y_{ij} denotes the age–period-specific value on the outcome variable for the ith age group for $i = 1, \dots, a$ age groups at the jth time period for $j = 1, \dots, p$ time periods;
- $\hat{R}_{ij}$ denotes the observed arrears rate of the i^{th} age group in the j^{th} period;
- O_{ij} denotes the observed number of arrears cases of the i^{th} age group in the j^{th} period;
- N_{ij}denotes the amount of risk time loan-years at risk of the i^{th} age group in the j^{th} period;
- α_i denotes the ith row age effect or the coefficient for the ith age group;
- η_j denotes the jth column period effect or the coefficient for the jth time period; and
- γ_k denotes the kth diagonal cohort effect or the coefficient for the kth cohort for $k = 1, \dots, (a + p - 1)$, with $k = a - i + j$;

The number of arrears cases for age group i in period j, indicated by O_{ij} is assumed to follow a Possion distribution with mean μ_{ij}. N_{ij} is assumed to be known and, therefore, a constant.

The logarithm of the expected number of arrears incidence can be expressed as a linear function of the independent variables:

$$ln(\mu_{ij}) = ln(O_{ij}) + \alpha_i + \eta_j + \gamma_k \quad (25)$$

Where:

- $ln(O_{ij})$ is an offset with a constant coefficient of 1 for each of the observations.

In this cohort model with three terms, only second derivatives of the effects are identifiable, therefore two levels and one value of the fist derivative must be fixed (Carstensen, 2007). Carstensen (2007) proposes two alternative principles for the choice of parameterization. The principle applied in this instance (an arbitrary decision) is that:

1. The age-function should be interpretable as log age-specific rates in cohort c_0 (the reference cohort) after adjustment for the period effect;
2. The cohort function is 0 at a reference cohort c_0, interpretable as log rate ratio relative to cohort c_0; and
3. the period function is 0 on average with 0 slope, interpretable as log rate ratios relative to the age-cohort prediction (residual log rate ratio).

Thus for the ALBA data the model was fitted so that age effects are presented as arrears rates for the reference vintage 2006-08-30 (an arbitrary decision). Additionally, origination effects represent arrears rate ratios relative to the 2006-08-30 reference vintage, whereas period effects are constrained to be 0 on average with 0 slope. The first choice fixes one constant (0 at 2006-08-30), and the third fixes a level (0 at 2006-08-30) and a slope (0 slope for the period function). The inclusion of the slope (drift) with the cohort effect makes the age-effects interpretable as cohort-specific arrears rates (longitudinal rates).

The extensive parameterization procedure is described in Carstensen (2007). The idea is to get three sets of columns in the design matrix that directly allows for computation of age, period, and cohort effects at any set of points wished. In this chapter the approximation procedure proposed by Carstensen (2007) is used. First, an age-cohort model is fitted. By omission of an explicit intercept and choosing a suitable reference for the cohort, the age effect will be log rates for the reference cohort and the cohort effect will be log rate ratios relative to this. The log of the fitted age and cohort values from this model is then used as an offset variable in a model with period-effects. The period effects from this model (also omitting an explicit intercept) are then used as the residual log RRs by period.

The estimates produced by this sequential procedure are not the maximum likelihood estimates from the cohort model, they are marginal age-cohort estimates and period estimates conditional on the estimates from the age-cohort model, but in practice they will be similar to the maximum likelihood estimates (Carstensen, 2007). For any chosen constraint there will be three estimated functions which sum to the fitted log rates.

The estimated age-curve is found by taking the unique rows of the age-part of the design matrix, where each row corresponds to an observed age in the data. This is then multiplied with the vector of age-parameters to give the curve of estimated log rates. Pre- and post-multiplication on the variance-covariance matrix of the age parameters gives the variances needed to construct confidence limits for the log rates. Finally, the log rates are transformed to the rate scale. The same procedure is used to obtain the arrears rate ratio curves for period and cohort.

Figure 10 shows the graphed results obtained from fitting the cohort model. 95% confidence limits are plotted together with the fitted values. The parameters in this model represent age-specific rates, that approximates the rates in the 2006-08-30 cohort, the cohort arrears rates relative to this cohort, and finally period "residual" arrears rates. But note that an explicit decision has been made as to how the period residuals are defined, namely as the deviations from the line between 2006-08-30 and 2010-05-30.

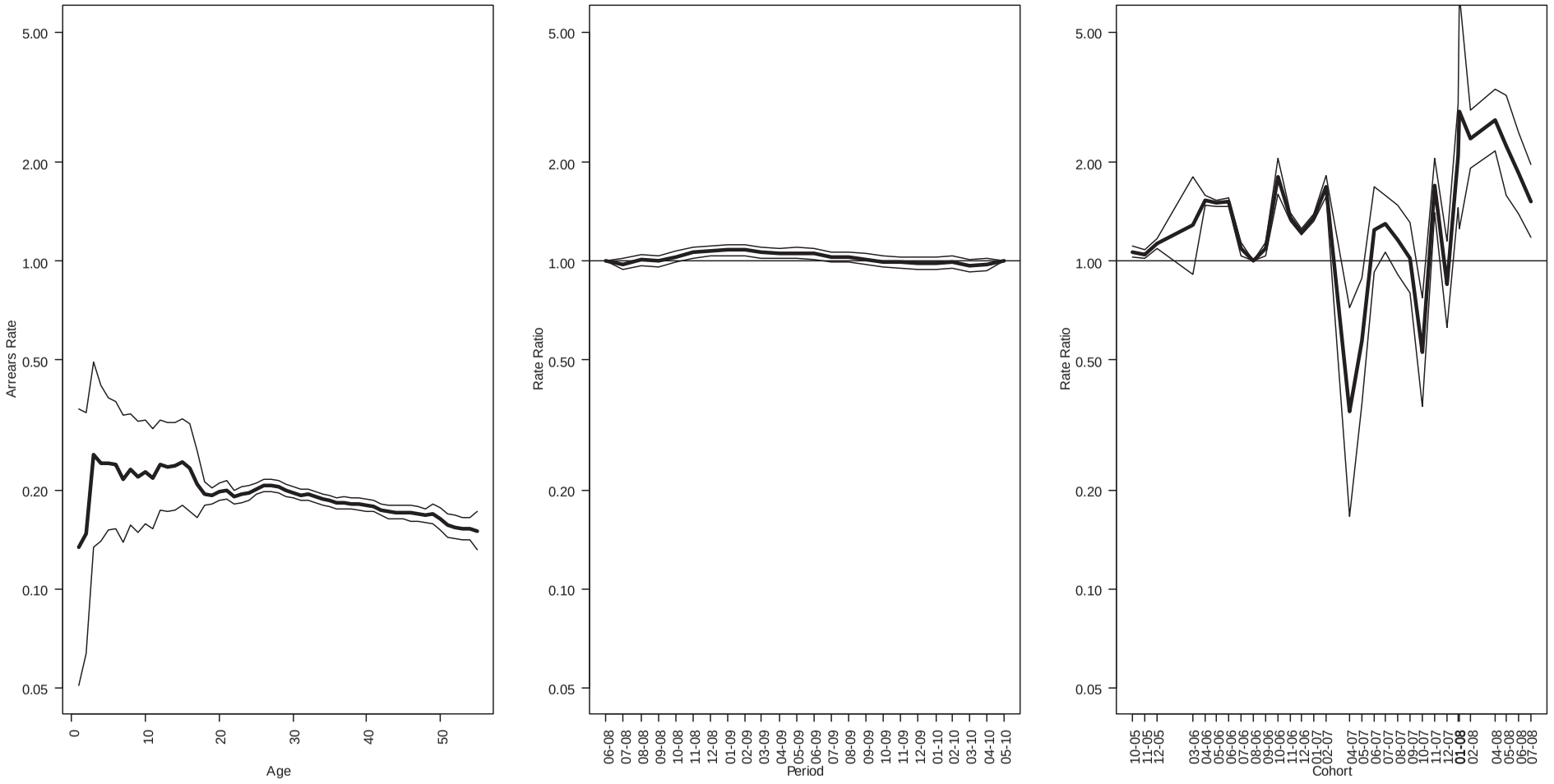


Figure 10: Estimated Effects from the Cohort Model (ALBA Arrears Rate)

Note. The curve in the left graph represents the age-specific rates per loan year at risk-for the reference period (2006-08-30). The curve in the middle graph shows the arrears rate ratios of periods relative to the reference period (2006-08-30). The curve in the right graph shows the arrears rate ratios of cohorts conditional on the estimated age and period effects. Fitted values are plotted together with 95% confidence limits.

5.3 Discussion

An interpretation of the results goes beyond the scope of this thesis. As was discussed in chapter 4 the conceptualization of what constitutes a cohort effect in combination with theory about the relationship between maturation, extrinsic, and origination effects and the problem under consideration should help to select the right model and, if applicable for the model under consideration, the constraints. In absence of theory the model used was arbitrarily selected. As a result the underlying assumptions of the model have not been tested, the constraints applied cannot be justified, and therefore, the results not interpreted. However, what can be done is pointing out several notions with respect to the modelling procedure.

Multiple assumptions and decisions have been made and constraints have been applied in the modelling process. First, are the assumptions that relate to the GLM framework in general and the Poisson regression model in specific. A discussion of these assumptions, how they affect the results, and how to test whether these are appropriate can be found elsewhere[19].

Second, since a generalized linear model has been used, additional identifying constraints were applied in order to get an estimation of the maturation, extrinsic, and origination effects. For the Holford based models, two levels and one value of the fist derivative must be fixed. The constraint applied was to fix a constant (cohort function 0 at a reference cohort, c_0), and to fix a level (make the period function 0 at a reference period, p_0) and a slope (0 slope for the period function). These constraints are very specific and have an effect on the estimated parameters. Depending on the problem at hand, the role of period and cohort could, for example, be interchanged. The point is that in cohort analysis based vintage analysis these decisions should be guided by theory regarding the relationship between maturation, extrinsic, and origination effects and the relative importance of these three effects. This is also important for the next issue.

