TECHNICAL UNIVERSITY OF MUNICH

TUM Data Innovation Lab

Data Driven Risk-Return Computation for Real Estate

Authors	B.Sc. Marina Lex, B.Sc. Alessandro Pesce, B.Sc. Maurice Schweitzer, M.Sc. M.Sc. Florian Wachter, B.Sc. Jan Watter
Mentor	M.Sc. Oliver Bachmann
Co-Mentor	Dr. Maximilian Engel
Project Lead	Dr. Ricardo Acevedo Cabra (Department of Mathematics)
Supervisor	Prof. Dr. Massimo Fornasier (Department of Mathematics)

Jul 2020

Contents

Table of Acronyms

Acronym	Description
CB	CapitalBay
DCF	Discounted Cash Flow
NPV	Net Present Value
IRR	Internal Rate of Return
KPI	Key Performance Indicator
ROI	Return of Investment
SDE	Stochastic Differential Equation
HPI	Housing Price Index
ETL	Extract Transform Load

Introduction

As current interest rates are diminishing and stock market investments are subject to severe volatility, the real estate investment market at least in Germany represents a profitable investment alternative. CapitalBay (CB), a real estate management and investment services company with roughly €4.5 billion assets under management, focuses on digital end-to-end solutions. One piece of CBs business model puzzle is the prediction of cash flows and risk figures [Capital Bay, 2020].

As with all investment alternatives, the return as well as risk are the two main key performance indicators (KPIs) for real estate investment as well. For single real estate objects (e.g. one single apartment, a single house) basic discounted cash flow (DCF), net present value (NPV) or internal rate of return (IRR) calculations might be sufficient. Those return models have the real estate purchase price and rent income as inputs in common. For investor, these KPIs should be accompanied by a certain risk measure. As neither vanilla NPV, IRR nor DCF are accounting for the risk of the investment, augmentations and techniques as Monte Carlo simulations or scenario analyses expand return models such that their output is not only one single return KPI but a distribution. With that, one widely used definition of risk is the standard deviation of e.g. IRR. The larger this measure is, the more riskier the investment is considered. As with return KPIs, other risk measures exist as well. In general, risk is the possibility of an undesired event multiplied the costs that occur when it materializes [Kaplan and Garrick, 1981], see Equation 1.

$$Risk = Possibility \times Costs \tag{1}$$

Next, KPIs as the Sharpe Ratio intertwine risk and return to make investment alternatives more comparable, Equation 2. Here $\bar{d}$ is the expected value of the return and σ_d denotes the standard variation of the return[Sharpe, 1994].

$$S = \bar{d}/\sigma_d \tag{2}$$

For portfolios of investment objects one can compute the investment allocation for all entities that maximizes return and minimizes risk.

In this project, our goal was to predict the house and rent prices for the biggest 149 cities in Germany for the next 10 years. Mainly, we tackled this with stochastic differential equations (SDE), see SDE Model for reference. At the same time, a risk score for every city shall be computed as well, for a detailed info on which measure we used and how we computed it, go to Risk Model. House and rent prices predicted by the SDE model are then fed into CapitalBays business plan algorithm that simulates multiple scenarios for each real estate object. As result an IRR distribution, both levered and unlevered, is computed for each city. These values are then compiled to one single scatter plot that plots risk vs. return for each city. Baseline for all computations is a data set that incorporates city and country-wide level data as well as micro and macro economic data. See Data for a detailed explanation. Finally, an investor can immediately see which city is in any case worse off and which city gives high returns with low associated risk. For the architecture that transforms data to scatter plot refer to Figure 1.

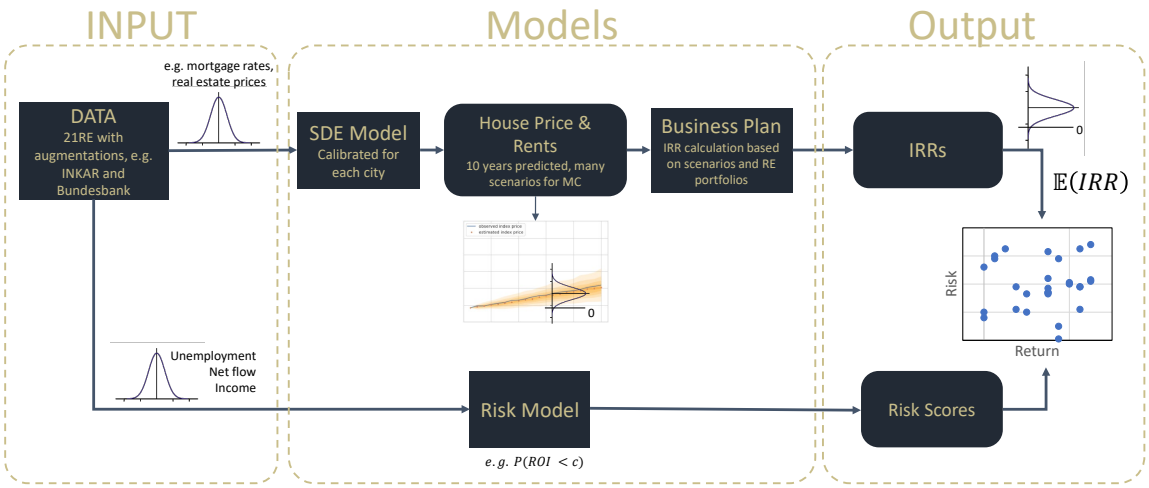


Figure 1. Model and Output Overview. Two streams for separate risk and return prediction lead to the final scatter plot.

