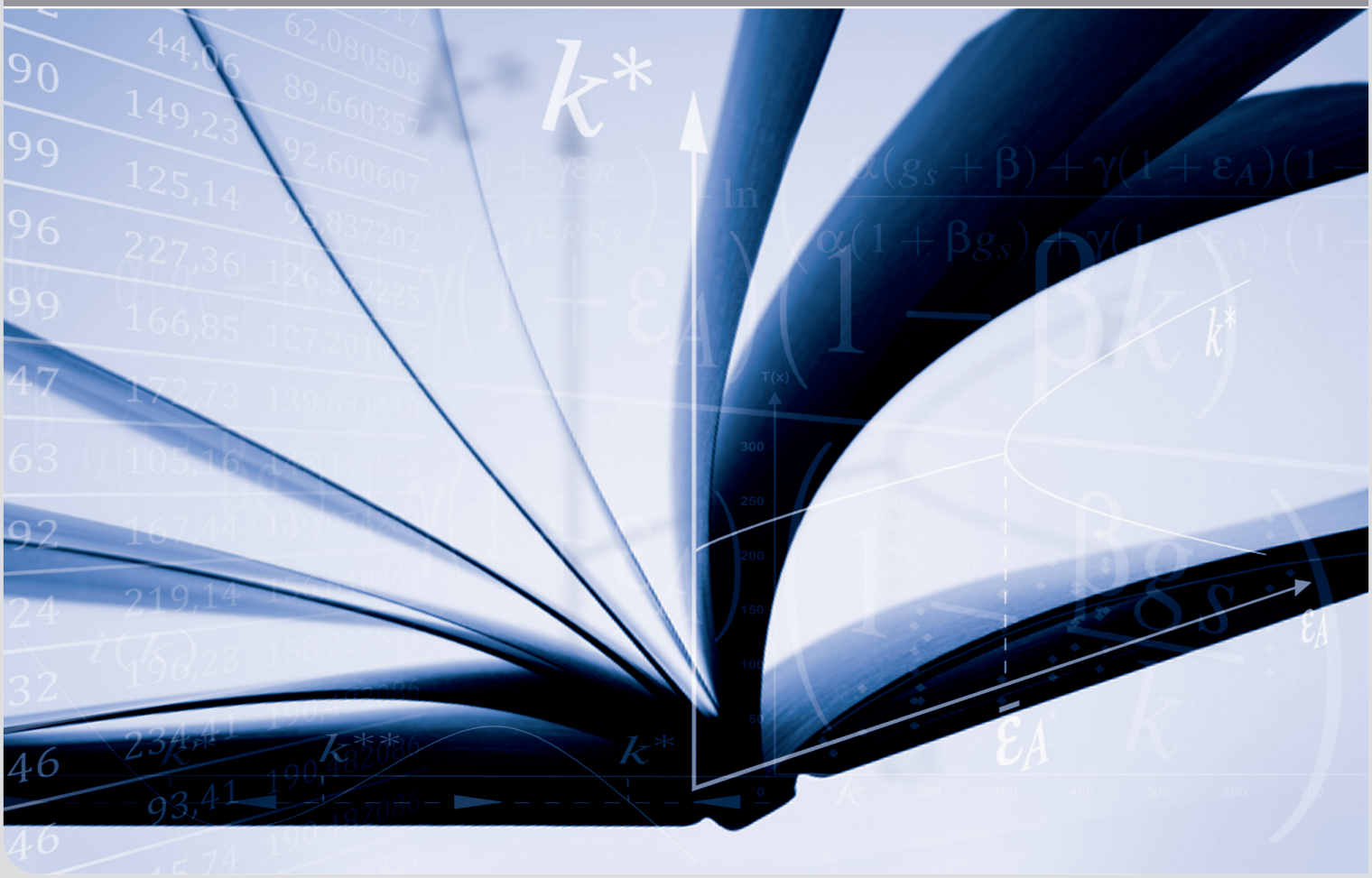
Karlsruher Institut für Technologie

Epidemiological spreading of mortgage default

by Jochen Schweikert, Markus Höchstötter

No. 112 | JANUARY 2018

WORKING PAPER SERIES IN ECONOMICS



Impressum

Karlsruher Institut für Technologie (KIT)
Fakultät für Wirtschaftswissenschaften
Institut für Volkswirtschaftslehre (ECON)

Kaiserstraße 12
76131 Karlsruhe

KIT – Die Forschungsuniversität in der Helmholtz-Gemeinschaft

Working Paper Series in Economics
No. 112, January 2018

ISSN 2190-9806

econpapers.wiwi.kit.edu

Epidemiological Spreading Of Mortgage Default

Jochen Schweikert

Lehrstuhl für Ökonometrie und Statistik

Schlossbezirk 12

D-76128 Karlsruhe

Dr. Markus Höchstötter

Lehrstuhl für Ökonometrie und Statistik

Schlossbezirk 12

D-76128 Karlsruhe

Karlsruhe, June 23, 2017

Abstract

This paper introduces mathematical models to capture the spreading of epidemics to explain the expansion of mortgage default events in the United States. Here we use the state of infectiousness and death to represent the subsequent steps of payment delinquency and default, respectively. Since the local economic structure influences regional unemployment that is a strong driver of mortgage default, we model interdependencies of regional mortgage default rates through employment conditions as well as vicinity. Based on a large sample between 2000 and 2014 of loan-level data, the estimation of key parameters of the model is proposed. The model's forecast accuracy shows an above average performance compared to well-known approaches like linear regression or logit models. The key findings may be useful in understanding the dynamics of mortgage defaults and its spatial spreading.

Keywords

Mortgage default; Epidemics; Spatial spreading

JEL classifications

R30

We thank Melanie Schienle and Monika Piazzesi for their helpful comments.

1 Introduction

The recent credit crisis of 2007 has resulted in a rapid decline of building prices and consequently of mortgages values due to increased mortgage default risk. The negative impacts not only influenced the U.S. economy, but resulted in a worldwide recession. Therefore the understanding of how mortgages become infected by each other and consequently delinquent and default consecutively, has been neglected so far to the best of our knowledge. Credit risk is highly influenced by the dependence between defaults. Researchers concentrated on the individual default probabilities, but a pool of mortgage loans and derivatives written on those mortgages are riskier if the defaults appear at similar dates. Therefore default dependence is a main concern to risk management or pricing of mortgaged backed securities. Important research has been conducted on simultaneous default such as, for example, in Duffie et al. (2009). However, only a few studies, for example Cowan and Cowan (2004) and Hillebrand et al. (2012) consider dependencies in the mortgage market by serial correlation.

The failure of models to capture the dynamics and interdependencies of mortgage performance reveals the need for new models based on a deeper understanding of mortgage default characteristics. To our knowledge, no model has been suggested, yet, that explicitly explains the spreading of mortgage default. Our approach is to capture the dynamics of mortgage default in terms of a compartment model for epidemic diseases.

There is some evidence that mortgage defaults are influenced by payment difficulties of debtors living in the surrounding area. Like the infection of a disease, Goodstein et al. (2011) show that default rates increase after the information that another mortgage within the neighbourhood defaulted started to spread. They explain this behaviour through fallen psychological barriers. For example Chan et al. (2013) observe the same effect and can show that an increase of foreclosure nearby result in decreasing house prices and increasing default probabilities. Therefore the use of compartment models normally capturing the course of a disease seems appropriate to explain mortgage default spreads.

In this paper we analyse the default dependence between 2000 and 2014 within a large data set of individual U.S. mortgages. The data include both pre-crisis and post-crisis times.

First, we propose a model that compares the states of mortgage contracts during their life cycles to a classical Susceptible-Infected-Recovered model (SIR model) first described by Ker-

mack and McKendrick (1927). The SIR model originally divides the population in different subgroups, called compartments, depending on if they are infected. We assign the basic ideas of the SIR model to capture the dynamics of the local mortgage market inside a U.S. county. Second, we determine rates that measure the amount of mortgages which change their states of infection through statistical methods for survival analysis. The Cox model is used to describe the time how long mortgage contracts stay in a specific compartment (see Cox, 1972). Since macroeconomic variables are strong drivers of mortgage defaults, we include unemployment rates, house prices, the mean credit score (FICO) within a county and the spread between the average mortgage rate within the county and the national mean. These variables are well-known factors influencing mortgage payments (see Elul et al., 2010; Divino and Rocha, 2013; Danis and Pennington-Cross, 2008).

Third, we model the interdependencies between several local mortgage markets that are far apart through industrial similarities proposed by Feser et al. (2005) because there is evidence that regional unemployment rates are influenced by the economic structure inside the area (Weiler, 2001).

The question we address is whether epidemic models are able to determine the dynamics of mortgage markets and to which extent the dependence of mortgage default rates can be measured through the proposed approach of industrial similarities among several counties. In addition, we examine how accurate the model predicts real data and compare the results to earlier research.

The paper is organized as follows. We shortly introduce the classical SIR-model in section 2, its similarity to the possible states of mortgage loans and statistical models for survival analysis. Then we discuss housing prices, unemployment rates, mortgage rates and their influence on the default probability of individual mortgages, as well as the geographic measure of economic structures in section 3. After the theoretical setup in section 4, we present the data used in our work in chapter 5. After section 6 presents the empirical results the final section concludes the paper.

2 Epidemic models

Epidemiological investigation has a long history in research dating back many centuries, however, the mathematical models were introduced in a series of seminal papers by Kermack and McKendrick in the 1920s, see for example Matthews and Woolhouse (2005).