Third, in sociology and epidemiology, age is believed to be the predominant factor in most cohort studies. Numerous epidemiologic studies, therefore, adopted the following model fitting sequence (Carstensen, 2007; Clayton & Schifflers, 1987; Holford, 1991):

1 The cohort analysis starts with fitting an age model;
2 If an age-only model does not yield a good fit, fitting an age-drift model is the next step;
3 If an age-drift model is not adequate to explain the data, an age-period model and an age-cohort model is considered; and
4 A full three-factor model is implemented if neither of the two-factor models is well-fitted.

This stepwise fitting procedure is usually implemented in the modeling phase in one go, after which the best results in terms of deviance are compared (see for example step 9 in the R code in Appendix C). Residuals of the selected two-factor model are fitted against the remaining variable of interest. The parameter estimates for the third factor obtained in this way are conditional only (Carstensen, 2007). This implies that with the sequential order of model fitting, the last factor should be the least influential of the three. In the cohort analysis carried out in this chapter, the epidemiologic stepwise fitting procedure has been applied, while a different sequence would result in different estimates of the maturation, extrinsic and origination effects. Therefore, for a cohort analysis based vintage analysis that makes use of this procedure one has to properly adjust this fitting sequence to what are considered the most and least important factors.

[19] The critical assumptions of the GLM framework are discussed by Breslow (1996). The assumptions of the Poisson regression by Gardner Mulvey and Shaw (1995).

Fourth, for the parameterization of the cohort model fitted in this chapter, a procedure was used which approximates the solution obtained by a full parameterization procedure. An indication of how the obtained solution deviates from the full parameterization procedure is missing and seems not to have been tested by Carstensen (2007). This deviation should be studied in more detail before the approximation is further used.

Fifth, any tabulation of data represents an information loss due to the rounding as a result of binning the observations per time interval (Carstensen, 2007). Therefore, the intervals should be as small as possible. Most epidemiologic and sociologic research create coarse cohorts (usually 1 or 5 year cohorts) due to the limitations of the population data that is available (Carstensen, 2007). In that respect structured credit loan tapes contain abundant data, since most structured credit loan tapes are based on monthly or quarterly updated historic data, and the population is usually considered to be the loan tape under investigation or the loan database available to the analyst.

What can be learned from the application of the cohort model to the mortgage data set and the points mentioned in the discussion above is that it is not difficult to come up with estimates for the maturation, extrinsic, and origination effects. The challenge is to produce sensible estimates of the effects. Besides, because of the difference between cohort analysis in sociology and epidemiology on the one hand, and structured credit markets on the other, any application of methodology from sociology and epidemiology should be approached with care.

The final chapter will provide an answer to the central question of this thesis: How can cohort analysis improve the vintage analysis of loans?

6 Conclusion, Recommendations, and Limitations

In this chapter an answer is provided to the central research question of this thesis: How can cohort analysis improve the vintage analysis of loans?

Section 6.1 presents a conclusion with respect to the central research question. Section 6.2 provides recommendations for further research. Section 6.3 discusses the limitations of this thesis.

6.1 Conclusion

The objective of this thesis was to explore whether and how cohort analysis can improve the vintage analysis techniques that are used for the analysis of loans by structured credit market participants. To reach this objective, an exploratory study was conducted. The purpose of this exploratory research was to move towards a clearer understanding of how cohort analysis and vintage analysis are carried out, to develop ideas of what are significant lines of relation, and ultimately to learn how cohort analysis can contribute to vintage analysis.

The relation between cohort analysis and vintage analysis becomes clear when vintage analysis is viewed as a problem of how to unravel maturation, extrinsic, and origination effects on structured credit performance. The reason for adopting this perspective is that the data tables on which the vintage graphs are based, that are created to study vintages, are duplicates of the tables that serve as the starting point in cohort analysis. When vintage analysis is viewed from a cohort analysis perspective, cohort analysis potentially has a lot to contribute to the solution of this problem of how to unravel vintage, maturation, and extrinsic effects on structured credit performance.

On the basis of a literature study a classification of a plethora of models available to deal with the separation of these three effects was provided in chapter 3. However, an important reminder of the limitations that affect all generalized linear models used with the purpose of unraveling age, period, and cohort, or in the case of vintage analysis of structured credit, vintage, extrinsic, and maturation effects was the model identification problem. This is the reason for the wide range of models available. The model identification problem is the result of a lack of understanding of the theory behind the problem that is analyzed. And theory is what structured credit is short of. Therefore, a hurdle that needs to be taken before the potential of cohort analysis can be used, is fundamental research into the specific causal mechanisms that underlie the performance of structured credit that can be measured and analyzed. Only then the true potential of cohort model based vintage analysis can be maximized.

As was noted in section 1.4, the main academic value of explorative studies is generating research questions and hypotheses for additional investigation. These are presented in the next section.

6.2 Recommendations for Further Research

Although there seems to be a general consensus among structured credit market participants about what drives structured credit performance, the body of literature regarding the relationships between the different mechanisms that drive structured credit performance is, as far as the author is aware, small. An important question about the applicability of cohort analysis to vintage analysis by structured credit market participants, that was left unanswered in this thesis, relates to the substantive importance of origination effects. Do maturation patterns, for collateral that has the same origination date, respond only to extrinsic effects, so that the experience of a vintage can be described completely by maturation effects and the effects of the periods its members live through? Or do they respond to

additional, origination effect as well? And, if so, how does the origination effect manifests itself? With respect to these questions two alternative hypotheses can be formulated:

Hypothesis 1a: Maturation and extrinsic effects should be treated as confounders of the origination effect.

Hypothesis 1b: Maturation and extrinsic effects should be treated as effect modifiers of the origination effect.

A suggested starting point for further research in this respect is an article by Winship and Harding (2008) in which a theory is presented on how to specify the mechanisms by which aging, period-related changes, and cohort-related change processes act on the dependent variable of interest. Although the focus is on cohort models in sociology, Winship and Harding note that their approach can also be applied to other problems in which there are substantively distinct but linearly dependent or, more generally, functionally dependent explanatory variables processes acting on the dependent variable, as is the case in cohort analysis based vintage analysis.

Once the substantive importance of origination effects is determined, the cohort analysis 'toolbox' becomes available to aid vintage analysis by structured credit market participants. The model overview in Table 4 in section 3.4 can serve as a guide to the cohort models available. Chapter 5 of this thesis can provide insight into the implementation process. Several directions of research can be pursued from here.

First, it was observed that a thorough interdisciplinary literature review on cohort analysis is missing from the academic literature. The categorization provided in section 3.4 can serve as a starting point in this respect, but is, due to the scope of thesis, not complete. A thorough interdisciplinary literature review can add to the cohort analysis discussion in general and the cohort based vintage analysis methods in particular by listing the models available to the analyst and the merits and drawbacks of the models.

Second, an appealing aspect of unraveling maturation, extrinsic, and origination effects is forecasting structured credit performance. Most private-sector analysts arrive at their estimates by extrapolating the performance – defaults, loss severities, and total loss rates – of each vintage, based on its own history as well as the typical progression pattern through time (Greenlaw, et al., 2008). A cohort analysis based vintage analysis that separates the historical maturation, extrinsic, and origination effects could serve as the basis for a better model to forecast vintage performance. This would involve the following steps. First the loan tape data is processed to produce the key rates per (preferably monthly) vintage. Next, the historic maturation, extrinsic, and origination effects are estimated. Then, scenario's for the future are created. Finally, the key rates are forecasted and the resulting output is used in the rating and valuation models. The implicit hypothesis that is made regarding forecasting is that cohort analysis based vintage analysis allows for better estimates of the input parameters for structured credit rating and valuation models. This hypothesis should be explicitly tested:

Hypothesis 2a: Input parameters estimates for structured credit rating and valuation models that are obtained by cohort models based vintage analyses, provide better results than input parameters that are based on expert judgment.

Hypothesis 2b: Input parameters estimates for structured credit rating and valuation models that are obtained by cohort models based vintage analyses, provide better results than input parameters that are based on vintage extrapolation.

Finally, a direction of research that might be of interest is to investigate the usefulness of concepts and models from biostatistics and epidemiology for structured credit and finance in general. During the research the author has come aware of many similarities between the different fields. The study of people developing diseases during their lifetime which is sometimes based on their health status and sometimes seems to be random, shows many similarities with the study of default and prepayments of loan pools. Concepts such as the incidence rate that for a dynamic population, take account of the effect of individuals that may have been at risk for different lengths of time before developing a disease, are useful for structured credit as well since structured credit often deals with dynamic loan pools and, as far as the author is aware, do not correct for the amount of time a loan is in the data under study.

6.3 Research Limitations

This thesis has several limitations. These are listed below.

First, and foremost, is the explorative nature of this interdisciplinary research. This research is mainly conducted via a literature study. Therefore, the researcher has been the research instrument, and, due to the open-ended nature of explorative research, determined the direction of the research. The researcher has no experience in the sociologic and epidemiologic fields. This means that important literature could have been missed or misinterpreted. This potential handicap has been tried to overcome by the thoroughness of the literature study. Wherever possible, results have been confirmed in multiple sources.