Data

Data is the new gold, an often heard sentence in companies these days. Real estate is no exception and as data becomes more accessible and algorithms evolve one needs to leverage on that to stay ahead of competition. Merging traditional features (e.g. property features and market performances) with nontraditional features (e.g. average rating of bars within half mile distance) heavily impact the predicted annual growth of real estate evaluations as Asaftei, Doshi, Means, and Sanghvi, 2018 point out. In this project we also make use of traditional and nontraditional factors to ultimately predict not only return but risk as well.

Data Collection

As main data source 3 tables from CapitalBays cooperation partner *21RE* are used. Both rent and sale prices of real estate objects (traditional data) and nontraditional components are included within those tables. The detailed description of this and all other data sets is described in the next subsection. As 21REs dataset is focused on price and rent data and only basic macro economic data and other basic values are offered, we enrich it with datasets from the public available Indikatoren und Karten zur Raum- und Stadtentwicklung (INKAR) [BBR, 2020a] collection. One manually created table *LinkTable.csv* is used during the extract, transform and load (ETL) process to link city names with 21RE respectively INKAR data. Other minor datasets from e.g. Bundesbank are used, whenever this is the case we denote it in the respective section.

Data Description

LinkTable.csv. When thinking of relational databases, the star schema might be a description how we merged our data sets. In our center (the star) *LinkTable.csv* is our starting point for all other appended data.

Variable	Description	Example Value
City	City Name	Aachen
AGS	Inique City Identifier for 21RE datasets	5334002
n	Number of tiles for this city	3776
INKAR	City name in the INKAR data	Städteregion Aachen

Table 1
LinkTable.csv data: This table is used to link AGS (unique city IDs in the 21RE) with INKAR datasets (INKAR) and one commonly used city name (City)

ags_data.csv. This 21RE table holds a few basic real estate and non-real estate variables on a yearly basis for each city, see Table 4.

cid_income.csv. Provided by 21RE as well, this table only holds information about average annual income per city and year. Table 5

cid_prices.csv. The heart of our data, rent and sale prices for offered real estate objects (Provided by 21RE). Mainly apartments and flats are taken into account since all data is scraped from internet real estate portals. Hence all figures are based on offers, the true sale and rent prices are most probably different. That information is only known by notaries and the tax authority. This data is not on city level but on city tile level (200m x 200m). For our analysis, we aggregated the variables to city level by taking the mean, seeTable 6.

INKAR variables. Since 21RE data sets focus mainly on real estate information, real estate rent and sale prices are influenced by other macro and micro economical variables as well, we include various tables from INKAR, some with yearly info (Table 8), and others only with data from the last recent year (see Table 7).

Data Transformation

As first step we load the link table, cid_income, ags_data and cid_prices. For the price table quarters and full years are derived from the existing year variable (e.g. 2018.75 as 4th quarter of 2018 is more difficult to merge than having two separate variables for both). Then, the data is aggregated on city level, hence taking the mean over all city tiles and per quarter respectively, if flagged, per year. The same is done for the income table. Subsequent we merge income table with link table and price table. Now we merge every INKAR table to the prior created merged table. If flagged, we interpolate data on quarterly basis since INKAR only provides yearly data. Last but not least the same merge and interpolation process is done with the ags data table. Figure 2 shows this process. Auxiliary computations (e.g. fixing enconding faults, creating the link table) and data cleansing were done semi-manually with tools such as Tableau and Excel PowerQuery. The routines are programmed such that additionally needed INKAR files can be downloaded and if saved in the corresponding folder, the routine is automatically loading and merging it. The output is one single data frame, ready to be processed by followed algorithms or saved.

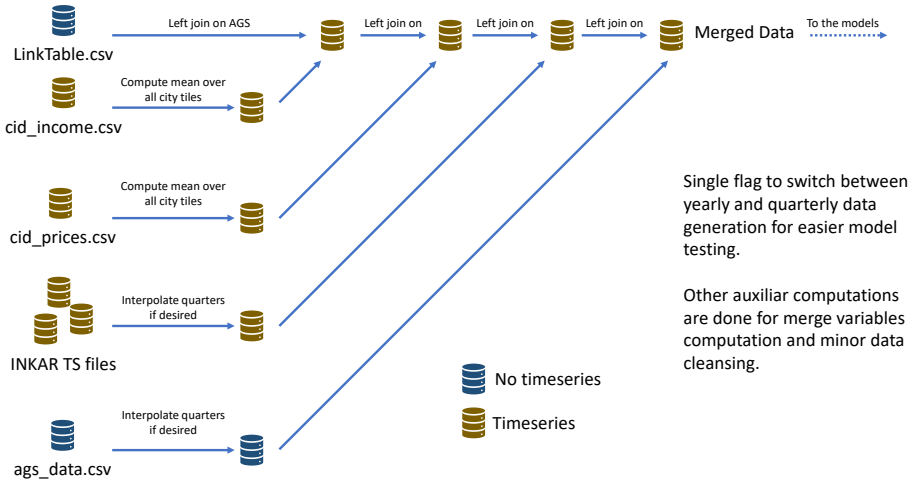


Figure 2. Linking 21RE tables with INKAR files was done step by step. Quarterly interpolation can be switched on or off, the interpolation method can be chosen as well.