Today, models with multiple facets and in a variety of complexities exist. The models may be categorized as deterministic or stochastic depending on their specification. Brauer et al. (2008) provide a broad overview on mathematical models for disease analysis. Deterministic, or compartmental models, try to capture the dynamics on a large scale. They are therefore very suitable for investigating the average evolution of a total population, whose individuals are categorized into different subgroups, called compartments, according to their status (Matthews and Woolhouse, 2005).

The basic framework of the majority of the epidemiological models is formed by the deterministic SIR model which assumes that people are homogneous, i.e. each individual has the same likelihood to become infected if exposed and is expected to experience the same severity of infection. Further, the population is assumed to be well-mixed, which yields equal exposure within each subpopulation. These assumptions allow to employ the mean-field methods often deployed in physics to derive the results. Mean field theory assumes strict homogeneity. The following introduction is based on the description in Keeling and Eames (2005) and Britton (2010), but may be found elsewhere in slightly different notation. The name SIR is derived from the possible model states. $\{S(t), t \geq 0\}$ denotes the amount of susceptible people at time t, i.e. the number of people that are endangered to become infected, $\{I(t), t \geq 0\}$ describes the infectious or infected part of the population, in other words, those that are able to spread the disease and $\{R(t), t \geq 0\}$ denotes the group that has recovered from the disease. As an extension the natural death of a person is usually modelled as a state called $\{D(t), t \geq 0\}$ which means that the total population alive at time t, $\{N(t), t \geq 0\}$, is estimated through $N(t) = S(t) + I(t) + R(t), \forall t$.

Exchanges between the compartments in a deterministic model are driven by transition rates. People are born with a rate of μ and die with a rate of τ per unit of time. A person is able to die no matter in which compartment he is in. All susceptible people are infected with the force of infection β and recover at rate γ. Transition is the only possible way from S to I

and from I to R. Let all rates be denominated as number of people per unit of time. If one assumes born children to be healthy, the model dynamics may be represented through a set of differential equations as in Brauer et al. (2008):

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \mu \cdot N(t) - \beta \cdot S(t) \cdot I(t) - \tau \cdot S(t) \tag{1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t) - \tau \cdot I(t) \tag{2}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma \cdot I(t) - \tau \cdot R(t) \tag{3}$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = \tau \cdot S(t) + \tau \cdot I(t) + \tau \cdot R(t) \tag{4}$$

In equation 1 the dynamics of the susceptible people is described by a constant birth rate $\mu \cdot N(t)$ that increases the amount of state $S(t)$. With rate $\beta \cdot S(t) \cdot I(t)$ people get infected and die with rate $\tau \cdot S(t)$ that both decrease the amount of susceptibles. Equation 2 describes the amount of the infected population where $\beta \cdot S(t) \cdot I(t)$ newly infected people join this state at every time span. They recover at rate $\gamma \cdot I(t)$ which increases the amount of recovered people in equation 3, then. Infected and recovered people die at rate $\tau \cdot I(t)$ and $\tau \cdot R(t)$. Equation 4 captures all deaths in every period and is the sum of all people dying while being in one of the other states.

Burnside et al. (2016) adapt the epidemiological approach to the housing market. They model the dynamics of the fraction of agents with different views about future house prices. A few agents changing their expectations through random social interactions can result in a housing market boom. Their work is based on Piazzesi and Schneider (2009) who show that a small amount of investors that are optimistic about future house prices is sufficient to cause a price increase.

Statistical models used for survival analysis were developed in order to analyze the time-to-occurence of a certain event as well as the circumstances related to that event. Survival analysis approaches are used to study time-to-failure characteristics of machine components, time-to-death in clinical studies, or in the estimation of the incubation time of diseases. For a deeper understanding of survival analysis, we refer to Elandt-Johnson and Johnson (1999), Lawless (2011), as well as Kalbfleisch and Prentice (2011).

Since we are interested in modelling how long mortgage contracts stay in a specific state and when they leave to another one according to our model, the transition rates are considered in terms of survival analysis approaches.

The essential term in survival analysis we use is the hazard function, or hazard rate, $h(t)$. In contrast to the unconditional survival function, the hazard function is based on the conditional probability, of observing the event of interest in the next small time step $[t, t + \Delta t]$, given that the event has not yet happened by time t. The hazard rate is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}\left[t \leq T < t + \Delta t \mid T \geq t\right] \tag{5}$$

In order to capture explanatory variables, Cox (1972) proposed a semi-parametric model for the hazard function, whereas he defines the hazard function as

$$h(t, x) = h_0(t) \cdot \exp\left[x_1\beta_1 + ... + x_p\beta_p\right] = h_0(t) \cdot \exp\left[x^T\beta\right] \tag{6}$$

with β being the vector of coefficients and $h_0(t)$ being the baseline hazard, the hazard function with all covariates being equal to zero. Since the baseline hazard is not of parametric form, the model is called to be a semi-parametric model. Because of its flexibility to handle time-dependent parameters, as well as frailty terms, the Cox model is the most popular approach for modelling the relation between covariates and censored data.

There are a lot of studies involving Cox models to describe the time till a mortgage defaults (see e.g. Deng et al., 1996; Quercia and Stegman, 1992). The main focus is to estimate the influence of loan-level characteristics and macroeconomic factors on the time between mortgage origination and termination.

3 Drivers of mortgage default

This section briefly summarizes some of the various drivers of mortgage default used in our model that are investigated in earlier research.

It has already been proved that a reduced unemployment rate is followed by a falling probability of mortgage default. This is due to an increased likelihood that the monthly mortgage payments can't be served if the borrower gets unemployed. It can even be seen that the default rate increases only if the unemployment rate rises (see for example Elul et al., 2010).

A second driver that is considered here is the house price that is negative correlated to the frequency of defaults (Danis and Pennington-Cross, 2008). If the house price declines over the lifetime of the mortgage contract the difference between it and the outstanding loan amount gets negative which is a reason for the borrower to default (Elul et al., 2010).

When a mortgage is originated, the lender offers a mortgage rate to the borrower which is influenced by the interest rate level of the financial market and the borrower characteristics to capture the borrowers creditworthiness. Divino and Rocha (2013) imply an increased default probability after the interest rate drops, because the contract can be switched to one with lower interest rate.

The area around the real estate bought by the mortgage borrower also influences the default probability. Chan et al. (2013) show a negative influence of foreclosures in the neighbourhood due to dropping house prices. If a mortgage in the neighbourhood defaults the stigma of an own default isn't seen as tragic as before which increases the default rate due to a psychological reasons (Goodstein et al., 2011).

The history of loan payments is a strong driver to predict the ability to pay loan payments in the future. In the US the creditworthiness is captured by a single number called FICO-score (Danis and Pennington-Cross, 2008). It has been shown that mortgage contracts whose borrower has got a high score are less likely to default (Bajari et al., 2013).

4 Theoretical setup

To compare the epidemic model to the mortgage market, the life cycle of a mortgage contract needs to be observed.

Initially a mortgage is originated, its payments are assigned through the mortgage contract and payments can be assumed to be on time. The mortgage is seen as susceptible (state S) and called current. If the payments are delinquent so that the payments are at least one month behind schedule the mortgage is seen as infected (state I). Then there are three possibilities that can happen. First, if the debtor is in delay on multiple monthly payments, the contract is considered defaulted by the lender, and, in terms of an epidemic model, dead due to the infection (state D). Second, the borrower can sell the real estate, pay back the outstanding debt and therefore leave the model in a recovered state (state R). Third, the borrower pays back all outstanding payments so that the mortgage contract returns to state S and is seen as susceptible again the next point of time.

Before the epidemic mortgage market model is introduced there are some assumptions that are made both to simplify the theoretical structure and to adjust properties of the mortgage market to the epidemiological model.

The mortgage market is modelled as a state space model and the parameters are estimated by an iterated filtering algorithm proposed by Ionides et al. (2006). Further details on state space models can be found for example in Robert H. Shumway and Stoffer (2011).

We assume a homogenous credit pool, which means that the amount of outstanding debt, mortgage rate and monthly mortgage payments are assumed to be equal for all mortgages. U.S. counties are the geographical unities since they are "the largest territorial division for local government within a state of the United States"' (Merriam-Webster, 2016). A lot of studies use ZIP-Codes to include spatial effects in its models which is crucial according to Grubesic (2008). To eliminate data issues, only counties are considered where more than 50 mortgages are originated between 2000 and 2014 according to the used data introduced in the next section.

Furthermore, a default of a mortgage is only possible if the loan has been delinquent before and a defaulted mortgage leaves the model and isn't considered for the rest of the observed timespan. A mortgage is able to be prepaid, e.g. the outstanding debt is fully paid before

termination, both in a delinquent state and in a state where the monthly payments are on schedule.

During the whole life cycle, a mortgage isn't considered to be in more than one state. The mortgage's payments are either on schedule or delayed, but not both. The cause of termination is considered to be unique, too. Either the mortgage's outstanding debt is fully repaid or the debtor fails to make payments multiple times so that the contract is considered defaulted by the lender.