Second, the internal validity of this research can be improved. Although the author feels that the general principles are accurately covered, the timespan within which this research was conducted resulted in that several themes could not be explored as thorough as the author had hoped for. For examples, Moody's structured credit rating methodology was discussed in-depth to show how vintage analysis is used in the rating process of CRAs. This could be complemented with an in-depth discussion of the structured credit rating methodology of DBRS, Fitch, and S&P to strengthen the argument. Although, in general, vintage analysis is used in the rating process in the same manner, subtle differences could have been uncovered. Besides, SCIs methods were explored on the basis of a narrow set of practical literature and experience of the author. To improve the internal validity, this could have been confirmed through interviews with major SCIs.

Third is the scope of this thesis. A relation between cohort analysis and vintage analysis was presented due to author's observation that cohort analysis suits the dual-time nature of vintage analysis. There might be other fields of study that can add to the discussion presented in this thesis which have not been considered here.

Besides reflecting on the research' limitations, the author has provided the schedule of this thesis in Appendix D to provide further insight into the research process. Furthermore, the author's reflections on his personal learning objectives for this thesis can be found in Appendix E.

References

AB Alert. (2009, July 10). Rating-Agency Shares Track Downward. *Asset-Backed Alert*.

AB Alert. (2011a, July 8). Rating Agencies Feel Ranking Shakeup. *Asset-Backed Alert*.

AB Alert. (2011b, January 14). S&P Holds On as DBRS Surges. *Asset-Backed Alert*.

Alexander, K., Eatwell, J., Persaud, A., & Reoch, R. (2007). Financial Supervision and Crisis Management in the EU. Brussels: European Parliament's Committee on Economic and Monetary Affairs.

Allen, M. P. (2004). *Understanding Regression Analysis*. New York, NY: Springer-Verlag.

Altman, E. I., & Rijken, H. A. (2004). How Rating Agencies Achieve Rating Stability. *Journal of Banking & Finance, 28*(11), 2679-2714.

Ashcraft, A., Goldsmith-Pinkham, P., Hull, P., & Vickery, J. (2011). Credit Ratings and Security Prices in the Subprime MBS Market. *American Economic Review, 101*(3), 115-119.

Ashcraft, A., Goldsmith-Pinkham, P., & Vickery, J. (2010). MBS Ratings and the Mortgage Credit Boom *Federal Reserve Bank of New York Staff Reports* (Vol. 449). New York, NY: Federal Reserve Bank of New York.

Barrett, J. C. (1973). Age, Time and Cohort Factors in Mortality from Cancer of the Cervix. *Journal of Hygiene, 71*(2), 253-259.

Barrett, J. C. (1978). The Redundant Factor Method and Bladder Cancer Mortality. *Journal of Epidemiology and Community Health, 32*(4), 314-316.

Ben-Israel, A., & Greville, T. N. E. (2003). *Generalized Inverses: Theory and Applications* (2 ed. Vol. 15). New York, NY: Springer-Verlag.

Benmelech, E., & Dlugosz, J. L. (2009). The Credit Rating Crisis. In D. Acemoglu, K. Rogoff & M. Woodford (Eds.), *NBER Macroeconomics Annual 2009* (Vol. 24, pp. 161-207). Chicago, IL: University of Chicago Press.

Berzuini, C., & Clayton, D. (1994). Bayesian Analysis of Survival on Multiple Time Scales. *Statistics in Medicine, 13*(8), 823-838.

Bhattacharya, A. K., & Fabozzi, F. J. (1996). *Asset-Backed Securities*. New York, NY: John Wiley & Sons.

Breeden, J. L. (2007). Modeling Data with Multiple Time Dimensions. *Computational Statistics & Data Analysis, 51*(9), 4761-4785.

Breslow, N. (1996). Generalized Linear Models: Checking Assumptions and Strengthening Conclusions. *Statistica Applicata, 8*, 23-41.

Bryer, L. G., Lebson, S. J., & Asbell, M. D. (2011). *Corporate Intellectual Property Management in the 21st Century: A shift in Strategic and Financial Mangement*. New York, NY: John Wiley & Sons.

Burns, P., & Stanley, A. (2001). Managing Consumer Credit Risk *Federal Reserve Bank of Philadelphia Payment Cards Center Discussion Paper* (pp. 1-7).

Carstensen, B. (2007). Age–Period–Cohort Models for the Lexis Diagram. *Statistics in medicine, 26*(15), 3018-3045.

Carstensen, B., Plummer, M., Laara, E. & Hills, M. (2011). Epi: A Package for Statistical Analysis in Epidemiology. R package (Version 1.1.24). Retrieved from http://CRAN.R-project.org/package=Epi.

Chambers, D. R., Kelly, M. A., & Lu, Q. (2010). The Role of the Constant Recovery Assumption in the Subprime Bubble. *The Journal of Alternative Investments, 13*(1), 30-40.

Chie, W. C., Chen, C. F., Lee, W. C., Chen, C. J., & Lin, R. S. (1995). Age-Period-Cohort Analysis of Breast Cancer Mortality. *Anticancer Research, 15*(2), 511-515.

Choudhry, M., Joannas, D., Landuyt, G., Pereira, R., & Pienaar, R. (2010). *Capital Market Instruments: Analysis and Valuation* (3 ed.). New York, NY: Palgrave Macmillan.

Clayton, D., & Schifflers, E. (1987). Models for Temporal Variation in Cancer Rates. II: Age-Period-Cohort Models. *Statistics in Medicine, 6(4)*, 469-481.

CML Research. (2006). Adverse Credit Mortgages. *Housing Finance, 10*, 1-12.

Cornaggia, J., Cornaggia, K. R., & Hund, J. (2011). Credit Ratings across Asset Classes: A ≡ A? *SSRN eLibrary*, 1-66.

Crouhy, M. G., Jarrow, R. A., & Turnbull, S. M. (2008). The Subprime Credit Crisis of 2007. *The Journal of Derivatives, 16*(1), 81-110.

de Vaus, D. A. (2001). *Research Design in Social Research*. London: Sage Publications.

Decarli, A., & La Vecchia, C. (1987). Age, Period and Cohort Models: Review of Knowledge and Implementation in GLIM. *Rivista di Statistica Applicata, 20*, 397-410.

Fabozzi, F. J. (2000). *Investing in Asset-Backed Securities*. New Hope, PA: Frank J. Fabozzi Associates.

Fannie Mae. (May 6, 2011). Form 10-Q Credit Supplement. Retrieved August 8, 2011 http://phx.corporate-ir.net/phoenix.zhtml?c=108360&p=irol-secQuarterly&control_SelectGroup=Quarterly%20Filings.

Faria, J. C. (2011). Resources of Tinn-R GUI/Editor for R Environment (Version 2.3.7.1). Ilheus, Brasil: UESC.

Farkas, G. (1977). Cohort, Age, and Period Effects upon the Employment of White Females: Evidence for 1957–1968. *Demography, 14*(1), 33-42.

Fienberg, S. E., & Mason, W. M. (1979). Identification and Estimation of Age-Period-Cohort Models in the Analysis of Discrete Archival Data. In K. F. Schuessler (Ed.), *Sociological Methodology* (Vol. 10, pp. 1-67). San Fransisco, CA: Jossey-Bass.

Firebaugh, G., & Davis, K. E. (1988). Trends in Antiblack Prejudice, 1972-1984: Region and Cohort Effects. *American Journal of Sociology, 94*(2), 251-272.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models. *Psychological Bulletin, 118*(3), 392-404.

Ghent, A. C., & Kudlyak, M. (2011). *Recourse and Residential Mortgage Default: Theory and Evidence from US States*. Federal Reserve Bank of Richmond Working Paper 09-10R.

Glenn, N. D. (1976). Cohort Analysts' Futile Quest: Statistical Attempts to Separate Age, Period and Cohort Effects. *American Sociological Review, 41*(5), 900-904.

Glenn, N. D. (1994). Television Watching, Newspaper Reading, and Cohort Differences in Verbal Ability. *Sociology of Education, 67*(3), 216-230.

Glenn, N. D. (2005). *Cohort analysis* (2 ed.). London: Sage Publications.

Goodman, L. (2007, May 20-23). *The Changing Face of the Mortgage Market.* Paper presented at the MBA National Secondary Conference, New York.

Graybill, F. A. (1961). *An introduction to linear statistical models: Volume 1*. New York, NY: McGraw-Hill.

Greenlaw, D., Hatzius, J., Kashyap, A. K., & Shin, H. S. (2008). Leveraged Losses: Lessons from the Mortgage Market Meltdown *U.S. Monetary Policy Forum Report No. 2* (pp. 21-34): Rosenberg Institute, Brandeis International Business School and Initiative on Global Markets, Universtiy of Chicago Graduate School of Business.

Harding, D. J., & Jencks, C. (2003). Changing Attitudes Toward Premarital Sex: Cohort, Period, and Aging Effects. *The Public Opinion Quarterly, 67*(2), 211-226.

Hayre, L. (2001). *Salomon Smith Barney Guide to Mortgage-Backed and Asset-Backed Securities*. New Jersey, NY: John Wiley & Sons.

Heckman, J., & Robb, R. (1985). Using Longitudinal Data to Estimate Age, Period and Cohort Effects in Earnings Equations. In W. M. Mason & S. E. Fienberg (Eds.), *Cohort analysis in social research: Beyond the identification problem* (1 ed., pp. 137–150). New York, NY: Springer-Verlag.

Hill, C. (2004). Regulating the Rating Agencies. *Washington University Law Quarterly, 82*, 43-95.

Holford, T. R. (1983). The Estimation of Age, Period and Cohort Effects for Vital Rates. *Biometrics, 39*(2), 311-324.

Holford, T. R. (1991). Understanding the Effects of Age, Period, and Cohort on Incidence and Mortality Rates. *Annual Review of Public Health, 12*(1), 425-457.

Holford, T. R. (1992). Analysing the Temporal Effects of Age, Period and Cohort. *Statistical Methods in Medical Research, 1*(3), 317-337.