Risk Model

Overview

The central question of the Risk Model is about quantifying risk when investing in the German real estate market. More precisely the goal is to identify variables that explain the risk (risk factors) and to calculate a final risk score for each city in the given dataset. The Risk Model is divided into two separate sub-models. The first submodel considers variables with time series and runs a logistic regression. The second submodel deals with mostly constant variables which we refer to as location factors, since we assume that they are constant over time but vary across cities. The outputs of these two models get weighted and summed up to obtain an overall final risk score.

The starting question is how to define risk in the context of investing in real estate, while also considering the data available to us. Many variables capture the risk of an investment and many are also subjective, which makes this task a difficult one. A good introduction for real estate investing and risk management is given by [Manganelli, 2014]. For the logistic regression, i.e. the first model, we concluded that risk is the possibility of an event with undesired effects occurring. Undesired effects for the investor would be that the return on investment (ROI) gets negative or below a certain threshold u. The ROI describes the net profit over a period divided by the cost of investment and is a performance measure, generally speaking. We explain how we define the ROI in our context in the chapter Logit Model.

The assumption of the second model, the Location Factor Model, is that some variables are constant over time for a city and therefore have a stabilising effect on the real estate market in this city. In this model the risk factor, which is the dependent variable again, is captured by the Sharpe ratio, which is explained in the chapter . The workflow for the Risk model was iterative: developing different risk factors, refining the dataset, adding new data, combining different variables etc. In the following sections we explain how we started out and what approaches we took to obtain better results as our understanding of the

topic grew. We divided our workstream in a data pipeline, the different regressions and post-processing as subtasks.

Data Pipeline

As mentioned above, we divide our variables into two groups for the two models respectively. The Logit model gets all variables where a time series is available with two exceptions where the variance over time is very small.
The two exceptions are *Ein_und_Zweifamilienhäuser* and *Großunternehmen* which get fed into the Location Factor Model together with the other variables that are considered constant over time but varying across cities. The variables with time series get merged in the golden source as described in Data Transformation. However before running the Risk Models, we do some more preprocessing as explained in the following. It should be noted that the 21st dataset is our main dataset and it determines the timeframe. It spans the years from 2011-2017 and is available in quarterly time steps for the price and rent data and yearly available for the other variables. The variables from INKAR are on a yearly basis only, and contain varying time series. We implement a boolean which allows to switch between yearly and quarterly time frames. For a few INKAR variables the years 2011 or 2017 are missing. Here we apply a backward- or forward-fill as shown in Figure 33 (appendix) or an interpolation depending on the variable.

We continue with the calculation of the dependent variable, which shall explain risk. Note that for a logistic regression we need a binary outcome variable. The first approach we implemented is given by following equation:

$$grossprofit := \frac{Rentalincome}{PurchasingPrice} < u \tag{3}$$

where u is a certain threshold. As the project went on, we define a more sophisticated way to calculate the profit:

$$profit_{c,t} = \frac{rent_{c,t} + purchase_{c,t+1}}{purchase_{c,t}} - 1 \tag{4}$$

where c stands for the city and t for the timestep. All rent- and purchasing prices are in euro per square meter. There were several reasons why we decided to adapt the definition of profit for the risk model, as will be explained in the following paragraph.
If we used the industry standard profit definition, large cities have a systematically higher risk, because they have a systematically lower grossprofit. A higher risk for the largest German cities is however not in line with previous research on risk in real estate markets over Germany, see [Bulwiengesa, 2019]. Furthermore our definition of profit also incorporates the purchasing price of the next year after a given year. Therefore it also incorporates the risk of a change in purchasing prices from one year to another and not just the ratio of rent to purchasing prices. Especially in large cities with a more active real estate market, this seems to be a factor which should be included, since it is important for investors who plan to buy and sell real estate on a more regular basis. We assume that the holding period of an asset from professional investors is shorter compared to the holding period of smaller or private real estate investors.

For the rent price (and therefore also for the profit and the thresholds) we implement one- and two-period lags as well. The rationale behind this is that the contractual rent prices probably lag behind the market rent prices, because of existing / older rental contracts.
For the independent factors we include the growth values and calculated them as follows:

$$\frac{\Delta x}{\Delta t} = \frac{x_t - x_{t+1}}{x_{t+1}} \tag{5}$$

where x stands for an independent factor. Doing this calculation, one NaN value will be generated for each city over the time series, which we interpolate again. The table Data Pipeline shows the final dataframe for the regression schematically.

City	Year	Grossprofit	c-levels (Regressand)	21st+INKAR Macrodata	Growth-variables	Quadrant-Data
Aachen	2011		0			
Aachen	2012		1			
...	...	...	...	...	...	...
Wuerzburg	2015		1			
Wuerzburg	2016		1			
Wuerzburg	2017		0			

Table 2
Data Pipeline for Log-Regression

Quadrants

When analysing and visualising the data in the very beginning of the project, we realised a very interesting fact when comparing the rent and purchasing prices on tile level for the different cities. There were significant differences between the price to rent ratio in the different tiles of a city. A rational investor that is only interested in the industry standard profit calculation will always choose the real estate object with the lowest price to rent ratio, because the lower the price to rent ratio is for given real estate object at a certain point in time, the higher the grossprofit. A real estate object in a tile with price p and rent r will ceteris paribus always be dominated by a real estate object with price p and rent $r + \epsilon$. Assuming real estate objects that are solely defined by rent and price, then the long term equilibrium of a real estate market in a given city should be defined by a constant price to rent ratio over all tiles. Based on this train of thought we define four different quadrants.