The parameter estimation of the state space model requires the definition of five compartments:

Compartment B: Equivalent to the level of infectious bacteria in Bertuzzo et al. (2010), this state corresponds to an unobserved system state that tries to include the mortgage dynamics not captured by explanatory variables, i.e. the systematic distress in the given local mortgage market

Compartment S: Pool of performing (current) mortgages which payments are on schedule

Compartment I: Pool of delinquent mortgages which payments are at least one month behind schedule

Compartment P: Pool of prepaid mortgages, hence the balance has been prepaid

Compartment D: Pool of defaulted mortgages, that means the mortgagor seized and did not resume servicing until property was repossessed and foreclosed

The general model for the US, i.e., several counties, with discrete equidistant time increments is given by the following system of differential equations. In order to capture the dynamics, we formulate an adapted version of the compartment model for each county i in the form

$$\frac{\partial B_i(t)}{\partial t} = -\mu_{B_i} \cdot B_i(t) + \lambda_i(t) + \sum_{k=1}^{N} \omega_{i,k} \cdot \delta_{i,k} \cdot \phi_{i,k}(B_k(t)) \tag{7}$$

$$\frac{\partial S_i(t)}{\partial t} = \mu_i - \beta_i \cdot B_i(t) \cdot S_i(t) + \gamma_i(t) \cdot I_i(t) - \nu_i^S(t) \cdot S_i(t) \tag{8}$$

$$\frac{\partial I_i(t)}{\partial t} = \beta_i \cdot B_i(t) \cdot S_i(t) - \gamma_i(t) \cdot I_i(t) - \nu_i^I(t) \cdot I_i(t) - \alpha_i(t) \cdot I_i(t) \tag{9}$$

$$\frac{\partial P_i(t)}{\partial t} = \nu_i^I(t) \cdot I_i(t) + \nu_i^S(t) \cdot S_i(t) \tag{10}$$

$$\frac{\partial D_i(t)}{\partial t} = \alpha_i(t) \cdot I_i(t) \tag{11}$$

The model of Cox (1972) is used to describe transition rates as follows

$$\lambda_i(t) = a_{i0} \cdot (I_i^{Data} + dD_i^{Data}) \cdot e^{a_{i1}\cdot \mathrm{ALQ}_i(t) + a_{i2}\cdot \mathrm{HOUSE}_i(t) + a_{i3}\cdot \mathrm{SPREAD}_i(t) + a_{i4}\cdot \mathrm{FICO}_i(t) + \epsilon_{i2}} \tag{12}$$

$$\nu_i^S(t) = b_{i0} \cdot e^{b_{i1}\cdot \mathrm{ALQ}_i(t) + b_{i2}\cdot \mathrm{HOUSE}_i(t) + b_{i3}\cdot \mathrm{SPREAD}_i(t) + b_{i4}\cdot \mathrm{FICO}_i(t) + \epsilon_{i1}} \tag{13}$$

$$\nu_i^I(t) = c_{i0} \cdot e^{c_{i1}\cdot \mathrm{ALQ}_i(t) + c_{i2}\cdot \mathrm{HOUSE}_i(t) + c_{i3}\cdot \mathrm{SPREAD}_i(t) + c_{i4}\cdot \mathrm{FICO}_i(t) + \epsilon_{i3}} \tag{14}$$

$$\alpha_i(t) = d_{i0} \cdot e^{d_{i1}\cdot \mathrm{ALQ}_i(t) + d_{i2}\cdot \mathrm{HOUSE}_i(t) + d_{i3}\cdot \mathrm{SPREAD}_i(t) + d_{i4}\cdot \mathrm{FICO}_i(t) + \epsilon_{i4}} \tag{15}$$

$$\gamma_i(t) = e_{i0} \cdot e^{e_{i1}\cdot \mathrm{ALQ}_i(t) + e_{i2}\cdot \mathrm{HOUSE}_i(t) + e_{i3}\cdot \mathrm{SPREAD}_i(t) + e_{i4}\cdot \mathrm{FICO}_i(t) + \epsilon_{i5}} \tag{16}$$

Since the above approach is modelled as a state space model, the states need to be linked to the observed data by a measurement model. We assume a normal distribution with the observed amount of mortgages in each states as mean values and standard deviations σ_i^S, σ_i^I, σ_i^P and σ_i^D for each state and county respectively.

$ALQ_i(t)$ describes the unemployment rate and $HOUSE_i(t)$ house prices in county i at time t. $SPREAD_i(t)$ is the difference between the weighted average mortgage rate in county i at time t and the national mean rate of 30 year fixed-rated mortgages at time t. $FICO_i(t)$ is the average FICO score of all considered mortgages within county i at time t. I_i^{Data} and dD_i^{Data} are normal distributed random variables with the observed amount of delinquent and newly defaulted mortgages in the previous month $t-1$ as mean values and σ_i^I, σ_i^D as standard deviations.

The error terms $\epsilon_{i1}, ..., \epsilon_{i5}$ are assumed as independent and normal distributed with $\tau_{i1}, ..., \tau_{i5}$ as standard deviations. $\phi_{i,k}(x)$ is considered as a function where x is weighted by the inverse of the distance between county i and county j in miles. $\delta_{i,k}$ captures the dependencies between two counties and is described below.

Equation 7 describes the unobserved system state that tries to capture the systematic distress in a given local mortgage market in county i at time t by accounting for contagion effects through linked counties. With a constant rate of $\mu_{B_i} \cdot B_i(t)$ the bacteria state decreases and is inspired by the cholera epidemic model in Bertuzzo et al. (2010). Infected and previously defaulted mortgages in the neighbourhood contribute to the concentration of state $B_i(t)$ by influencing $\lambda_i(t)$ through equation 12. $\sum_{k=1}^{N} \omega_{i,k} \cdot \delta_{i,k} \cdot \phi_{i,k}(B_k(t))$ connects the bacteria state of county i to bacteria states of linked counties by applying the approach described below.

The dynamics of susceptible mortgages whose payments are on schedule is described by equa-

tion 8. In every period newly originated mortgages that are assumed to be healthy enter the state with rate μ_i. Mortgages that become delinquent with rate $\beta_i \cdot B_i(t) \cdot S_i(t)$ decreases state S in equation 8 and increase state I that is described through equation 9. As it can be seen, the rate at which mortgages become delinquent is highly influenced by the bacteria state B. On the opposite, with rate $\gamma_i(t) \cdot I_i(t)$ delayed mortgage payments are repaid so that the payments are on schedule again. This decrease state I and increase state S. $\gamma_i(t)$ is defined in equation 16.

The amount of mortgages that outstanding debt is fully paid back before termination is estimated by $\nu_i^S(t) \cdot S_i(t)$ which decreases the amount of current mortgages and increases the amount of prepaid mortgages captured by state R in equation 10. $\nu_i^S(t)$ is modelled by a Cox model and can be seen in equation 13.

All mortgages that are infected and their payments are behind schedule are prepaid with rate $\nu_i^I(t) \cdot I_i(t)$ that is captured by equation 14 which decreases state I and increases state R.

Rate $\alpha_i(t) \cdot I_i(t)$ describes the amount of delinquent mortgages that are considered as defaulted. $\alpha_i(t)$ is defined by equation 15 and increases chronological sequence of defaulted mortgages in equation 11.

In figure 1 is a flowchart that illustrates the model.

The question that remains is how to link two counties so that their dependencies can be included in state B or alternatively: How should $\delta_{i,k}$ be estimated ?

Industry sectors aren't evenly distributed throughout the US landscape. They are settled in regions, e.g. automotive sector in Michigan.

Ellison et al. (2007) summarize as a main factor that industry sector are settled in agglomerations through its falling transport costs. Delgado et al. (2016) develop an algorithm that combines all sectors of the 6-digit North American Industry Classification System (NAICS) in 51 traded industries whose occurrence are agglomerated and 16 local industries that are spatial dispersed. Since we are interested in industries that can be mainly found in specific regions, local industries are left out in our study. Their basic idea is to search for co-location patterns, input-output links and similarities in labour occupations to define value chains that are similar in their demands, offers, knowledge and technology. The definition of the value chains and its corresponding 6-digit NAICS code is available upon request.

The question where these agglomerations are found is estimated by the work of Feser et al. (2005) who define a measure to address the geographic occurrence of industrial sectors.
The G-statistic measures the deviation in standard deviations from the mean. For a given value chain u the G-value for county i

$$\mathrm{G}^*_{u,i} = \frac{\sum_j w_{ij}\epsilon_j - W_i\bar{\epsilon}}{s\sqrt{\frac{nS_{1i}-W_i^2}{n-1}}} \tag{17}$$

, given value chain u. w_{ij} is the spatial weight, which defines the neighbouring counties j to i (i.e. either binary adjacency matrix or centroid distance-based approach which decrease as the distance between the centres or two regions increases). Further, let $W_i = \sum_j w_{ij}$ and $\bar{\epsilon} = \sum_j \frac{\epsilon_j}{n-1}$, $S_{1i} = \sum_j w_{ij}^2$ and $s^2 = \left(\frac{\sum_j \epsilon_j^2}{n-1}\right) - (\bar{\epsilon})^2$.
There are multiple possibilities to determine both independent variables of G^*_i and the spatial weight matrix w_{ij}. Feser et al. (2005) propose the use of the residuals after they regressed the value chain employment $\widehat{y}$ on total (export oriented) employment (x) with coefficient (β) derived from national averages (i.e. $\widehat{y} = \beta x + \epsilon$) for each county. Therefore, ϵ_j is the residual of county j for a given value chain and

$$w_{ij} = \begin{cases} \frac{\epsilon_i\epsilon_j}{\sum_j \epsilon_i\epsilon_j} & \text{, county i and county j share a border or i=j} \\ 0 & \text{, else} \end{cases} \tag{18}$$

given the values for the G-statistic are available for all counties N and for all value chains P under considerations. The following methodology may be used to derive the function $\delta_{i,k}$. If the G-value of county n for value chain i is greater than a cutoff value c for any chain i, then the two vertices will be connected. The connection is recorded in a matrix E. For any two counties $k, l \in 1, ..., N$ edges are governed by

$$\delta_{i,k} = \begin{cases} \frac{\sum_u 1_{\mathrm{G}^*_{u,i}>c} \cdot 1_{\mathrm{G}^*_{u,k}>c}}{P} & \text{, if } \sum_u 1_{\mathrm{G}^*_{u,i}>c} \cdot 1_{\mathrm{G}^*_{u,k}>c} \geq 1, i \neq k \\ 0 & \text{, else} \end{cases} \tag{19}$$

The matrix E may be updated on a regular basis in order to consider changes in the underlying industrial structure. The updating frequency is to be determined, as the process is deemed

to be computationally costly. Therefore the G-statistics is computed once for annual means between 2000 and 2014.