Holford, T. R. (2004). Temporal Factors in Public Health Surveillance: Sorting Out Age, Period, and Cohort Effects. In R. Brookmeyer & D. F. Stroup (Eds.), *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance* (pp. 99-126). New York, NY: Oxford University Press.

Holford, T. R. (2005). Age-Period-Cohort Analysis. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics: 8-Volume Set* (2 ed.). New Jersey, NY: John Wiley & Sons.

Holford, T. R., Zhang, Z., & McKay, L. A. (1994). Estimating Age, Period and Cohort Effects Using the Multistage Model for Cancer. *Statistics in medicine, 13*(1), 23-41.

Hu, J. C. (2011). *Asset Securitization: Theory and Practice* (Vol. 679). New Jersey, NY: John Wiley & Sons.

Hull, A., & White, J. (2010). The Risk of Tranches Created from Residential Mortgages. *Financial Analysts Journal, 66*(5), 54–67.

Jobst, A. (2008). What is Securitization. *Finance & Development, 45(3)*, 48-49.

Kahn, J. R., & Mason, W. M. (1987). Political Alienation, Cohort Size, and the Easterlin Hypothesis. *American Sociological Review, 52*(2), 155-169.

Keyes, K. M., & Li, G. (2010). A Multiphase Method for Estimating Cohort Effects in Age-Period Contingency Table Data. *Annals of Epidemiology, 20*(10), 779-785.

Keyes, K. M., Utz, R. L., Robinson, W., & Li, G. (2010). What is a cohort effect? Comparison of Three Statistical Methods for Modeling Cohort Effects in Obesity Prevalence in the United States, 1971-2006. *Social Science & Medicine, 70*(7), 1100-1108.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (2007). *Applied Regression Analysis and other Multivariable Methods* (4 ed.). Belmont, CA: Thomson Brooks/Cole.

Knoke, D., & Hout, M. (1976). Reply to Glenn. *American Sociological Review, 41*(5), 905-908.

Kothari, V. (2006). *Securitization: The Financial Instrument of the Future* (3 ed.). Singapore: John Wiley & Sons (Asia).

Kregel, J. (2008). Using Minsky's Cushions of Safety to Analyze the Crisis in the US Subprime Mortgage Market. *International Journal of Political Economy, 37*(1), 3-23.

Kupper, L. L., Janis, J. M., Karmous, A., & Greenberg, B. G. (1985). Statistical Age-Period-Cohort Analysis: A Review and Critique. *Journal of Chronic Diseases, 38*(10), 811-830.

Land, K. C. (2011). Age-Period-Cohort Analysis: New Models, Methods, and Empirical Analyses. *Indiana University 2010-2011 Workshop in Methods*. Bloomington, Indiana.

Lang, W. W., & Jagtiani, J. A. (2010). The Mortgage and Financial Crises: The Role of Credit Risk Management and Corporate Governance. *Atlantic Economic Journal, 38*(2), 123-144.

Lazaridis, A. (1986). A Note Regarding the Problem of Perfect Multicollinearity. *Quality & Quantity, 20*(2), 297-306.

Lee, W. C., & Lin, R. S. (1994). Interactions between Birth Cohort and Urbanization on Gastric Cancer Mortality in Taiwan. *International Journal of Epidemiology, 23*(2), 252-260.

Lee, W. C., & Lin, R. S. (1996). Autoregressive Age–Period–Cohort Models. *Statistics in Medicine, 15*(3), 273-281.

Li, G., & Baker, S. P. (2012). *Injury Research: Theories, Methods, and Approaches*. New York, NY: Springer.

López-Abente, G., Pollán, M., & Jiménez, M. (1993). Female Mortality Trends in Spain due to Tumors Associated with Tobacco Smoking. *Cancer Causes and Control, 4*(6), 539-545.

Making Securitization Work for Financial Stability and Economic Growth. (2009) *Statement Nº 20*. Santiago de Chile: Shadow Financial Regulatory Committees of Asia, Australia-New Zealand, Europe, Japan, Latin America, and the United States.

Mason, K. O., Mason, W. M., Winsborough, H. H., & Poole, W. K. (1973). Some Methodological Issues in Cohort Analysis of Archival Data. *American Sociological Review, 38*(2), 242-258.

Mason, W. M., & Fienberg, S. E. (1985a). *Cohort Analysis in Social Research: Beyond the Identification Problem*. New York, NY: Springer-Verlag.

Mason, W. M., & Fienberg, S. E. (1985b). Introduction: Beyond the identification problem. In W. M. Mason & S. E. Fienberg (Eds.), *Cohort analysis in social research: Beyond the identification problem* (pp. 1-8). New York, NY: Springer-Verlag.

Mason, W. M., Mason, K. O., & Winsborough, H. H. (1976). Reply to Glenn. *American Sociological Review, 41*(5), 904-905.

Mason, W. M., & Wolfinger, N. H. (2001). Cohort Analysis. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 151-228). Oxford: Elsevier Science.

Mason, W. M., & Wolfinger, N. H. (2001). Cohort Analysis. In J. S. Neil & B. B. Paul (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 2189-2194). Oxford: Pergamon.

McNally, R., Alexander, F. E., Staines, A., & Cartwright, R. A. (1997). A Comparison of Three Methods of Analysis for Age-Period-Cohort Models with Application to Incidence Data on Non-Hodgkin's Lymphoma. *International Journal of Epidemiology, 26*(1), 32-46.

Megginson, W. L., & Smart, S. B. (2008). *Introduction to Corporate Finance*: Cengage Learning.

Melennec, O. (2000). Asset Backed Securities: Practical Guide for Investors (pp. 10). London: Société Générale European ABS Group.

Moody's. (2005). Historical Default Data Analysis for ABS Transactions in EMEA *International Structured Finance Rating Methodology*. London: Moody's Investors Service.

Moody's. (2009). Data Requirements for Australian ABS *International Structured Finance Rating Implementation Guidance*. Sydney: Moody's Investors Service.

Münkel, H. (2006). Asset-Backed Securities: It's as easy as this! A Practical Factbook (C. M. G. M. Research, Trans.). Munich: Bayerische Hypo- und Vereinsbank AG.

Myers, D., & Lee, S. W. (1998). Immigrant Trajectories into Homeownership: a Temporal Analysis of Residential Assimilation. *International Migration Review, 32*(3), 593-625.

Nakamura, T. (1986). Bayesian Cohort Models for General Cohort Table Analyses. *Annals of the Institute of Statistical Mathematics, 38*(1), 353-370.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General), 135*(3), 370-384.

O'Brien, R. M. (2000). Age Period Cohort Characteristic Models. *Social Science Research, 29*(1), 123-139.

O'Brien, R. M., Stockard, J., & Isaacson, L. (1999). The Enduring Effects of Cohort Characteristics on Age-Specific Homicide Rates, 1960-1995. *American Journal of Sociology, 104*(4), 1061-1095.

O'Brien, R. M. (2010). The Age–Period–Cohort Conundrum as Two Fundamental Problems. *Quality & Quantity, 45(6)*, 1429-1444.

Osmond, C., & Gardner, M. J. (1982). Age, Period and Cohort Models Applied to Cancer Mortality Rates. *Statistics in Medicine, 1*(3), 245-259.

Pagano, M., & Volpin, P. (2010). Credit Ratings Failures and Policy Options. *Economic Policy, 25*(62), 401-431.

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing (Version 2.13.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Raynes, S., & Rutledge, A. (2003). *The Analysis of Structured Securities: Precise Risk Measurement and Capital Allocation*. Oxford: Oxford University Press.

RBS, & UniCredit. (2011). Driver Nine Asset-Backed Security Marketing Material.

Reymen, I. M. M. J. (2001). *Improving Design Processes through Structured Reflection: A Domain-independent Approach.* Ph.D Disseration, Eindhoven University of Technology, Eindhoven. Retrieved from http://www.dbnl.org/tekst/groo004meth01_01/colofon.php.

Robertson, C., & Boyle, P. (1986). Age, Period and Cohort Models: The Use of Individual Records. *Statistics in Medicine, 5*(5), 527-538.

Robertson, C., & Boyle, P. (1998). Age–Period–Cohort Analysis of Chronic Disease Rates. I: Modelling Approach. *Statistics in Medicine, 17*(12), 1305-1323.

Robertson, C., Gandini, S., & Boyle, P. (1999). Age-Period-Cohort Models: A Comparative Study of Available Methodologies. *Journal of Clinical Epidemiology, 52*(6), 569-583.

Rodgers, W. L. (1982a). Estimable Functions of Age, Period, and Cohort Effects. *American Sociological Review, 47*(6), 774-787.

Rodgers, W. L. (1982b). Reply to Comment by Smith, Mason, and Fienberg. *American Sociological Review, 47*(6), 793-796.

Roush, G. C., Holford, T., Schymura, M., & White, C. (1987). *Cancer Risk and Incidence Trends. The Connecticut Perspective*. Washington, D.C.: Hemisphere Publishing Corporation

Roush, G. C., Schymura, M. J., Holford, T. R., White, C., & Flannery, J. T. (1985). Time Period Compared to Birth Cohort in Connecticut Incidence Rates for Twenty-Five Malignant Neoplasms. *Journal of the National Cancer Institute, 74*(4), 779-788.

Royse, D. D. (2010). *Research Methods in Social Work* (6 ed.). Belmont, CA Thomson Brooks/Cole.

Ruane, J. M. (2005). *Essentials of Research Methods: A Guide to Social Science Research*. Oxford: Blackwell Publishing.

Ryder, N. B. (1965). The Cohort as a Concept in the Study of Social Change. *American Sociological Review, 30*(6), 843-861.

Saunders, A., & Allen, L. (2010). *Credit Risk Measurement in and out of the Financial Crisis: New Approaches to Value at Risk and Other Paradigms*. New Jersey, NY: John Wiley & Sons.