We call the lower right quadrant the "Growth Quadrant", because ceteris paribus the purchasing prices in that quadrant should rise to reach the market equilibrium of the price to rent ratio. Correspondingly we call the upper left quadrant "Shrink Quadrant", because we expect the purchasing prices in those tiles to shrink. Obviously the rents could also change over time to reach the assumed price to rent ratio equilibrium and in reality we expect a mix of shrinking (rising) rent prices and rising (shrinking) purchasing prices for the Growth (Shrink) Quadrant following our assumption of a long term equilibrium with a constant price to rent ratio over all tiles of a city.

We test this assumption with an exploratory analysis for different German cities by comparing the development of the purchasing prices of tiles that were part of the shrink

Figure 3. Rent vs Sales Prices in Munich for the 1st quarter of 2011 with a horizontal line for the mean purchasing price and a vertical line for the mean rent price.

quadrants vs. the ones that were part of the growth quadrants in a given quarter. We also investigated how long this effect lasts on average for the different cities (see).

If this observation really has a significant impact on the prices in the different cities, it will probably also have a significant impact on the profit for each city in each year. We test this by building two features and include them in the Logit Model. First of all we define a given tile s in a city c as part of the growth quadrant in timestamp t (quarter or year) if

$$purchasingprice_{s,t} < mean(purchasingprices_{c,t}) \wedge rentprice_{s,t} > mean(rentprices_{c,t}) \quad (6)$$

and part of the shrink quadrant if

$$purchasingprice_{s,t} > mean(purchasingprices_{c,t}) \wedge rentprice_{s,t} < mean(rentprices_{c,t}) \quad (7)$$

For every y and city c we calculate the

$$GrowthShrinkRatio = \frac{\#tilesgrowth_{y,c}}{\#tilesshrink_{y,c}} \quad (8)$$

and include it as a factor for the logistic regression. Additionally we calculate the variance of the price to rent ratio for every city c and year y over all tiles and add it as a feature into the Logit Model as $PriceRentRatioVariance_{y,c}$.

Logit Model

We define risk in the Logit Model as the probability of the profit for a certain year y and city c being smaller than a certain threshold u.

$$profit_{c,y} < u \tag{9}$$

We label a given observation that does not hit the target u with 1 and one that surpasses u with zero and therefore end up with a binary dependent variable y for a logistic regression.

$$y = \alpha + \beta * X + \epsilon \tag{10}$$

In a first step we have to come up with reasonable level for u to use when building the dependent variable y. We test many different levels of u (see) and look for a level of u that labels roughly 1/3 of all observations as 1 in y to leave us with a not completely misbalanced dataset while still considering that risk is rare by definition. We tested country-wide and over time constant features for u, but also city and country specific thresholds, that change over time and from city to city. In total we tested 40 different levels. Out of all levels for u, 5% had the highest measure of certainty R^2. We therefore use a level of 5% for u to come up with a risk score in our model.

In a second step we define which independent variables to be in X. The Risk Model data pipeline delivers 143 possible features (absolut and growth values) to consider in the Logit Model. We perform more feature transformation to transform those variables into the best possible set of features to use in a Logistic Regression:.

Drop Missing Values: Some of the INKAR data was not available for a few cities. We dropped every column completely that had at least one missing value.

Remove correlated features: In a first step we remove one of two features with a Pearson Correlation Index higher than 0.8.

Include Interactions: We include interactions and add a variable for the product of every 2-tupel combination of variables. We call those new features according to following schema: $mult\_ < column1 > \_x\_ < column2 >$.

Include Lags: We calculate the lagged values for all variables and include them as a new variable. For every feature we consider lags of one and two years. We call those new features according to following schema: $< column1 > \_lagged\_1$ for a given value of the previous year and $< column1 > \_lagged\_2$ for a given value from two years before. We replace missing values with the mean over time of that value for given a city.

Scale Features: In order to increase comparability of the different features we then scale all features to be between 0 and 1 using a MinMaxScaler.

Remove correlated features again: After adding interactions and lagged features we again remove one of two features with a Pearson Correlation Index higher than 0.8.

Recursive Feature Elimination: In a last step we run a Recursive Feature Elimination Algorithm to identify the 20 factors that are most significant.

We end up with 20 independent features in X and run the logistic regression. See the results of the logistic regression for $u = 5\%$ in figure 34.

We define all features with a $p\_value < 5\%$ as significant. Those features seem to have a significant explanatory effect on whether the profit reaches a certain target or not, so we use these significant features to calculate a risk score for each year and all the cities. We multiply the estimated coefficients of the significant features with the observed relevant data of each city and year and transform the thereby predicted $log(\frac{p}{1-p})$ into the predicted probability of default p using following formula

$$p = \frac{e^{\beta_1 * X_1 + \beta_2 * X_2 + .. + \beta_n * X_n}}{e^{\beta_1 * X_1 + \beta_2 * X_2 + .. + \beta_n * X_n} + 1} \tag{11}$$

We thereby generate a yearly changing probability of default between 0 (low risk of default) and 1 (high risk of default). We will combine that probability of default with the risk score generated in the location factor model.