The boundary value for the G-statistic was set to $\Phi(1 - \frac{\alpha}{2})$, in which α was set to $0,005$ and Φ denotes the cumulative Gaussian normal distribution. As the G-statistic describes specialized counties, a smaller value α would result in a tighter definition of specialization. Hence, a higher deviation from the national average would be required to denote a county as 'specialized'. The G-statistic for every county and all value chains are computed and the results are available upon request.

As previously noted, Elul et al. (2010) have studied the effect of local unemployment rates on default. The study found that unemployment is a significant, systematic measure for defaults. Therefore, the test for the direct default channel was reduced to test for co-movement of unemployment rates within the counties identified through the value chains.

5 Data

The loan-level data has been obtained from three different sources. All contracts that have its maturity date after January 2000 are included.

First, Bloomberg L.P. (2015) was used to collect a sample from securitised mortgage loans in a series of mortgage backed securities. The data was downloaded in July 2015. A list of the firms and their products is available upon request. The data consists of 2,433,501 mortgage contracts that have originated prior to 2014.

Second, the loan performance data provided by Fannie Mae (2015) and available online has been another source. The data obtained in July 2015 consists of 7,963,189 individual loans. The performance files report the monthly payments of fixed-rated mortgages that are bought by Fannie Mae between January 2000 and March 2012. The duration is between 25 and 35 years and all documents are available at origination date.

Third, loan performance files consisting of 4,205,383 individual loans provided by Freddie Mac (2015) has been sampled online in December 2015. The duration of the fixed-rated mortgages is 30 years and all documents are available at origination date, too.

As previously mentioned, only counties with more than 50 mortgage contracts originated between 2000 and 2014 are considered which covers 1490 counties or over 90 % of the U.S. population according to U.S. Census Bureau (2010). The observation of every county starts, if more than 10 mortgages are originated to avoid data issues. A list of all considered counties is available upon request.

The unemployment rate has been provided by the US. Bureau of Labor Statistics (2015)[1].

Zillow Home Value Index was used to get a proxy of house price data on county level. Unfortunately not all counties considered are covered by the index. Therefore average values of neighbouring counties are used to get monthly house price values in case of missing information.

The mortgage rates of the financial market is provided by Freddie Mac (2016). To get the spread between the national mean and the regional interest rate level the following difference is estimated. In every county, we compute the weighted average coupon rate if more than 100 mortgages are active at the time. If that's not the case we use spatial Kriging to get an

[1] The data has been available online on the website of Federal Reserve Bank of St. Louis.

approximation and to avoid data issues. The national mean of 30-year fixed rated mortgages minus the weighted average regional coupon rate yields the spread mentioned earlier.

The FICO score is a well-known index to capture the creditworthiness of a borrower. We use the average FICO within a county from the loan-level data to obtain a proxy of the ability to pay for the whole county. Like before, to avoid data issues we only consider the mean value if more than 100 mortgages are active during the specific month. Otherwise, we use spatial Kriging to approximate the average FICO score within the county.

In both approximations a Gaussian Semivariogram Model is assumed to get monthly parameter for spatial Kriging. Then spatial mean values are computed for each county to get FICO and spread values at each considered county and date. Information about spatial Kriging can be found for example in Cressie (1992).

In table 1 descriptive statistics of the mortgage data used in our work is presented. As it can be seen most of the defaulted mortgages were originated between 2006 and 2009.

6 Empirical Analysis

First, the iterated filtering algorithm, proposed by Ionides et al. (2006), is performed for each county without contagion effects to or from other counties. Therefore, equation 7 is modified with $\omega_{i,k} = 0, \forall i, k$. The iterated filtering algorithm is performed 60 times with different starting points to estimate all parameters used in the model. While the first observation of every state in a county is used as the state's starting point, parameters are chosen uniformly distributed between $[-1; 1]$. We assume the parameters $a_{i0}, b_{i0}, c_{i0}, d_{i0}, \tau_{i1}, ..., \tau_{i5}, \sigma_i^S, \sigma_i^I, \sigma_i^P, \sigma_i^D$ for all i to be positive which means both they start uniformly distributed between $[0; 1]$ and transition rates are positive, too. Then the sum of all absolute deviations between the mean of simulated data and real observations for all dates and different compartments is estimated. The parameter vector that minimizes the sum of each compartment's absolute values is accepted as optimal. Due to the amount of parameters estimated, only the distribution of each parameter from all local mortgages market is presented in table 2. Furthermore, for every parameter that isn't assumed to be positive distribution tests are performed if the estimation's mean are equal to zero.

Although, parameter estimations are widely distributed including both positive and negative values, several insights can be deduced. Besides of c_3, the parameter distribution for all parameters from equation 13 and 14 that aren't assumed to be positive have got significantly negative means. This is in line with studies from section 3. If county's unemployment rate or spread increase or if house prices or the county's average FICO-score decrease, $\alpha_i(t)$ from equation 15 rise. $\gamma_i(t)$ from equation 16 increases if house prices and the interest rate spread rise, as well as if county's unemployment rates and the average FIC0-score decrease. The bacteria concentration is highly influenced by $\lambda_i(t)$ from equation 12 that is positively affected by county's house prices. It decreases if unemployment rates, spread or the average FICO-score rise.

Although the convergence of iterated filtering algorithms is presented in Ionides et al. (2011), we check how the parameter estimation is influenced by the starting vectors. Therefore, we simulate the time series 100 times based on our parameter vector that we assumed as optimal and estimate parameter vectors with the same settings described earlier based on the simulated time series. Then, a confidence interval at the 95%-level is estimated through the 100

newly estimated parameter vectors. 21 out of the 37 entries are located inside the confidence level.

Since we are interested how mortgage defaults are influenced by other local mortgage markets, we then estimate $\omega_{i,k}, \forall i, k$. Therefore, the optimal parameter vector described above along with $\omega_{i,k} = 0, \forall i, k$ is used as a starting point to estimate dependencies between different regions. The Iterated Filtering algorithm is performed 100 times. Since there are 104550 parameter values of $\omega_{i,k}$ that are estimated, only some descriptive statistics are presented here. $\omega_{i,k}$ is distributed between $-3,2 \cdot 10{-}4$ and $3,2 \cdot 10^{-4}$ with a mean value of $1,43 \cdot 10^{-5}$, a standard deviation of $4,59 \cdot 10^{-3}$ and a median of $-9,14 \cdot 10^{-8}$. A complete list of all values is available upon request.

After the performance of the iterated filtering algorithm, the complete model is simulated 500 times with the optimal parameter vector and mean values per compartment, county and date are computed to get 4 time series per county; one for each compartment.

For every county and compartment R^2-values between simulated values and real data are computed to show the share of variability of each compartment explained by the proposed model. Figure 2 shows the histograms of all R^2-values for each compartment and each county that consists of more than 50 mortgages due to our data. Since we are interested in modelling the whole mortgage market and the influence both between the defined compartments within a county and the dependencies from other regions, we show the explained variability not only of defaulted mortgages. Fluctuations of mortgages that are paid on schedule and explained by the model can be seen in figure 2a with a median value of 47%. Our model explains more than 59% of the variability of the share of delayed mortgages for 50% of all observed counties as it is shown in figure 2b. The share of prepaid and defaulted mortgages are explained through R^2-values with a median of 85% and 96%, respectively. This can be seen in figure 2c and 2d.

We used four different approaches from previous studies that capture mortgage defaults to compare to the performance of the presented epidemiological model. For this purpose we estimate all parameters for our model by use of the data till December 2012. Then we predict the behaviour of each county's four compartments and compare them to the real dynamics via R^2 and absolute residuals. The four concepts that we want to compare to our approach are:

1. Linear regression that is used for example in Agarwal et al. (2009); Beem Jr. (2014): We describe each of the four compartments through an equation with all variables defined in section 4. Parameters are estimated due to the data until December 2012 and the compartments are predicted till December 2014 for each county individually.

2. Linear regression of log-Odds that is used in Coleman et al. (2005); Misina et al. (2006) among others: We describe the log-Odds for the default and prepaid compartment through an equation with all variables defined in section 4. Parameters are estimated due to the data until December 2012 and the log-Odds are predicted till December 2014 for each county.

3. Multinomial logit model that is used for example in the study of Elul et al. (2010); Floros and White (2016): We describe the probability that a mortgage is prepaid, defaulted or right censored through a multinomial logit model with both all variables defined in section 4 and individual loan characteristics like the coupon rate, individual FICO-score, loan-to-value ratio and the unpaid balance at origination. Parameters are estimated due to all loans that are originated until December 2012 and a forecast is made for all loans originated after January 2013.