Seber, G. A. F. (2008). *A Matrix Handbook for Statisticians* (Vol. 746). New Jersey, NY: John Wiley & Sons.

SEC. (2008). Summary Report of Issues Identified in the Security and Exchange Commission Staff's Examinations of Select Credit Rating Agencies. New York City: United States Securities and Exchange Commission.

Selvin, S. (2004). *Statistical Analysis of Epidemiologic Data* (3 ed.). New York, NY: Oxford University Press.

Servigny, A., & Jobst, N. (2007). *The Handbook of Structured Finance*. New York, NY: McGraw-Hill.

Shahpar, C., & Li, G. (1999). Homicide Mortality in the United States, 1935–1994: Age, Period, and Cohort Effects. *American Journal of Epidemiology, 150*(11), 1213-1222.

Simon, H. A. (1962, 8/9 November). *Approaching the Theory of Management.* Paper presented at the Toward a Unified Theory of Management Conference, Los Angeles.

Smith, H. L., Mason, W. M., & Fienberg, S. E. (1982). Estimable Functions of Age, Period, and Cohort Effects: More Chimeras of the Age-Period-Cohort Accounting Framework: Comment on Rodgers. *American Sociological Review, 47*(6), 787-793.

Tabeau, E. (2001). A Review of Demographic Forecasting Models for Mortality. In E. Tabeau, A. v. d. Berg Jeths & C. Heathcote (Eds.), *Forecasting Mortality in Developed Countries: Insights from a Statistical, Demographic, and Epidemiological Perspective*. Dordrecht: Kluwer Academic Publishers.

Tang, T., & Kurashina, S. (1987). Age, Period and Cohort Analysis of Trends in Mortality from Major Diseases in Japan, 1955 to 1979: Peculiarity of the Cohort Born in the Early Showa Era. *Statistics in Medicine, 6*(6), 709-726.

Tango, T. (1988). Statistical Modelling of Lung Cancer and Laryngeal Cancer Incidence in Scotland, 1960-1979. [Letter]. *American Journal of Epidemiology, 128*(3), 677-678.

Tarone, R. E., & Chu, K. C. (1992). Implications of Birth Cohort Patterns in Interpreting Trends in Breast Cancer Rates. *Journal of the National Cancer Institute, 84*(18), 1402-1410.

Tarone, R. E., & Chu, K. C. (1996). Evaluation of Birth Cohort Patterns in Population Disease Rates. *American Journal of Epidemiology, 143*(1), 85-91.

Von Furstenberg, G. M., & Green, R. J. (1974). Home Mortgage Delinquencies: A Cohort Analysis. *The Journal of Finance, 29*(5), 1545-1548.

White, L. J. (2010). Markets: The Credit Rating Agencies. *The Journal of Economic Perspectives, 24*(2), 211-226.

Wickramaratne, P. J., Weissman, M. M., Leaf, P. J., & Holford, T. R. (1989). Age, Period and Cohort Effects on the Risk of Major Depression: Results from Five United States Communities. *Journal of Clinical Epidemiology, 42*(4), 333-343.

Winship, C., & Harding, D. J. (2008). A General Strategy for the Identification of Age, Period, Cohort Models: A Mechanism Based Approach. *Sociological Methods and Research, 36*(3), 362-401.

Yang, Y. (2005). *New Avenues for Cohort Analysis in Social Research.* Doctor of Philosophy Dissertation, Duke University.

Yang, Y. (2007). Age/Period/Cohort Distinctions. In K. S. Markides (Ed.), *Encyclopedia of Health & Aging* (pp. 20-22). Los Angeles, CA: Sage Publications.

Yang, Y. (2008). Social Inequalities in Happiness in the United States, 1972 to 2004: An Age-Period-Cohort Analysis. *American Sociological Review, 73*(2), 204.

Yang, Y. (2010). Aging, Cohorts, and Methods. In R. H. Binstock & L. K. George (Eds.), *Handbook of Aging and the Social Sciences*. London: Elsevier/Academic Press.

Yang, Y., Fu, W. J., & Land, K. C. (2004). A Methodological Comparison of Age-Period-Cohort Models: The Intrinsic Estimator and Conventional Generalized Linear Models. *Sociological Methodology, 34*(1), 75-110.

Yang, Y., & Land, K. C. (2008). Age–Period–Cohort Analysis of Repeated Cross-Section Surveys: Fixed or Random Effects? *Sociological Methods & Research, 36*(3), 297.

Yang, Y., Schulhofer-Wohl, S., Fu, W. J., & Land, K. C. (2008). The Intrinsic Estimator for Age-Period-Cohort Analysis: What It Is and How to Use It. *American Journal of Sociology, 113*(6), 1697-1736.

Appendix A. Market Shares of CRAs in the U.S. Structured Credit Markets

Table 6 shows the CRAs market shares in the U.S. structured credit markets. S&P has been the leader in U.S. structured credit ratings since 1997. Only in the first half of 2011, Moody's, which finished second place in market share in the U.S. structured credit markets for the last 13 years, overtook S&Ps leading position. A layer below Moody's and S&P, Fitch and DBRS are claiming the third and fourth place. Interesting to note is the change in ranking over time when only considering resident mortgage backed securities (RMBS). DBRS then climbs to second place.

Table 6

Rating-Agency Shares of U.S. Asset Backed Security (ABS) and Mortgage Backed Security (MBS) Issuance in the period 2009-2011

ABS/MBS	1H-2011 Issuance ($Mil.)	No. of Deals	Market Share (%)	2010 Issuance ($Mil.)	No. of Deals	Market Share (%)	2009 Issuance ($Mil.)	No. of Deals	Market Share (%)	2008 Issuance ($Mil.)	No. of Deals	Market Share (%)	2007 Issuance ($Mil.)	No. of Deals	Market Share (%)
S&P	61,386.9	77	64.0	148,327.3	264	73.9	138,953.5	191	72.3	160,365.1	239	91.4	992,842.8	1,435	95.7
Moody's	59,756.2	86	62.3	119,899.1	162	59.8	120,785.2	135	62.8	155,916.6	218	88.8	892,675.1	1,167	86.0
Fitch	52,136.8	63	54.3	87,169.6	117	43.5	117,002.9	170	60.8	117,512.6	156	67.0	598,636.1	743	57.7
DBRS	18,148.2	52	18.9	58,493.2	114	29.2	26,705.7	53	13.9	11,817.4	25	6.7	55,920.4	93	5.4
TOTAL	**95,969.6**	**162**	**100.0**	**200,601.7**	**373**	**100.0**	**192,296.4**	**306**	**100.0**	**175,520.2**	**289**	**100.0**	**1,037,449.3**	**1,576**	**100.0**
ABS															
S&P	55,316.9	73	70.6	106,194.4	196	74.7	118,850.0	144	82.3	139,655.6	178	92.5	579,667.8	892	96.9
Moody's	51,015.7	79	65.1	105,116.3	153	73.9	118,129.8	123	81.8	146,158.2	182	96.8	566,613.8	744	94.8
Fitch	46,417.3	56	59.2	72,440.1	100	51.0	87,543.2	89	60.6	106,379.2	118	70.4	344,362.0	424	57.6
DBRS	11,105.8	28	14.2	22,622.3	48	15.9	6,933.4	17	4.8	7,498.3	10	5.0	43,102.7	73	7.2
TOTAL	**78,390.7**	**128**	**100.0**	**142,177.9**	**262**	**100.0**	**144,421.7**	**180**	**100.0**	**151,029.3**	**207**	**100.0**	**597,991.1**	**992**	**100.0**
MBS															
S&P	8,740.5	7	49.7	42,132.8	68	72.1	20,103.5	47	42.0	20,709.5	61	84.6	413,175.0	543	94.0
DBRS	7,042.4	24	40.1	35,871.0	66	61.4	19,772.3	36	41.3	4,319.2	15	17.6	12,817.6	20	2.9
Moody's	6,070.0	4	34.5	14,782.9	9	25.3	2,655.4	12	5.5	9,758.4	36	39.8	326,061.3	423	74.2
Fitch	5,719.5	7	32.5	14,729.4	17	25.2	29,459.7	81	61.5	11,133.4	38	45.5	254,274.1	319	57.9
TOTAL	**17,579.0**	**34**	**100.0**	**58,423.8**	**111**	**100.0**	**47,874.7**	**126**	**100.0**	**24,490.9**	**82**	**100.0**	**439,458.2**	**584**	**100.0**

Note. From "Rating-Agency Shares of US ABS and MBS Issuance in the First Half" by Asset-Backed Alert, (2011a); "S&P Holds On as DBRS Surges" by Asset-Backed Alert, (2011b); and "Rating-Agency Shares Track Downward" by Asset-Backed Alert, (2009).

Appendix B. The Aggregator of Loans Backed by Assets (ALBA) Securitization Program

The dataset analyzed in this thesis is based on monthly updated loan level data from three United Kingdom non-conforming residential mortgage backed securities (RMBS): ALBA 2006-1, ALBA 2006-2 and ALBA 2007-1. The timespan of the data is from the period June 2008 to May 2010.