Location Factor Model

As stated before we include all possible explanatory factors that are not present as a time-series or have a very small variance in our Location Factor model. We generally assume two major sources of risk that investors face in the different German cities. One source being yearly changing variables that we include in the Logit Model. For the Location Factor Model we assume that different cities also posses differentiating features that stay more or less stable over time and make cities more or less attractive to companies and individuals and therefore also interesting to investors. Assume for example parliamentary districts that will be there no matter how the economy is developing, generating jobs and therefore having a stabilising positive effect on rent and purchasing prices. Other stabilising factors could be: Large international companies, green areas that improve quality of living, a good connection to highways, airports and large train stations and similar. All those features usually only change gradually over time, if at all.

To summarise we assume that those factors can decrease risk over time and make the profit in certain cities less dependent on short term fluctuations in the current economic situation. In the location factor model we therefore aim to identify exactly those features that have stabilising effects on the development of the profit in a certain city. We investigate two options of quantifying positive stability of profit development over time, which will serve as the dependent variable in a linear regression later on.

The variance of the growth of the profit We define a stable investment as generating a steadily growing profit over time. We therefore calculate the growth of the profit from one quarter to another for every city c and calculate the variance for it over the whole available timeframe T. A high variance represents a risky investment opportunity, a

low variance a stable and therefore less risky investment opportunity.

$$Var\left(\frac{profit_{c,qy} - profit_{c,qy-1}}{profit_{c,qy-1}}\right), \text{over all } qy \in T. \tag{12}$$

The disadvantage with this approach is that is does not differentiate between positive and negative growth. According to this measure, a constant negative growth of the profit would be as desirable as a constant positive growth. Furthermore a positive jump in the profit is certainly something desirable for the investor but is treated as *unstable* when regressing on this variable. We therefore calculate a second risk score to measure long term positive stability of an investment in a certain city.

The Sharpe Ratio To overcome the shortcoming of the first approach, we implement the Sharpe ratio as a dependent variable. In finance this ratio tries to capture the performance of an investment compared to a risk-free asset after adjusting for its risk. Generally speaking it is calculated by the difference of the returns of an investment and the risk-free return, divided by the volatility of the investment. We tried to proxy the risk-free rate as the median profit over Germany during the whole timeframe T, but ended up with mediocre statistical results. We therefore assume a risk-free return of 0 and just compare the mean profit of a city c over the whole timeframe T with the standard deviation of the profit over the whole timeframe. Schematically this can be written as:

$$\frac{mean(profit_{c,T})}{std(profit_{c,T})}, \text{over } qy \in T. \tag{13}$$

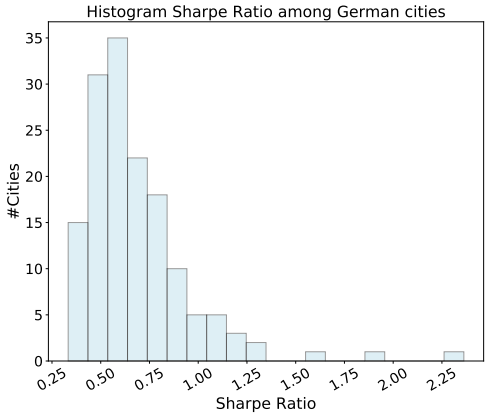


Figure 4. Sharpe Ratio over the whole timeframe for all German cities.

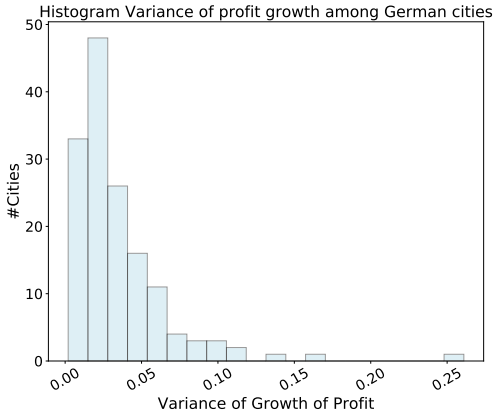


Figure 5. Variance of profit over the whole timeframe for all German cities.

We then run two linear cross-sectional regressions. One with the Sharpe Ratio, one with the variance of the profit growth as the dependent variable y and over time constant features X.

$$y = \alpha + \beta * X + \epsilon \tag{14}$$

In order to derive the best possible set of features X, we choose a similar approach as in the Logit Model. A description is given in the appendix.

Find the regression results using the Sharpe Ratio in figure 35 and using the the variance of the profit growth in figure 36.

We end up with a set of significant features with a $pvalue < 5\%$. We use those features to predict the Sharpe Ratio and the variance of the profit growth for all cities based on only the significant location factors. We then transform the predicted $SharpeRatio = SharpeRatio * (-1)$ to be able to compare the two Location Factor Model risk scores. A high Sharpe Ratio is actually desirable and not something we would define risky or unstable, whereas a high variance of the profit growth is rather unstable and therefore risky.
In a last step we normalise the predicted Sharpe Ratio and the predicted variance of the profit growth using a MinMaxScaler to end up with a Risk Score between 0 and 1.