4. Multinomial probit model that is used in e.g. Rebelo and Caldas (2010); Rajan et al. (2015): We describe the probability that a mortgage is prepaid, defaulted or right censored through a multinomial probit model with both all variables defined in section 4 and individual loan characteristics like the coupon rate, individual FICO-score, loan-to-value ratio and the unpaid balance at origination. Parameters are estimated due to all loans that are originated until December 2012 and a forecast is made for all loans originated after January 2013.

We define two prediction periods. The first period is 2013 and the second is both 2013 and 2014 to consider short-term and long-term forecast performance. We estimate the difference between R^2-values and absolute residuals of each compartment and each county of the four well-known models and our approach. Then we test if the mean of the distribution is equal to zero.

First, we compare our model to a the linear regression and show the results in table 3. When predicting the four compartments both one year and two years ahead the state of current and delayed mortgages are determined more precisely on average through a linear regression

than the epidemiological approach. The linear regression shows averagely higher R^2-values and lower absolute residuals. Both the compartments of prepaid and defaulted mortgages and the forecast of all four compartments combined are predicted in greater detail through the epidemiological approach. The R^2-values are significantly higher and the differences between the absolute residuals are significantly lower compared to the linear regression model. These effects can be seen on short and long-term forecasts.

Second, we compare our model to the linear regression of log Odds-ratio. Again, we predict the odds ratio one year and two years ahead and show the results in table 4. Here, we don't compare the four compartments, but the amount of defaulted or prepaid mortgages in a specific period compared to the amount of active mortgages, called default-rate or prepaid-rate. As it can be seen, the epidemiological approach predict the default-rate and the prepaid-rate more precisely than the linear regression of log odds both at the one year forecast and the two years forecast. The absolute residuals from our approach are on average highly significant lower in both periods. If we compare the differences between the R^2-values of both models we can just show significantly higher values of the default-rates at the one year forecast.

The multinomial logit and probit model both estimate the probability that a mortgage is prepaid, defaulted or is paid on schedule over the whole duration. To compare our approach to both models we predict the probability of each individual mortgage in 2013 and 2014 and compare the average probability in each county with the quotient between the average amount of defaulted or prepaid mortgages and the amount of active mortgages in the specific county. The results of the prediction accuracy between the logit, probit and the epidemiological approach is presented in table 5. Since the logit and probit model don't do any statement when a mortgage defaults or is prepaid, we only compare our model to them over the long-term period of two years. The results show that both the default-rate and the prepaid-rate is described on average more precisely through our approach. R^2-values over all counties for default rates are $11,88\%$ estimated through the logit model and $21,92\%$ through the probit approach. Our model shows a R^2 of $35,97\%$. Prepaid-rates show R^2-values of $0,13\%$ and $0,18\%$ for logit and probit models, compared to $0,05\%$ through our approach.

In Figure 3 and figure 4 the evolution of predicted and real values of delayed and defaulted mortgages are visualized. The two sample counties Boulder County in Colorado and San Diego County in California are picked because both they predict defaulted mortgages above

average if compared to other counties and sufficient mortgages were taken up within those counties according to our data. The real data from December 2012 is used as a starting point for the long-term prediction of two years.

In summary it can be said, therefore, that our model explains an above average proportion of the variance of mortgage defaults than previous studies.

7 Conclusion

With our approach, we can, for the first time, model the spatial contagion effect of mortgage defaults using an epidemiological approach. We introduced the connection between compartments of disease models and the corresponding states in the mortgage market. With the concept of G-statistic as a measure of geographic occurrence of industrial sectors we introduce an approach to capture the dependencies of other local mortgage markets that are far away but economically similar.

After origination a mortgage contract is assumed to be paid on schedule and is therefore seen as healthy or susceptible. During its lifetime the payments can get delinquent due to macroeconomic conditions which is comparable to an infected person and, at worst, defaults (this means the infected person dies). Another opportunity for borrower is to prepay the mortgage either in a healthy or infected state. These loans leave the model like recovered people in an epidemic approach.

Furthermore, we estimate the parameter vector with a big dataset of more that 14 million loans originated between 2000 and 2014 using the iterated filtering algorithm and describe the performance of the simulated compartments. We showed that our approach predicts future default and prepayment rates more precisely than previous concepts like linear regression, logit and probit models or linear regression of log odds.

In summary, our new approach to estimate the dynamics of mortgage default rates and its spatial contagion effect even between local markets that are far apart is a new strategy that shows an above-average performance and support the idea of using more flexible concepts in the mortgage market.

References

Agarwal, S., G. Amromin, I. Ben-david, S. Chomsisengphet, and D. D. Evanoff (2009): "Do Financial Counseling Mandates Improve Mortgage Choice and Performance? Evidence from a Legislative Experiment," *Federal Reserve Bank of Chicago, Working Paper*, [online] http://www.chicagofed.org/digital_assets/publicati...s/2009/wp2009_07.pdf [10.09.2016].

Bajari, P., C. Chu, D. Nekipelov, and M. Park (2013): "A DYNAMIC MODEL OF SUBPRIME MORTGAGE DEFAULT: ESTIMATION AND POLICY IMPLICATIONS," *Working Paper. National Bureau of Economic Research.*

Beem Jr., R. H. (2014): "Residential Mortgage Delinquency Rates: The Determinants of Default," *Issues in Political Economy*, 23, 59–75.

Bertuzzo, E., R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo (2010): "On spatially explicit models of cholera epidemics." *Journal of the Royal Society, Interface / the Royal Society*, 7, 321–33.

Bloomberg L.P. (2015): "Mortgage loan level data," *Bloomberg database. Karlsruher Institut für Technologie (KIT) - Lehrstuhl für Financial Engineering und Derivate.*

Brauer, F., P. Van den Driessche, and J. Wu (2008): *Mathematical Epidemiology*, vol. 1945 of *Lecture Notes in Mathematics*, Berlin, Heidelberg: Springer Berlin Heidelberg.

Britton, T. (2010): "Stochastic epidemic models: a survey." *Mathematical biosciences*, 225, 24–35.

Burnside, C., M. Eichenbaum, and S. Rebelo (2016): "Understanding Booms and Busts in Housing Markets," *Journal of Political Economy*, 124.

Chan, S., M. Gedal, V. Been, and A. Haughwout (2013): "The role of neighborhood characteristics in mortgage default risk: Evidence from New York City," *Journal of Housing Economics*, 22, 100–118.

Coleman, A., N. Esho, I. Sellathurai, and I. Thavabalan (2005): "Stress Testing Housing Loan Portfolios: A Regulatory Case Study," *Australian Prudential Regulation Authority Working Paper*.

Cowan, A. M. and C. D. Cowan (2004): "Default correlation: An empirical investigation of a subprime lender," *Journal of Banking and Finance*, 28, 753–771.

Cox, D. (1972): "Regression Models and Life-Tables," *Journal of the Royal Statistical Society*, Series B 3, 187–220.

Cressie, N. (1992): "STATISTICS FOR SPATIAL DATA," *Terra Nova*, 4, 613–617.

Danis, M. a. and A. Pennington-Cross (2008): "The delinquency of subprime mortgages," *Journal of Economics and Business*, 60, 67–90.

Delgado, M., M. E. Porter, and S. Stern (2016): "Defining clusters of related industries," *Journal of Economic Geography*, 16, 1–38.

Deng, Y., J. M. Quigley, R. Van Order, and F. Mac (1996): "Mortgage default and low downpayment loans: The costs of public subsidy," *Regional Science and Urban Economics*, 26, 263–285.

Divino, J. A. and L. C. S. Rocha (2013): "Probability of default in collateralized credit operations," *The North American Journal of Economics and Finance*, 25, 276–292.

Duffie, D., A. Eckner, G. Horel, and L. Saita (2009): "Frailty correlated default," *Journal of Finance*, 64, 2089–2123.

Elandt-Johnson, R. C. and N. L. Johnson (1999): *Survival Models and Data Analysis*, John Wiley & Sons.

Ellison, G., E. L. Glaeser, and W. Kerr (2007): "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," *Working Paper. National Bureau of Economic Research*.

Elul, R., N. Souleles, and S. Chomsisengphet (2010): "What Triggers Mortgage Default?" *American Economic Review*, 100.

Fannie Mae (2015): "Loan Performance Data," [online] https://loanperformancedata.fanniemae.com/lppub/index.html [24.07.2015].

Feser, E., S. Sweeney, and H. Renski (2005): "A Descriptive Analysis of Discrete U.S. Industrial Complexes*," *Journal of Regional Science*, 45, 395–419.

Floros, I. and J. T. White (2016): "Qualified residential mortgages and default risk," *Journal of Banking & Finance*, 70, 86–104.

Freddie Mac (2015): "Loan-Level Dataset," [online] https://freddiemac.embs.com/FLoan/Data/download.php [08.12.2015].

——— (2016): "Mortgage Rates Survey Archive," [online] http://www.freddiemac.com/pmms/pmms_archives.html [29.06.2016].

Goodstein, R., P. E. Hanouna, C. D. Ramirez, and C. W. Stahel (2011): "Are Foreclosures Contagious?" *SSRN Electronic Journal*.

Grubesic, T. H. (2008): "Zip codes and spatial analysis: Problems and prospects," *Socio-Economic Planning Sciences*, 42, 129–149.

Hillebrand, E., A. N. Sengupta, and J. Xu (2012): "Temporal correlation of defaults in subprime securitization," *Communications on Stochastic Analysis*, 6, 487–511.

Ionides, E., C. Bretó, and A. King (2006): "Inference for nonlinear dynamical systems," *Proceedings of The National Academy of Sciences of the USA*, 103, 18438–18443.