Non-conforming mortgages

RMBS have been introduced in chapter 1 and were further discussed in chapter 2. Non-conforming mortgages are a subcategory of residential mortgages. There is a broad range of lending where there is something about the borrower that does not fit the standard lending criteria expected by mortgage lenders. Non-conforming mortgages are traditionally offered by specialist lenders and originally developed for people who have experienced material and recent credit difficulties. As a result, the borrower is seen as a bigger risk by the lender. Therefore, non-conforming mortgages tend to have higher interest rates than standard loans and usually have higher charges and stricter conditions attached to them. Lenders adopt different approaches and there are no commonly agreed definitions or terminology, but broadly speaking we are talking about one or more of the following situations (CML Research, 2006):

- Borrowers having a poor credit record. A recent history of County Court Judgments (CCJs), loan arrears or defaults, rent arrears, decrees (in Scotland), bankruptcy, or individual voluntary agreements (IVAs) are the usual reasons for an adverse credit history.
- Borrowers having little or no credit record at all. For example, because they do not appear on the electoral roll, have had several home addresses in a short space of time, or have lived abroad until recently.
- Borrowers having an irregular or uncertain income, for example, because they are recently self-employed, have changed jobs frequently, have a job that relies heavily on commission income, or have more than one job. When borrowers cannot provide documentary evidence of their incomes, such loans will often be taken forward on a "self-certified" basis and lenders will employ a wide range of measures to guard against the fraudulent overstatement of incomes by the borrower. Lenders view self-cert mortgages in different ways, but many will treat those with LTVs of up to 75% as part of their mainstream business.

This categorization is visualized in Figure 11.

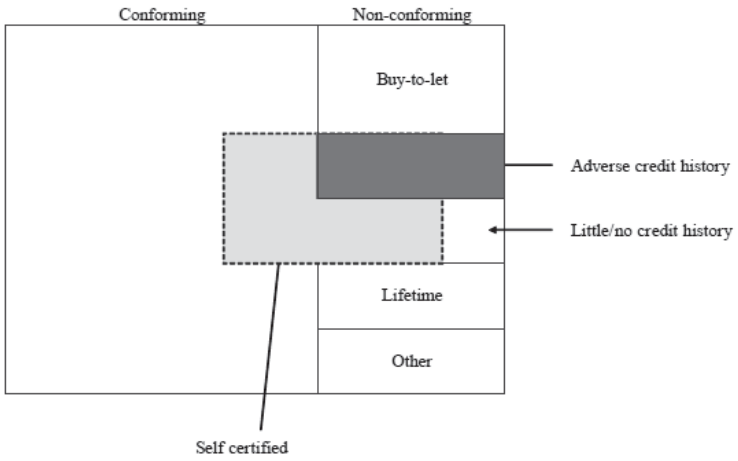


Figure 11: Conforming Versus Non-conforming Mortgages (CML Research, 2006).

Note that Figure 11 contextualizes the non-conforming part of the UK mortgage market, and is not intended to be a definitive sub-division of the market (CML Research, 2006).

ALBA 2006-1, ALBA 2006-2 and ALBA 2007-1

The three analyzed mortgage backed securities are from the Aggregator of Loans Backed by Assets (ALBA) program. The securities analyzed are the ALBA 2006-1, ALBA 2006-2 and ALBA 2007-1 deals.

ALBA is the securitization platform from Oakwood Homeloans Limited (OHL). OHL was created by Jason Miller in 2004 to acquire and trade portfolios of residential mortgage assets in the UK. The platform is used to source whole loan portfolios for securitization. Between January 2005 and May 2007 OHL acquired thirteen portfolios of residential mortgages from six different buyers (Amber Homeloans; Preferred Mortgages; GMAC-RFC, a wholly owned subsidiary of Minneapolis-based Residential Capital Corporation, in turn part of General Motors Acceptance Corporation; Kensington Group; and Edeus). The total value of OHL's acquisitions over this period was over £3 billion. In June 2006 Oakwood sold a majority stake (51%) in OHL to Credit Suisse London Branch.

The ALBA platform was launched by OHL in November 2005 with ALBA 2005-1. The first ALBA transaction featured loans originated by Platform Homes & Preferred Mortgages Limited.

ALBA 2006-1, the second securitization of OHL, was completed on June 16th, 2006. The collateral included in ALBA 2006-1 was originated by GMAC-RMC Limited and Kensington Mortgage Company Limited. The collateral consists of first ranking mortgage loans secured on residential properties in England, Wales, Scotland, and Northern Ireland. The £556 million securitization comprised around £349 million of near-prime residential mortgage assets, including some prime Buy-to-Let, from GMAC-RFC, and approximately £207 million of near-prime and light adverse assets from Kensington Mortgage Company, which were acquired using financing from Credit Suisse. Near-prime mortgages can be categorized between mainstream prime and non-conforming mortgages, and relate to borrowers whose recent credit problems have been minimal or where material problems were several years before the actual mortgage application.

ALBA 2006-2 is the third in the ALBA series of near-prime, non-conforming, and impaired credit mortgage securitizations from OHL and closed in November 2006. The collateral consists of residential mortgage loans originated by GMAC-RFC Limited, Kensington Mortgage Company Limited, and Money Partners Limited. The collateral consists of first ranking mortgage loans secured on residential properties in England and Wales. Of the loans, 44.85% were granted for re-mortgage purposes and of the obligors some 7.6% have been subject to a CCJ at some time (ALBA 2005-1, CCJs 7.4%). Self-certified mortgages make up just over 55% of the loan pool. Geographically, London and the South East dominate (40.90%) with the other three main regions being the North West (12.02%), South West (9.81%), and West Midlands (9.06%).

ALBA 2007-1 is the fourth issue from ALBA's asset-backed program. The structural features of this latest transaction closely resemble those of ALBA 2006-2. The mortgage loans have been purchased by OHL from originator GMAC-RFC Limited. This transaction is a securitization of near-prime and non-conforming residential mortgages originated and located in England, Wales and Scotland. Some 69.02% of the pool for this transaction by value corresponds to GMAC's prime and near-prime products. 6.01% of the borrowers have been the subject of at least one county court judgment, but only 0.60% of the loans are to borrowers who have been subject to a bankruptcy order or individual voluntary arrangement. 37.28% of the loans are to borrowers who self-certified their income stated on

their applications and 19.35% of the loans are for buy-to-let purposes. The majority of the notes were placed with investors in the UK with some interest also coming from Germany.

Table 7
ALBA Program Summary Statistics

	ALBA 2006-1	**ALBA 2006-2**	**ALBA 2007-1**
Closing Date	16 Jun 2006	17 Nov 2006	18 Jun 2007
Total Deal Size(Mil £)	556	538	975
Number of Loans	4172	4106	7018
Percentage of Buy to Let Loans	7.29%	17.24%	19.35%
Percentage of Self Certified Loans	69.71%	55.75%	37.28%
Weighted Average Original Loan To Value (%)	79.88%	82.25%	84.07%

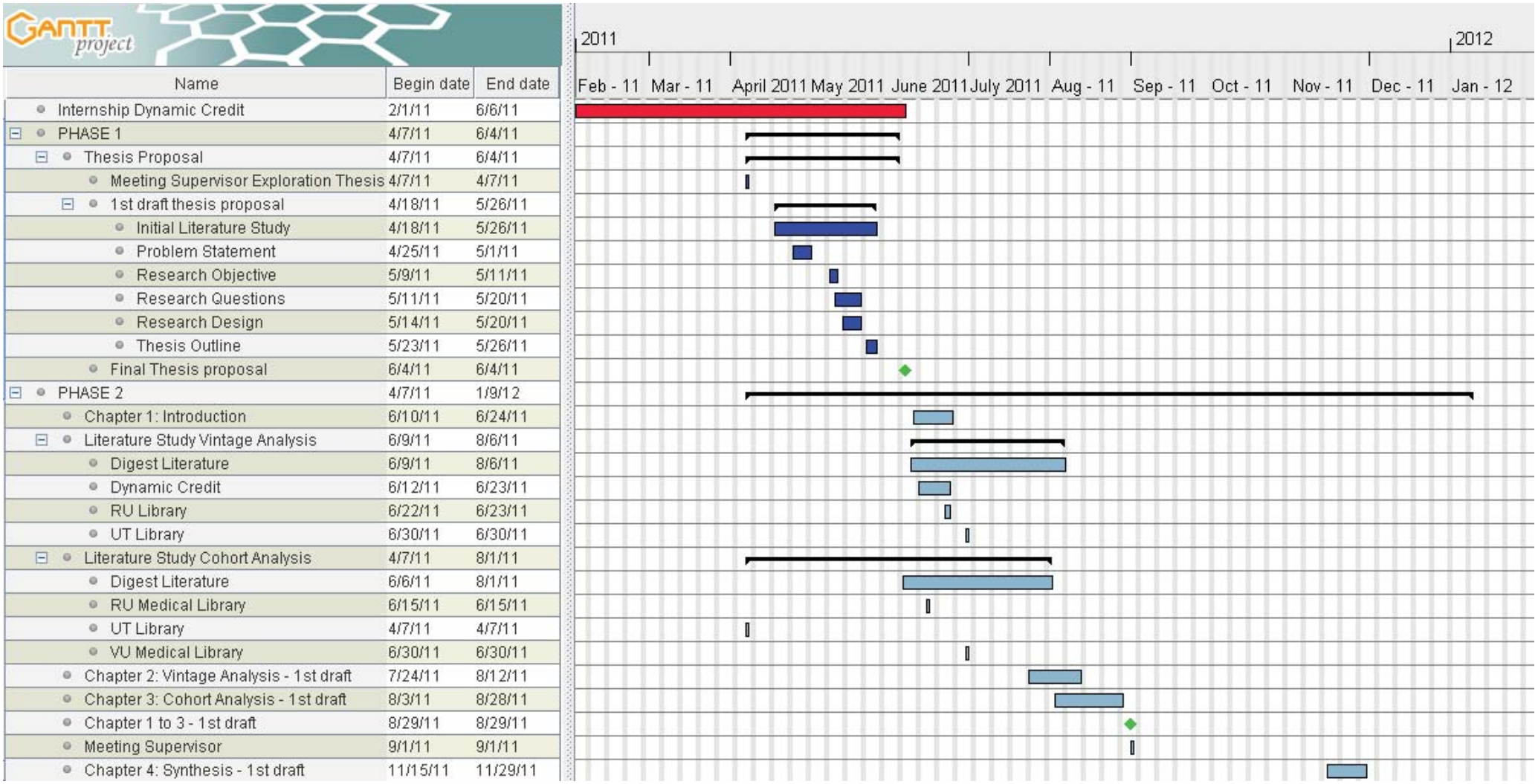
Appendix C. R Code - Cohort Analysis ALBA Data

```
#########################################################################
# Cohort analysis of ALBA mortgage data with the Epi package for        #
# epidemiological analysis in R                                         #
#                                                                       #
# Merijn Bosman 2011-11-01                                              #
#                                                                       #
# Epi package version 1.1.24 (2011-07-19) by Bendix Carstensen, Martyn  #
# Plummer, Esa Laara, Michael Hills et. al.                             #
#########################################################################

# 1. Load the Epi package
library( Epi )

# 2. The ALBA data is imported
alba <- read.table( "alba.csv", sep=";", header=TRUE)

# 3. Check whether the ALBA data was imported correctly
str( alba )

# 4. Attach the data to a dataframe to make objects within the dataframe
# easier to access
attach( alba )

# 5. Convert the date columns from factors to dates
as.Date(P,"%Y-%m-%d")
as.Date(C,"%Y-%m-%d")

# 6. Create simple tables from the loan data
table( A )
table( P )
table( C )

# 7. Create period-cohort tables from the loan data
O_table_nice <- stat.table( index=list( C,P ), sum( O ), data=alba,
margin=TRUE )
print( O_table_nice, digits=c( sum=0 ))
N_table_nice <- stat.table( index=list( C,P ), sum( N ), data=alba,
margin=TRUE )
print( N_table_nice, digits=c( sum=0 ))

# 8. Creates rates from the tables
O_table <- tapply( O, list( C,P ), sum )
N_table <- tapply( N, list( C,P ), sum )
R_table <- O_table/N_table
print( N_table, digits=c( sum=2 ))

# 9. Fit different generalized linear models for comparison
#1. Age model
m.A <- glm( O ~ factor( A )+ offset( log( N ) ), family=poisson, data=alba)
#2. Age-Period model
m.AP <- glm( O ~ factor ( A ) + factor ( P ) + offset( log( N ) ),
family=poisson, data=alba )
#3. Age-Cohort model
m.AC <- glm( O ~ factor ( A ) + factor ( C ) + offset( log( N ) ),
family=poisson, data=alba )
#4. Age-drift model
m.Ad <- glm( O ~ factor ( A ) + P + offset( log( N ) ), family=poisson,
data=alba )
#5. Age-period-cohort model
```