Aggregated Risk Model

So we end up with a probability of default from the Logit Model for every city and year and a constant risk score between 0 and 1 from the Location Factor model for every city. In a final step we combine both risk measures using a weight w.

$$AggregatedRiskScore = w * RiskScore + (1 - w) * ProbabilityOfDefault \tag{15}$$

Figure 6 visualises how the two models work together.

In our final Risk Model, we use a level of 5% for u in the Logit Model and we use the Sharpe Ratio as the dependent variable in the Location Factor Model, because of the better measure of certainty R^2 in the regression and due to the clear shortcomings of the variance of the profit growth described earlier.

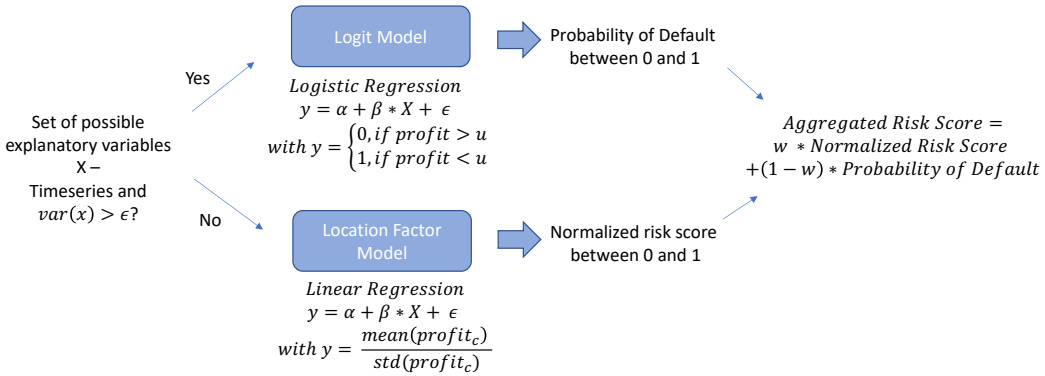


Figure 6. Overview Aggregated Risk Model

Outlook and Limitations

We had to take a few assumptions when building this model, which limit our findings to a certain extend.

Different way of calculating profit By deviating in the way we calculated the profit and thereby trying to capture the risk of a change in the purchasing price, we assume that real estate can be traded quite frequently, much like a liquid asset such as stocks or commodities. This may not always be the case. In reality investors might often be bound to a real estate object they bought and it might take weeks to sell it again, which makes it difficult to react to changes in the purchasing prices instantly. But since we are interested in risk, just use the profit as a proxy to quantify risk and are not interested in the profit per se, this approach is still the best possible solution, even though it is based on a rather strong assumption.

Quadrants Even though the quadrant idea seems logical, we did not prove a stable effect all over Germany of tiles from Growth Quadrants rising in purchasing prices and tiles from Shrink Quadrants shrinking. Please find a more detailed discussion and an evaluation of the results of the exploratory analysis in the appendix .

City unspecific level of u We tested many different levels for u in the Logit Model, but the city specific levels performed worse than constant levels. We therefore actually identified country wide risk drivers and not city specific risk drivers. The macroeconomic data from 21st, which we used in the Risk Model was only available on a yearly level. We therefore only had seven data points per city, one per year. In order to identify city specific risk drivers one could increase the sample size, try to get macroeconomic data on a quarterly level as well and then run the Logit Model on city level, ending up with city specific regression results and significant risk drivers.

Location Factor Model Since we aggregate all the price data into just one number for each city, we end up with a rather small dataset that just contains 149 observations, one for each city. Consequently the statistical characteristics of our linear regression are nowhere perfect. One could think of not running a regression, dispense finding the location factors and just incorporate the historic Sharpe Ratio or variance of the profit growth in the risk model. One could do this by just adding risk to cities with a past high variance of the profit growth (low Sharpe Ratio) and reducing the risk for cities with a past low variance of the profit growth (high Sharpe Ratio).

Data in General In general we heavily rely on only two data sources. We performed a robustness test of the rent prices we received from 21st using official data of the *BSSR*, but nevertheless one could think of investigating different data sources if possible and backtest data and models. Specifically a more up to date and longer time series could generate better regression results. Another idea would be to perform an Extreme Value Analysis to estimate the probability of unusual large jumps in the German real estate market. As mentioned before, the Gutachterausschüsse provide the true contractual data of rent and purchasing prices but not in a countrywide database. One could think of a collaboration with these institutions and perform an analysis for a single city using this valuable data.

SDE Model

The pattern of housing prices and their volatility can be used as an indicator to understand housing market dynamics. There exist a variety of models that can be used to study the dynamics of this market. Most of them are based on the renowned Black-Scholes-Merton model which is originally used to price financial options. Different studies show that applying those theories to the Real Estate market works as well. In this section we want to improve and refine the work done by our predecessors in Data Innovation Lab - Capital Bay WS 2019-2020 [Altschäffel, Dragoi, Mendez, Tamada, and Wiyoga, 2020] in order to optimize the prediction of house prices and rent prices for the residential market in Germany for the next ten years. These predictions will be used to feed the business plan of Capital Bay.