Ionides, E. L., A. Bhadra, Y. Atchadé, and A. King (2011): "Iterated filtering," *Annals of Statistics*, 39, 1776–1802.

Kalbfleisch, J. D. and R. L. Prentice (2011): *The Statistical Analysis of Failure Time Data*, John Wiley & Sons.

Keeling, M. J. and K. T. D. Eames (2005): "Networks and epidemic models." *Journal of the Royal Society, Interface / the Royal Society*, 2, 295–307.

Kermack, W. O. and A. G. McKendrick (1927): "A Contribution to the Mathematical Theory of Epidemics," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 115, 700–721.

Lawless, J. (2011): *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd ed.

Matthews, L. and M. Woolhouse (2005): "New approaches to quantifying the spread of infection." *Nature reviews. Microbiology*, 3, 529–36.

Merriam-Webster (2016): "County," [online] http://www.merriam-webster.com/dictionary/counties [31.03.2016].

Misina, M., D. Tessier, and S. Dey (2006): "Stress Testing the Corporate Loans Portfolio of the Canadian Banking Sector," *Bank of Canada Working Paper*, 2006-47.

Piazzesi, M. and M. Schneider (2009): "Momentum traders in the housing market: survey evidence and a search model," *American Economic Review P&P*, 99, 406–411.

Quercia, R. G. and M. a. Stegman (1992): "Residential Mortgage Default : A Review of the Literature *," *Journal of Housing Research*, 3, 341–380.

Rajan, U., A. Seru, and V. Vig (2015): "The failure of models that predict failure: Distance, incentives, and defaults," *Journal of Financial Economics*, 115, 237–260.

Rebelo, J. and J. V. Caldas (2010): "DEFAULT MORTGAGE PROFILE : A MICRO ANALYSIS OF THE PORTUGUESE CASE," *Portugese Journal of Management Studies*, XV, 109–126.

Robert H. Shumway and D. S. Stoffer (2011): *Time Series Analysis and Its Applications With R Examples*, Springer-Verlag New York, 3rd ed.

US. Bureau of Labor Statistics (2015): "Unemployment in States and Local Areas (all other areas)," [online] https://fred.stlouisfed.org/tags/series?ob=pv&od=desc&t=county%3Bmonthly%3Bunemployment [05.10.2015].

U.S. Census Bureau (2010): "County Adjacency File," [online] https://www.census.gov/geo/reference/county-adjacency.html [25.11.2014].

Weiler, S. (2001): "Unemployment in regional labor markets: Using structural theories to understand local jobless rates in West Virginia," *Industrial and Labor Relations Review*, 54, 573–592.

8 Tables

Table 1: Descriptive statistics of the used data: the amount of observed mortgages, the share of defaulted and prepaid mortgages per origination year till the end of 2014.

Origination year	Amount	Share of prepaid mortgages in %	Share of defaulted mortgages in %
1995	2	0	100
1996	8	0	75
1997	40	0	82,5
1998	432	0	79,4
1999	198576	95,19	2,05
2000	531500	96,11	2,32
2001	1161116	95,55	1,85
2002	1249121	92,67	2,01
2003	895079	86,28	2,51
2004	705582	79,51	4,25
2005	794234	72,97	7,31
2006	1241833	61,35	16,29
2007	987064	55,35	21,06
2008	1195790	60,85	16,72
2009	1712289	54,22	10,73
2010	711539	55,6	0,84
2011	622896	45,16	0,5
2012	972210	18,15	0,2
2013	999383	7,13	0,17
2014	623379	3,38	0,03
Total	14602073	58,9	6,68

Table 2: Descriptive statistics of parameter estimation
Descriptive statistics of the parameter estimation without any contagion to or from other counties. The distribution of absolute residuals and log-likelihoods for each county are presented, too. The distribution of starting values for the Iterated Filtering algorithm are given, as well as the minimum, maximum, mean, standard deviation, median and the lower and upper quartile of the estimated parameters. $\mathcal{U}(0;1)$ describes a uniform distribution between 0 and 1. All parameters are described between equation 7 and equation 16. $^{*}p < 0,05$, $^{**}p < 0,01$ and $^{***}p < 0,001$ shows the significance level if the hypothesis that the mean of the parameter is equal to zero can be rejected.

parameter	starting values	min	max	mean	std.	median	25%-quantile	75%-quantile
μ_B	$\in \mathcal{U}(0;0,01)$	$-0,048$	$0,4845$	$0,1436^{***}$	$0,1073$	$0,138$	$0,0467$	$0,2175$
a_0	$\in \mathcal{U}(0;1)$	0	$5701,1179$	$11,9757$	$162,2451$	$0,0656$	$0,0035$	$0,7225$
a_1	$\in \mathcal{U}(-1;1)$	$-21,8823$	$19,0543$	$-0,4804^{*}$	$4,9222$	$-0,3089$	$-3,5627$	$2,6522$
a_2	$\in \mathcal{U}(-1;1)$	$-17,6773$	$16,4966$	$0,0688$	$5,2092$	$0,0786$	$-3,2539$	$3,5217$
a_3	$\in \mathcal{U}(-1;1)$	$-16,7574$	$16,2802$	$-1,3328^{***}$	$5,1004$	$-1,407$	$-4,8581$	$1,9386$
a_4	$\in \mathcal{U}(-1;1)$	$-18,4288$	$14,2373$	$-1,6141^{***}$	$4,832$	$-1,7233$	$-4,7448$	$1,6113$
b_0	$\in \mathcal{U}(0;1)$	0	$1244,0863$	$3,6140$	$45,538$	$0,0028$	$0,0001$	$0,0544$
b_1	$\in \mathcal{U}(-1;1)$	$-19,8978$	$17,9656$	$-1,6851^{***}$	$5,3843$	$-1,4422$	$-5,1567$	$1,8142$
b_2	$\in \mathcal{U}(-1;1)$	$-18,9368$	$16,3412$	$-1,224^{***}$	$5,3914$	$-1,2228$	$-4,6627$	$2,4684$
b_3	$\in \mathcal{U}(-1;1)$	$-15,7788$	$14,0072$	$-0,4213^{*}$	$4,8996$	$-0,3786$	$-3,7302$	$2,9093$
b_4	$\in \mathcal{U}(-1;1)$	$-20,4202$	$14,2985$	$-2,329^{***}$	$4,994$	$-2,4471$	$-5,6623$	$0,9278$
c_0	$\in \mathcal{U}(0;1)$	0	$256,604$	$0,8503$	$9,3001$	$0,056$	$0,0011$	$0,4933$
c_1	$\in \mathcal{U}(-1;1)$	$-20,613$	$15,4185$	$-0,9446^{***}$	$5,209$	$-0,9756$	$-4,2658$	$2,5447$
c_2	$\in \mathcal{U}(-1;1)$	$-21,0172$	$17,549$	$-1,3304^{***}$	$5,155$	$-1,2898$	$-4,5564$	$2,1796$
c_3	$\in \mathcal{U}(-1;1)$	$-19,4739$	$17,5729$	$-0,034$	$4,9214$	$-0,1081$	$-3,2959$	$3,1625$
c_4	$\in \mathcal{U}(-1;1)$	$-21,9419$	$10,1632$	$-4,9443^{***}$	$4,753$	$-4,9097$	$-7,8229$	$-2,0036$
d_0	$\in \mathcal{U}(0;1)$	0	$1229,8495$	$4,62$	$56,1177$	$0,1765$	$0,0181$	$0,6369$
d_1	$\in \mathcal{U}(-1;1)$	$-18,6083$	$26,2288$	$1,1069^{***}$	$5,8314$	$0,9133$	$-2,8203$	$4,7567$
d_2	$\in \mathcal{U}(-1;1)$	$-18,3982$	$18,305$	$-0,652^{***}$	$5,2332$	$-0,5952$	$-4,1475$	$2,7115$
d_3	$\in \mathcal{U}(-1;1)$	$-19,2549$	$22,7965$	$0,4222^{*}$	$5,1507$	$0,3634$	$-2,8417$	$3,6781$
d_4	$\in \mathcal{U}(-1;1)$	$-18,8977$	$11,6279$	$-4,0278^{***}$	$4,2121$	$-4,3557$	$-6,7781$	$-1,4783$
e_0	$\in \mathcal{U}(0;1)$	0	$3673,3285$	$9,8503$	$113,7003$	$0,3061$	$0,0312$	$0,7815$
e_1	$\in \mathcal{U}(-1;1)$	$-18,0942$	$16,0386$	$-1,6741^{***}$	$4,8261$	$-1,8034$	$-4,8131$	$1,5316$
e_2	$\in \mathcal{U}(-1;1)$	$-18,0664$	$19,5672$	$0,2969$	$5,1675$	$0,362$	$-3,1509$	$3,6631$
e_3	$\in \mathcal{U}(-1;1)$	$-15,2089$	$17,7133$	$1,092^{***}$	$4,8661$	$1,1946$	$-2,0059$	$4,2625$
e_4	$\in \mathcal{U}(-1;1)$	$-18,1969$	$14,7393$	$-1,309^{***}$	$4,3014$	$-1,0416$	$-3,795$	$1,1859$
τ_1	$\in \mathcal{U}(0;0,1)$	0	$0,1269$	$0,0488$	$0,0292$	$0,0476$	$0,0238$	$0,0731$
τ_2	$\in \mathcal{U}(0;0,1)$	0	$0,1170$	$0,0488$	$0,0288$	$0,0476$	$0,0236$	$0,0722$
τ_3	$\in \mathcal{U}(0;0,1)$	$0,0001$	$0,1239$	$0,0494$	$0,03$	$0,0475$	$0,0237$	$0,0744$
τ_4	$\in \mathcal{U}(0;0,1)$	$0,0001$	$0,1182$	$0,0506$	$0,0296$	$0,0502$	$0,0248$	$0,0755$
τ_5	$\in \mathcal{U}(0;0,1)$	0	$0,1206$	$0,0503$	$0,0295$	$0,0491$	$0,0255$	$0,0749$
μ	$\in \mathcal{U}(0;1)$	0	$0,0109$	$0,0043$	$0,0016$	$0,004$	$0,0033$	$0,0054$
β	$\in \mathcal{U}(0;1)$	$-0,2252$	$1,2045$	$0,5523^{***}$	$0,3039$	$0,5706$	$0,2906$	$0,8119$
ρ^S	$\in \mathcal{U}(0;1)$	$0,0016$	$0,8521$	$0,19$	$0,1093$	$0,1829$	$0,1047$	$0,2601$
ρ^I	$\in \mathcal{U}(0;1)$	$0,0006$	$0,6281$	$0,1146$	$0,0805$	$0,096$	$0,0517$	$0,1637$
ρ^P	$\in \mathcal{U}(0;1)$	$0,0157$	$0,7959$	$0,2055$	$0,097$	$0,2004$	$0,1399$	$0,2612$
ρ^D	$\in \mathcal{U}(0;1)$	$0,0003$	$0,6017$	$0,1025$	$0,0742$	$0,0876$	$0,0446$	$0,1455$
likelihood		$-1538,2513$	$1769,6048$	$611,9083$	$321,3004$	$516,8854$	$387,4436$	$773,8229$
residuals		$6,135$	$63,0931$	$21,5549$	$8,7291$	$19,6861$	$14,377$	$27,7738$