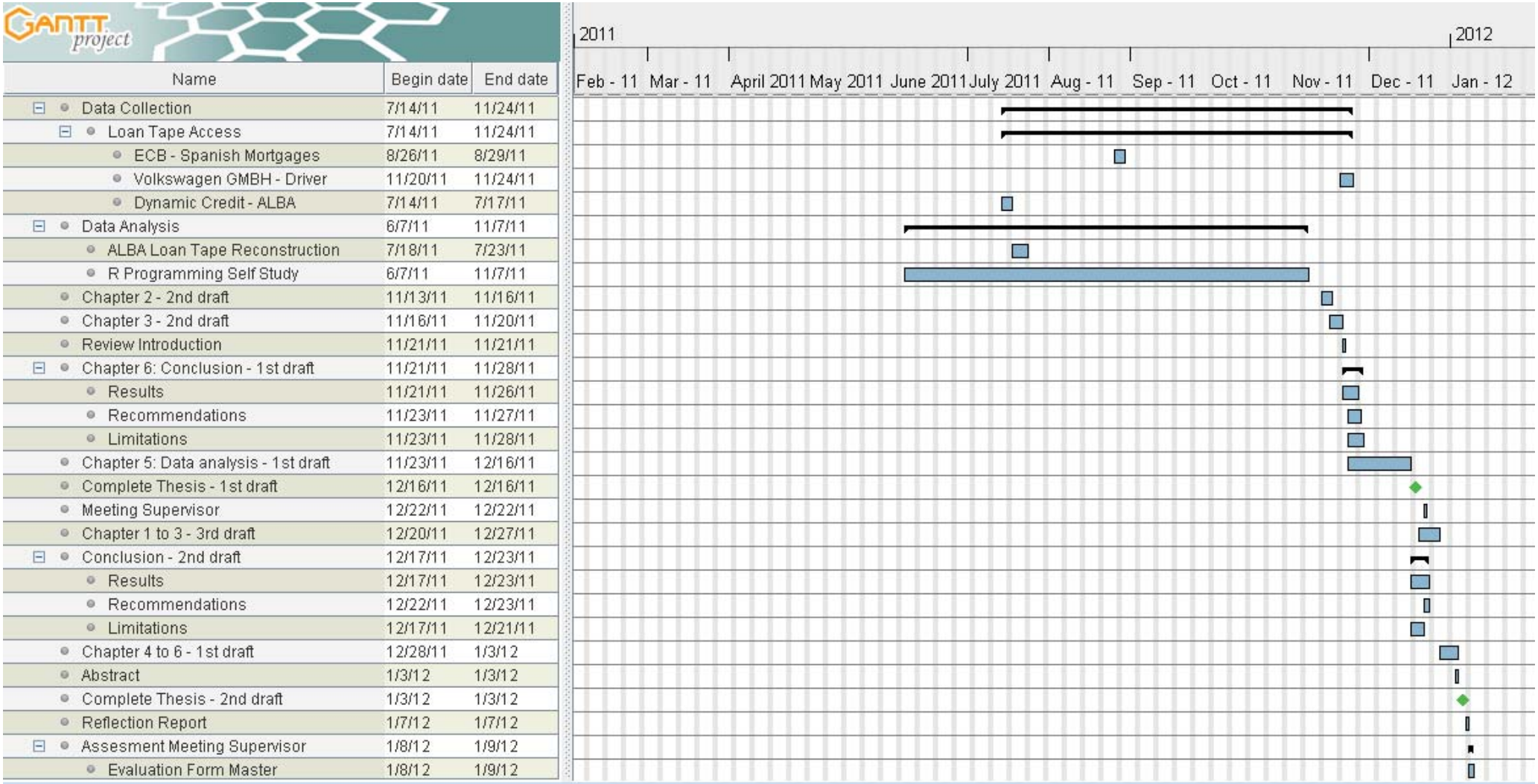
```
m.APC <- glm( O ~ factor( A ) + factor( P ) + factor( C ) + offset( log( N
) ), family=poisson, data=alba )

# 10. Test the different models by comparing their deviance
# The successive tests refer to:
#   1. Linear effect of period/cohort
#   2. Non-linear effect of period
#   3. Non-linear effect of cohort (in the presence of period)
#   4. Non-linear effect of period (in the presence of cohort)
#   5. Non-linear effect of cohort
anova( m.A, m.Ad, m.AP, m.APC, m.AC, m.Ad, test="Chisq" )

# 11. Fit models where some of the factor levels are merged or sorted as
the first one
alba$Pr <- Relevel(factor(alba$P), list("first-last"=c("2008-06-30","2010-
05-30") ) )
alba$Cr <- Relevel(factor(alba$C), "2006-08-30")
with(alba, table (P,Pr))
with(alba, table(C,Cr))
m.APC1 <- glm(O~-1 + factor(A) + factor(Pr) + factor(Cr) + offset (log(N)),
family=poisson, data=alba)
m.APC1$coef

# 12. Extract the age, period and cohort parameters with confidence limits
from the cohort model
ci.lin( m.APC1, subset="A", Exp=TRUE, alpha=0.1)[,5:7]
A.eff <- ci.lin( m.APC1, subset="A", Exp=TRUE, alpha=0.1)[,5:7]
rbind( c(1,1,1), ci.lin( m.APC1, subset="P", Exp=TRUE, alpha=0.1)[,5:7],
c(1,1,1) )
P.eff <- rbind( c(1,1,1), ci.lin( m.APC1, subset="P", Exp=TRUE,
alpha=0.1)[,5:7], c(1,1,1) )
C.ref <- match ("2006-08-30", levels( with (alba, factor(C)) ) )
rbind( c(1,1,1), ci.lin( m.APC1, subset="C", Exp=TRUE, alpha=0.1)[,5:7]
)[c(2:C.ref,1,C.ref:(nlevels(alba$Cr)-1)),]
C.eff <- rbind( c(1,1,1), ci.lin( m.APC1, subset="C", Exp=TRUE,
alpha=0.1)[,5:7] )[c(2:C.ref,1,C.ref:(nlevels(alba$Cr)-1)),]

# 13. Plot the parameters with confidence limits
A.pt <- sort( unique ( alba$A ) )
P.pt <- sort( unique ( alba$P ) )
C.pt <- sort( unique ( alba$C ) )
par( mfrow=c(1,3), las=2 )
matplot ( A.pt, A.eff, xlab="Age", ylab="Rates", log="y", type="l", lty=1,
lwd=c(3,1,1), col="black" )
matplot ( as.Date(P.pt), P.eff, xlab="Period", xaxt="n", ylab="Arrears
Rate", log="y", type="l", lty=1, lwd=c(3,1,1), col="black" )
datelabels <- format(as.Date(P.pt), "%m-%y")
axis(1,at=as.Date(P.pt),labels=datelabels)
abline ( h=1)
matplot ( as.Date(C.pt), C.eff, xlab="Cohort", xaxt="n", ylab="Arrears
Rate", log="y", type="l", lty=1, lwd=c(3,1,1), col="black" )
datelabels <- format(as.Date(C.pt), "%m-%y")
axis(1,at=as.Date(C.pt),labels=datelabels)
abline ( h=1 )

# 14. Rescale the rates to make the plot more informative
par( mfrow=c(1,3), las=2 )
matplot ( A.pt, A.eff, xlab="Age", ylab="Rates", ylim=c(0.05,5), log="y",
type="l", lty=1, lwd=c(3,1,1), col="black" )
matplot ( as.Date(P.pt), P.eff, xlab="Period", xaxt="n", ylab="Arrears
Rate", ylim=c(0.05,5), log="y", type="l", lty=1, lwd=c(3,1,1), col="black")
```

```
datelabels <- format(as.Date(P.pt), "%m-%y")
axis(1,at=as.Date(P.pt),labels=datelabels)
abline ( h=1)
matplot ( as.Date(C.pt), C.eff, xlab="Cohort", xaxt="n", ylab="Arrears
Rate", ylim=c(0.05,5), log="y", type="l", lty=1, lwd=c(3,1,1), col="black")
datelabels <- format(as.Date(C.pt), "%m-%y")
axis(1,at=as.Date(C.pt),labels=datelabels)
abline ( h=1 )
```