Theoretical Model/ Research

As a starting point, Data Innovation Lab - Capital Bay WS 2019-2020 used the model defined in [Yilmaz and Selcuk-Kestel, 2018]. Analyzing this model, a stochastic process is considered in which the h_t House Price Index is defined recursively and dependent on the interest rate r_t. The dependency between interest rate and housing prices is based on the economic reasoning that a considerable part of the housing market is financed through debt and therefore the house prices naturally incorporate the cost of borrowing. Due to the fact that interest rate dictates house prices, the dependency between h_t and r_t is modeled as a mean-reverting recursive process.
In this model a natural long-term equilibrium level for h_t (denoted by μ_h) and r_t (denoted by μ_r) is assumed. At every step, the process including an added noise term converges towards the mean. The dynamics of the model are described by the following system:

$$\frac{dh_t}{h_t} = \lambda(\mu_h - r_t)dt + \sigma_h dZ_t^h, \tag{16}$$

$$dr_t = k(\mu_r - r_t)dt + \sigma_r dZ_t^r. \tag{17}$$

We have seen in the work of [Altschäffel et al., 2020] that this model is a good choice to describe the dynamics of the housing market. Nevertheless, there are certain points that can be improved to reflect the actual behavior of the market precisely and therefore perform a more accurate prediction.

Since the model is based on Black-Scholes-Merton, we decided to improve and refine the model following the same stream of ideas. For the sake of generality, we decided to consider the volatility as a time dependent variable and to incorporate a term that describes possible jumps in the price, resulting in an improvement for the dynamics of the housing price model. (Nevertheless, an analysis of the data from the past years will result in the decision of assuming a constant volatility when implementing the actual model. Additionally, we will not consider jumps for the Rent Price Index in the parameter estimation. The reason for this is that history displayed less and smaller jumps as the Housing Price

Index, which makes it even harder to calibrate the parameters due do the limited amount of timesteps. Further explanation can be found in the section describing the parameter estimation.)

Considering the factors mentioned above and comparing different studies on option pricing and volatility with the data available for the residential market, we noticed a huge similarity between the characteristics of stock and housing prices. Thus, we decided to choose the Bates model as one of the best fitting models: a stochastic volatility model with price jumps for the stock pricing. In this model, the dynamics of the underlying asset are driven by both: a Heston stochastic volatility and a compound Poisson jump process. The application of the model, in its general expression, to the residential house market to estimate and predict the housing price, results in the following dynamics:

$$dh_t = h_t \lambda_h (\mu_h - r_t)dt + h_t \sqrt{v_t} dZ_t^h + h_t dH_t^h, \tag{18}$$

$$dv_t = k_v(\mu_v - v_t)dt + \sigma_v \sqrt{v_t} dZ_t^v + dH_t^v, \tag{19}$$

$$dr_t = k_r(\mu_r - r_t)dt + \sigma_r dZ_t^r, \tag{20}$$

$$dZ_t^h dZ_t^v = \rho_1 dt, \tag{21}$$

$$dZ_t^h dZ_t^r = \rho_2 dt, \tag{22}$$

where h_t denotes the Housing Price Index, v_t the volatility of the index following a Cox-Ingersoll-Ross (CIR) process and r_t the mortgage rate defined by a generalized Ornstein-Uhlenbeck (OU) process. Z_t^h, Z_t^v, Z_t^r are correlated Brownian motions with correlations ρ_1 and ρ_2. H_t^h is a compound Poisson process with intensity λ and independent identical distributed jumps.

In addition to this system of stochastic differential equations describing the Housing Price Index, we also want to consider the housing rent market. On the basis of the housing price dynamics, we model a system including rent and aim in determining an economic factor which heavily affects the rental market.
Therefore, we define three equations to describe the complete dynamics, that can be added to the system of eq. 18-22:

$$dm_t = m_t(\mu_m + k_{m1}\frac{dh_{t-l_1}}{h_{t-l_1}} + k_{k2}\frac{df_{t-l_2}}{f_{t_2}})dt + m_t \sigma_m dZ_t^m + m_t dH_t^m, \tag{23}$$

$$df_t = k_r(\mu_f - f_t)dt + \sigma_f dZ_t^f, \tag{24}$$

$$dZ_t^m dZ_t^f = \rho_3 dt, \tag{25}$$

where m_t denotes the Rent Price Index and f_t an additional factor affecting the Rent Index (i.e.vacancy) defined by a generalized Ornstein-Uhlenbeck (OU) process. Z_t^m, Z_t^f are correlated Brownian motions with correlation ρ_3, and H_t^m is a compound Poisson process.

Discrete time model

For the implementation of the model and due to the limited availability of data, we use a discretized version of the model, where the process runs in step-wise increments. Considering $1 \leq i \leq T$, we index the time series as t_i and each time step is written as $\Delta t = t_{i+1} - t_i$. We denote by $\Delta Z_t^h = Z_{t_{i+1}}^h - Z_{t_i}^h$ the difference of the Brownian motion of the House Price Index at each time-step and by $\Delta H_t^h = H_{t_{i+1}}^h - H_{t_i}^h$ the difference of the compounded Poisson process at each time-step. This method can be applied to the other factors analogously.