Table 3: Performance of epidemiological approach compared to linear regression
The distribution of the difference between R^2 of predicted compartments between the epidemiological approach and the linear regression is estimated. Furthermore the absolute residuals of the predicted compartments between the epidemiological approach and the linear regression is estimated, too. $^{*}p < 0,05$, $^{**}p < 0,01$ and $^{***}p < 0,001$ shows the significance level if the hypothesis that the mean of the difference is equal to zero can be rejected.

compartment	period	difference	mean	median
Current	2013	R^2	$-0,0136$	$0,0044$
Default	2013	R^2	$0,3784^{***}$	$0,3539$
Delayed	2013	R^2	$-0,0156$	$0,0032$
Prepaid	2013	R^2	$0,3548^{***}$	$0,3079$
Current	2013 & 2014	R^2	$-0,1504^{***}$	$-0,0898$
Default	2013 & 2014	R^2	$0,4111^{***}$	$0,3842$
Delayed	2013 & 2014	R^2	$-0,1921^{***}$	$-0,1217$
Prepaid	2013 & 2014	R^2	$0,3772^{***}$	$0,3327$
Default	2013	Residuals	$-0,5994^{***}$	$-0,5096$
Delayed	2013	Residuals	$0,2067^{***}$	$0,1464$
Prepaid	2013	Residuals	$-0,9196^{***}$	$-0,8099$
all	2013	Residuals	$-1,0871^{***}$	$-0,9638$
Default	2013 & 2014	Residuals	$-1,5703^{***}$	$-1,3677$
Delayed	2013 & 2014	Residuals	$0,3983^{***}$	$0,3733$
Prepaid	2013 & 2014	Residuals	$-2,2909^{***}$	$-2,0625$
all	2013 & 2014	Residuals	$-3,0133^{***}$	$-2,6652$

Table 4: Performance of epidemiological approach compared to linear regression of log odds
The distribution of the difference between R^2 of predicted default-rates and prepaid-rates between the epidemiological approach and the linear regression of log odds is estimated. Furthermore the absolute residuals of the predicted default-rates and prepaid-rates between the epidemiological approach and the linear regression of log odds is estimated, too. $^{*}p < 0,05$, $^{**}p < 0,01$ and $^{***}p < 0,001$ shows the significance level if the hypothesis that the mean of the difference is equal to zero can be rejected.

compartment	period	difference	mean	median
default-rate	2013	R^2	$0,0166$	$0,0011$
prepaid-rate	2013	R^2	$-0,0022$	0
default-rate	2013	Residuals	$-0,2725^{***}$	$-0,0039$
prepaid-rate	2013	Residuals	$-0,3873^{***}$	$-0,0174$
default-rate	2013 & 2014	R^2	$-0,0021$	0
prepaid-rate	2013 & 2014	R^2	$-0,0121$	$-0,0009$
default-rate	2013 & 2014	Residuals	$-0,6637^{***}$	$-0,0126$
prepaid-rate	2013 & 2014	Residuals	$-0,929^{***}$	$-0,0504$

Table 5: Performance of epidemiological approach compared to multinomial logit and probit model
The distribution of the difference between the absolute residuals of the predicted default-rates and prepaid-rates between the epidemiological approach and the logit and probit model is estimated. $^{*}p < 0,05$, $^{**}p < 0,01$ and $^{***}p < 0,001$ shows the significance level if the hypothesis that the mean of the difference is equal to zero can be rejected.

model	compartment	period	difference	mean	median
logit	default-Rate	2013 & 2014	Residuals	$-0,1460^{***}$	$-0,0121$
probit	default-Rate	2013 & 2014	Residuals	$-0,0887^{***}$	$-0,0551$
logit	prepaid-Rate	2013 & 2014	Residuals	$-0,1734^{***}$	$-0,0904$
probit	prepaid-Rate	2013 & 2014	Residuals	$-0,2033^{***}$	$-0,1131$

9 Figures

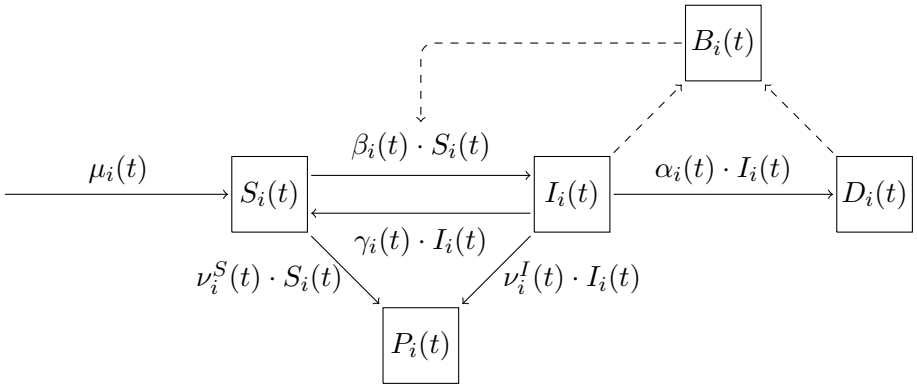
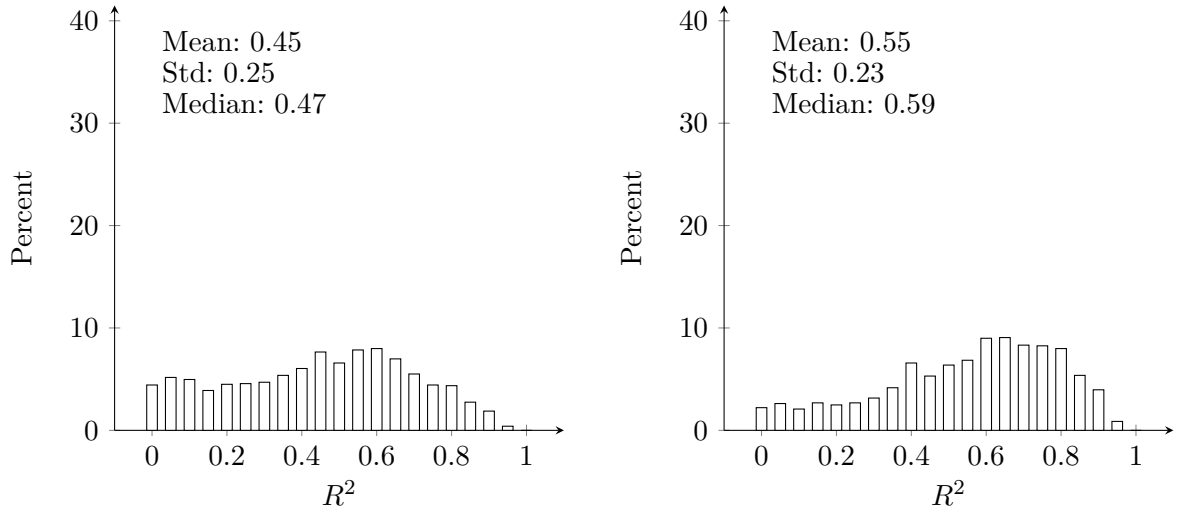
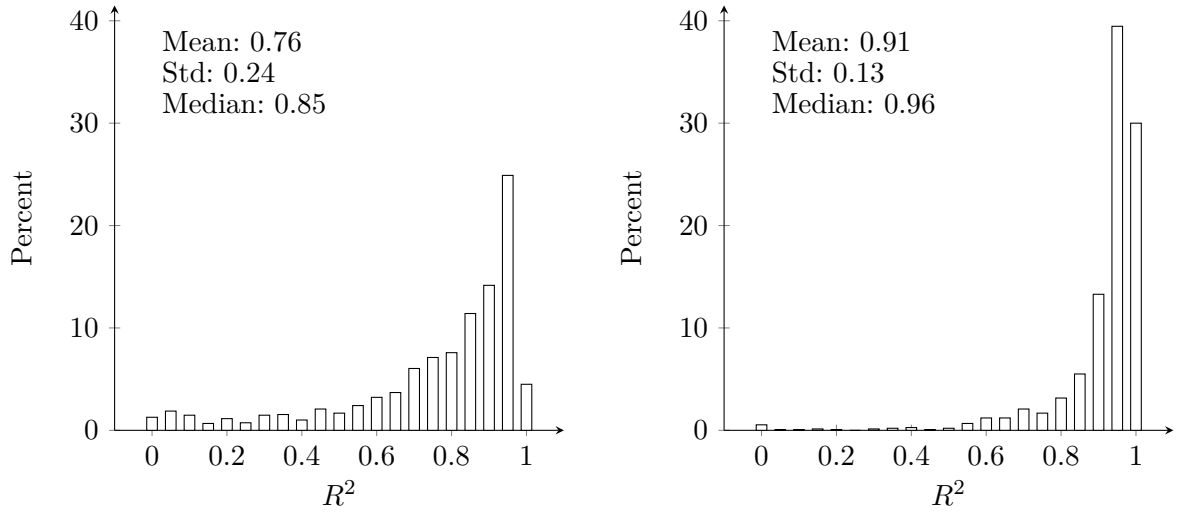