Appendix D. Schedule Master Thesis



Name	Begin date	End date
Internship Dynamic Credit	2/1/11	6/6/11
PHASE 1	4/7/11	6/4/11
Thesis Proposal	4/7/11	6/4/11
Meeting Supervisor Exploration Thesis	4/7/11	4/7/11
1st draft thesis proposal	4/18/11	5/26/11
Initial Literature Study	4/18/11	5/26/11
Problem Statement	4/25/11	5/1/11
Research Objective	5/9/11	5/11/11
Research Questions	5/11/11	5/20/11
Research Design	5/14/11	5/20/11
Thesis Outline	5/23/11	5/26/11
Final Thesis proposal	6/4/11	6/4/11
PHASE 2	4/7/11	1/9/12
Chapter 1: Introduction	6/10/11	6/24/11
Literature Study Vintage Analysis	6/9/11	8/6/11
Digest Literature	6/9/11	8/6/11
Dynamic Credit	6/12/11	6/23/11
RU Library	6/22/11	6/23/11
UT Library	6/30/11	6/30/11
Literature Study Cohort Analysis	4/7/11	8/1/11
Digest Literature	6/6/11	8/1/11
RU Medical Library	6/15/11	6/15/11
UT Library	4/7/11	4/7/11
VU Medical Library	6/30/11	6/30/11
Chapter 2: Vintage Analysis - 1st draft	7/24/11	8/12/11
Chapter 3: Cohort Analysis - 1st draft	8/3/11	8/28/11
Chapter 1 to 3 - 1st draft	8/29/11	8/29/11
Meeting Supervisor	9/1/11	9/1/11
Chapter 4: Synthesis - 1st draft	11/15/11	11/29/11



Name	Begin date	End date
Data Collection	7/14/11	11/24/11
Loan Tape Access	7/14/11	11/24/11
ECB - Spanish Mortgages	8/26/11	8/29/11
Volkswagen GMBH - Driver	11/20/11	11/24/11
Dynamic Credit - ALBA	7/14/11	7/17/11
Data Analysis	6/7/11	11/7/11
ALBA Loan Tape Reconstruction	7/18/11	7/23/11
R Programming Self Study	6/7/11	11/7/11
Chapter 2 - 2nd draft	11/13/11	11/16/11
Chapter 3 - 2nd draft	11/16/11	11/20/11
Review Introduction	11/21/11	11/21/11
Chapter 6: Conclusion - 1st draft	11/21/11	11/28/11
Results	11/21/11	11/26/11
Recommendations	11/23/11	11/27/11
Limitations	11/23/11	11/28/11
Chapter 5: Data analysis - 1st draft	11/23/11	12/16/11
Complete Thesis - 1st draft	12/16/11	12/16/11
Meeting Supervisor	12/22/11	12/22/11
Chapter 1 to 3 - 3rd draft	12/20/11	12/27/11
Conclusion - 2nd draft	12/17/11	12/23/11
Results	12/17/11	12/23/11
Recommendations	12/22/11	12/23/11
Limitations	12/17/11	12/21/11
Chapter 4 to 6 - 1st draft	12/28/11	1/3/12
Abstract	1/3/12	1/3/12
Complete Thesis - 2nd draft	1/3/12	1/3/12
Reflection Report	1/7/12	1/7/12
Assesment Meeting Supervisor	1/8/12	1/9/12
Evaluation Form Master	1/8/12	1/9/12


GANTT project
Name
Begin date
End date
Announcement Colloquium 1/8/12 1/8/12
PHASE 3 1/10/12 1/29/12
Final Version Master Thesis 1/10/12 1/22/12
Preparation Viva Voce 1/23/12 1/29/12
Review Thesis 1/23/12 1/28/12
Create Presentation 1/25/12 1/29/12
PHASE 4 1/30/12 1/30/12
Viva Voce 1/30/12 1/30/12
2011
2012
Feb - 11
Mar - 11
April 2011
May 2011
June 2011
July 2011
Aug - 11
Sep - 11
Oct - 11
Nov - 11
Dec - 11
Jan - 12

Appendix E. Reflection Report Master Thesis

Personal learning objectives

The personal learning objectives I established before starting this thesis were:

1. Learn how to establish relationships between seemingly unconnected subjects in an academic way
2. Learn how to program in R
3. Adopt a helicopter perspective with respect to structured credit valuation and rating.

In the next section I will elaborate and reflect upon these learning objectives.

Degree to which the learning objectives are achieved

The first leaning objective relates to the research design of the thesis. For my bachelor thesis, I opted for a replicative study. I noticed that the amount of share buybacks in the Netherlands grew in tenfold in the last decade and that the last study into this phenomenon in the Netherlands was conducted more than 10 years ago when this was a new phenomenon in the country. Literally thousands of studies are already conducted to investigate the returns before and around the date that a share buyback is announced to the public and I chose to investigate this in the Netherlands. In my eyes, a replicative study is a safe research design, because it is relatively easy to set well defined boundaries for the research. Also the outline of such a research is relatively straightforward. Although I consider this to be an interesting academic exercise and believe that replication is necessary if not critical for the development of all areas of science, I was a little disappointed after conducting the research. I attributed this to the nature of the research and made a mental note to myself to try a different research design for my master thesis.

This thesis is explorative in nature. I relate cohort analysis to vintage analysis and thereby establish relationships between structured credit, demography, sociology and epidemiology. I felt more satisfied after the literature study in the various areas of science than how I felt after writing my bachelor thesis, so in that respect I consider this learning objective to be reached. The dissatisfying part for me, however, was the stage after the literature study. Where I initially hoped to conduct a full-blown data analysis, I did not manage to complete this. The main factor in this was the outcome of this thesis (that more theory is needed, before it makes sense to conduct cohort based vintage analyses).

Looking back on the process of writing this thesis I feel that conducting an explorative for a master thesis is difficult. First, I, as a researcher, defined the boundaries of the research. Every new insight obtained has the potential to alter the direction of the research. Comparing this to a replicative study, such as the I did for my bachelor thesis, makes the research process difficult and time consuming. This made it also harder to stay focused on the end result, since the end result, as was the case in this research, can change multiple times. Second, investigating areas of science in which I had limited knowledge (sociology, demography and, even more difficult, epidemiology), required me put far more time than I had expected into conducting a proper literature study. Every article became a piece of a puzzle and only in the last weeks of writing my thesis I started to see the full image.

In sum, I feel that this learning objective has been reached at the cost of investing far more time and energy in the project than I initially planned.

The second learning objective, to learn how to program in R, is a personal interest that needed a fire starter, such as the writing of a thesis, to start off. R is known for its steeps learning curve. This can be

attributed to the open source nature of the language. There are many introductory tutorials, but not many are elaborate. Also, most of the tutorials are written from a viewpoint that is specific to the area of science the writer is working in. The result is that it requires more effort from the side of the R student. Besides the tutorials, R is in my eyes Spartan compared to other programs, such as SPSS. What I mean by that is that programs such as SPSS allow the user to conduct statistical analyses with the click of a few buttons (one only needs to understand the procedure to obtain output); R needs a command for every step the user wants to make. This implies that before you can do even the simplest analysis in R, you need to invest a considerable amount of time into the tutorials. The advantage is the power and flexibility the user gets in return.

I feel that the learning objective of learning R is partly reached. As was mentioned above in the discussion of my first learning objective, I did not manage to spend the amount of time on the data analysis that I had hoped for. Nonetheless, I have the feeling that I now know the basics. I have spent several weekends identifying useful tutorials, reading them, playing around with importing small datasets, preparing graphs and conduct simple statistical analysis. Also, I have created a list with "to read" materials. I believe that this make it easy to continue my study in R and in that respect I come to the conclusion that this learning objective has been partly reached.

The third learning objective, adopting a helicopter perspective with respect to structured credit valuation and rating, relates to my "pre-work" for my master thesis. I have accepted an internship at a well respected structured credit asset manager/advisor. During this internship I learned a lot about the details of specific structured credit product categories and the processes involved in the valuation and rating of these products. However, during my internship I did not have the time to adopt the helicopter perspective and see what similarities and differences between the different product categories exist and how these differences come about.

I feel that this learning objective has been reached. During my literature study, I have identified several high-quality sources that helped me to obtain the helicopter perspective. For example, the review of the structured credit rating methodologies and the work of Ashcraft et al (2011; 2010) that I used in chapter 2 of this thesis helped me to establish the relationships between the different rating methodologies and allowed me to see the differences and similarities between them. Next to that, the process of the literature study made me feel more confident to explain the differences between the product categories and I consider myself to be better able to identify the main assumptions in valuation and rating approaches. Therefore, I come to the conclusion that this learning objective has been reached.