Figure 1: Flowchart of local mortgage market
Description of the local mortgage market through a flowchart. $\mu_i(t)$ is the rate of newly originated mortgages that increase the amount of susceptible loans $S_i(t)$ between $[t-1, t)$. $\beta_i(t)$ shows the amount of newly delinquent loans and $\alpha_i(t)$ describes the rate of delinquent loans $I_i(t)$ that default and therefore stay in state $D_i(t)$ for the rest of the observation period. $\gamma_i(t)$ shows the extent to which loans get back to their scheduled payments and $\nu_i^s(t)$, $\nu_i^I(t)$ are rates at which mortgages are prepaid and increase $P_i(t)$ depending from which state ($S_i(t)$ or $I_i(t)$) they come from. $B_i(t)$ is the unobserved system state that represents systematic distress.



(a) Current mortgages

(b) Delayed mortgages



(c) Prepaid mortgages

(d) Defaulted mortgages

Figure 2: Histogram of R^2-values
The figure shows the R^2 distribution of each county in the specific compartment. Subfigure 2a shows the variability of all mortgages that are paid on schedule explained by the proposed model. Subfigure 2b shows the R^2 distribution of all delayed mortgages. Subfigure 2c and 2d shows the explained variability of prepaid and defaulted mortgages, respectively.

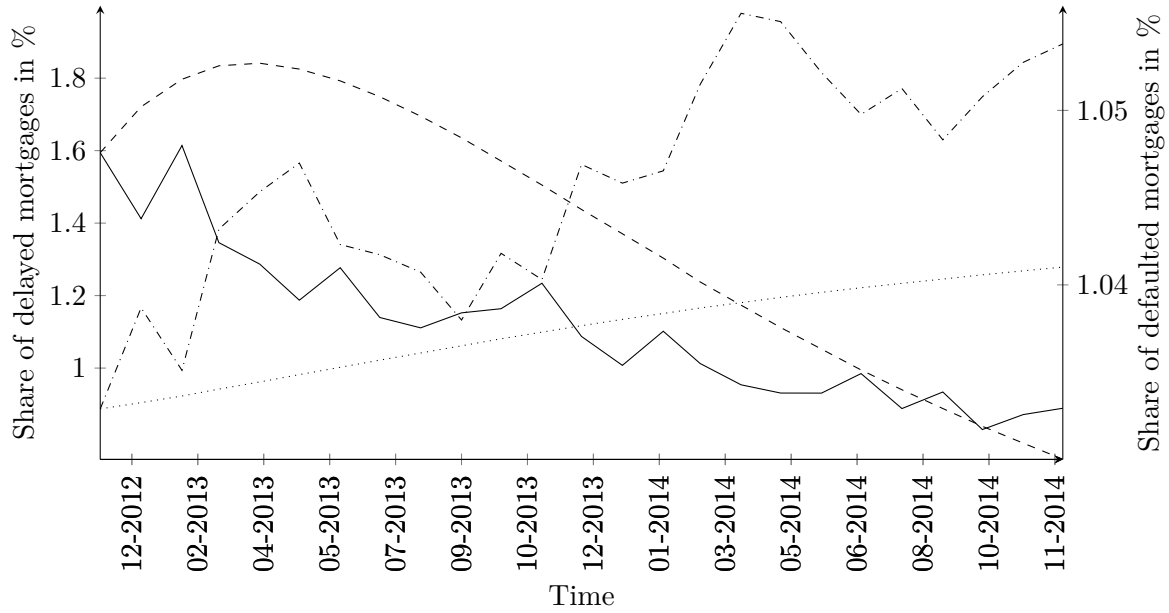


Figure 3: Performance of predicted evolution in Boulder County (Colorado)
The figure shows the predicted and real values of both delayed and defaulted share of mortgages in Boulder County (Colorado) between January 2013 and December 2014. Delayed mortgages are shown on the left y-axis while defaulted mortgages are shown on the right y-axis. The solid line describes the evolution of the real share of delayed mortgages and the dashed line shows the predicted values. The dash-dotted line displays the share of defaulted mortgages according to the data and the dotted line describes the predicted values.

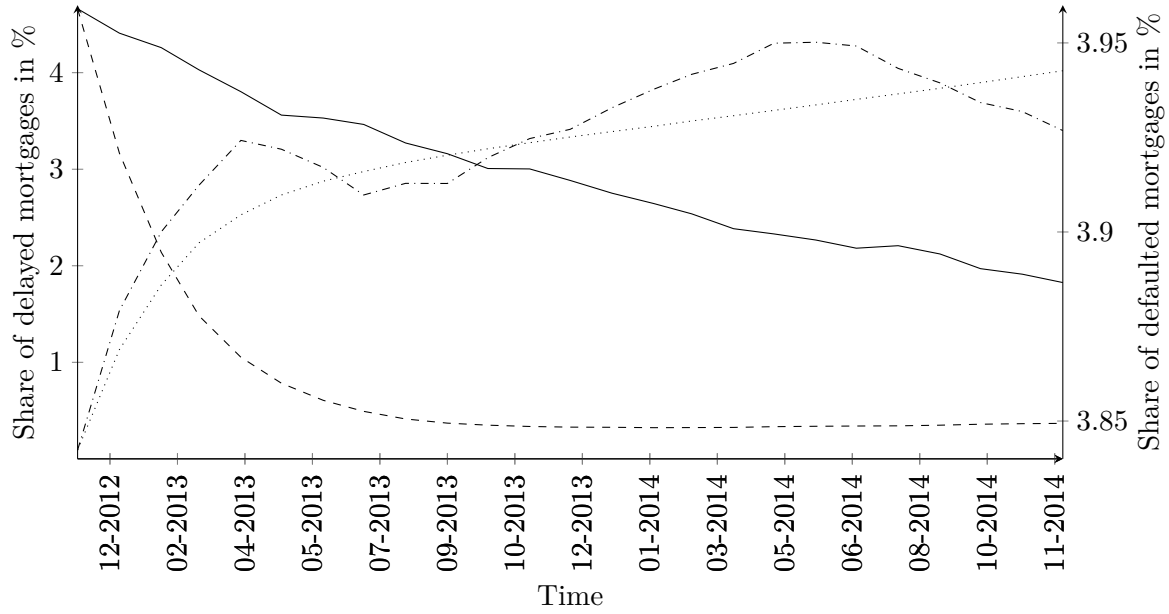


Figure 4: Performance of predicted evolution in San Diego County (California)
The figure shows the predicted and real values of both delayed and defaulted share of mortgages in San Diego County (California) between January 2013 and December 2014. Delayed mortgages are shown on the left y-axis while defaulted mortgages are shown on the right y-axis. The solid line describes the evolution of the real share of delayed mortgages and the dashed line shows the predicted values. The dash-dotted line displays the share of defaulted mortgages according to the data and the dotted line describes the predicted values.

Working Paper Series in Economics

recent issues

No. 112 *Jochen Schweikert and Markus Höchstötter:* Epidemiological spreading of mortgage default, January 2018

No. 111 *Armin Falk and Nora Szech:* Diffusion of being pivotal and immoral outcomes, December 2017

No. 110 *Leonie Kühl and Nora Szech:* Physical distance and cooperativeness towards strangers, November 2017

No. 109 *Deniz Dizdar, Benny Moldovanu and Nora Szech:* The multiplier effect in two-sided markets with bilateral investments, November 2017

No. 108 *Andranik S. Tangian:* Policy representation by the 2017 Bundestag, September 2017

No. 107 *Andranik S. Tangian:* Policy representation by German parties at the 2017 federal election, September 2017

No. 106 *Andranik S. Tangian:* Design and results of the third vote experiment during the 2017 election of the Karlsruhe Institute of Technology student parliament , September 2017

No. 105 *Markus Fels:* Incentivizing efficient utilization without reducing access: The case against cost-sharing in insurance, July 2017

No. 104 *Andranik S. Tangian:* Declining labor–labor exchange rates as a cause of inequality growth, July 2017

No. 103 *Konstanze Albrecht, Florentin Krämer and Nora Szech:* Animal welfare and human ethics: A personality study, June 2017

No. 102 *Jannis Engel and Nora Szech:* A little good is good enough: Ethical consumption, cheap excuses, and moral self-licensing, March 2017