Rewriting the system of equations 18-25 into discrete time steps using the Euler-Maruyama method leads to the following formulas:

$$\begin{aligned}
h_{t_{i+1}} &= h_{t_i} + h_{t_i}\lambda_h(\mu_h - r_{t_i})\Delta t + h_{t_i}\sqrt{v_{h_{t_i}}}\Delta Z_t^h + h_{t_i}\Delta H_t^h, \\
v_{t_{i+1}} &= v_{t_i} + k_v(\mu_v - v_{t_i})\Delta t + \sigma_v\sqrt{v_{t_i}}\Delta Z_t^v + \Delta H_t^v, \\
r_{t_{i+1}} &= r_{t_i} + k_r(\mu_r - r_{t_i})\Delta t + \sigma_r\Delta Z_t^r, \\
m_{t_{i+1}} &= m_{t_i} + m_{t_i}(\mu_m + k_{m1}\frac{\Delta h_{t-l_1}}{h_{t-l_1}} + k_{m2}\frac{\Delta f_{t-l_2}}{f_{t-l_2}})\Delta t + m_{t_i}\sigma_m\Delta Z_t^m + m_{t_i}\Delta H_t^m, \\
f_{t_{i+1}} &= f_{t_i} + k_f(\mu_f - f_{t_i})\Delta t + \sigma_f\Delta Z_t^f.
\end{aligned} \tag{26}$$

After a careful analysis of the Housing Price Index volatility, we observed a relatively steady behavior in most of the cities, with only few peaks for some cities, see Figure 7. For this reason, we decided to simplify the dynamics of the system neglecting the time dependent stochastic volatility.

The numerical solutions of the equations are estimated using the Implicit Euler Method due to its strong property of convergence and stability, see [Higham and Kloeden, 2006].

$$\begin{aligned}
h_{t_{i+1}} &= h_{t_i} + (1-\theta)h_{t_i}\lambda_h(\mu_h - r_{t_i})\Delta t + \theta h_{t_{i+1}}\lambda_h(\mu_h - r_{t_i})\Delta t + h_{t_i}\sqrt{v_h}\Delta Z_t^h + h_{t_i}\Delta H_t^h, \\
&= \frac{h_{t_i} + (1-\theta)h_{t_i}\lambda_h(\mu_h - r_{t_i})\Delta t + h_{t_i}\sqrt{v_h}\Delta Z_t^h + h_{t_i}\Delta H_t^h}{1 - \theta\lambda_h(\mu_h - r_{t_i})\Delta t}, \\
r_{t_{i+1}} &= r_{t_i} + k_r(\mu_r - r_{t_i})\Delta t + \sigma_r\Delta Z_t^r,
\end{aligned}$$

here θ is a parameter in $[0, 1)$.

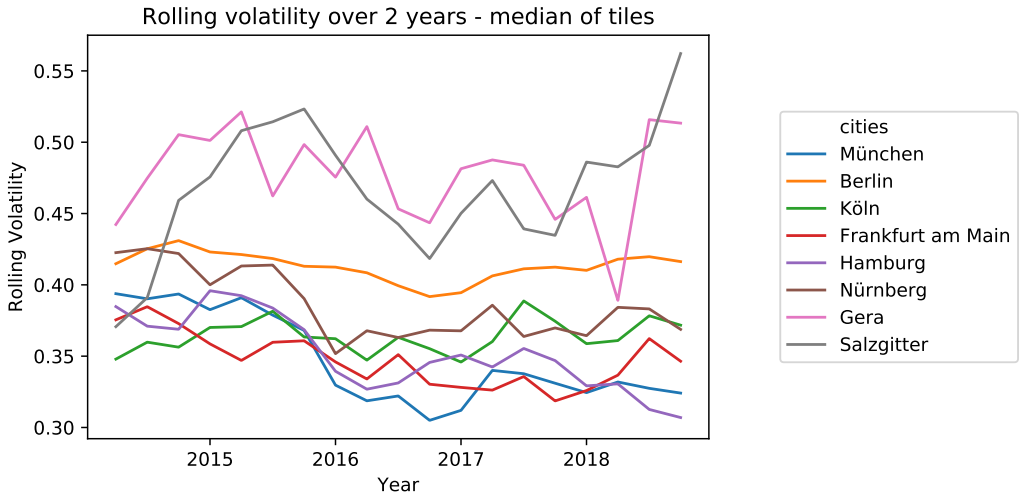


Figure 7. This figure shows the rolling volatility over two years by taking the median value of all tiles of the Housing Price Index displaying eight cities in Germany.

Analogously using the Implicit Euler Method for the Rent Price, we derive

$$\begin{aligned} m_{t_{i+1}} &= m_{t_i} + (1-\theta)m_{t_i}(\mu_m + k_{m1}\frac{\Delta h_{t-l_1}}{h_{t-l_1}} + k_{m2}\frac{\Delta f_{t-l_2}}{f_{t_2}})\Delta t \\ &\quad + \theta m_{t_i}(\mu_m + k_{m1}\frac{\Delta h_{t-l_1}}{h_{t-l_1}} + k_{m2}\frac{\Delta f_{t-l_2}}{f_{t_2}})\Delta t + m_{t_i}\sigma_m \Delta Z_t^m + m_{t_i}\Delta H_t^m, \\ f_{t_{i+1}} &= f_{t_i} + k_r(\mu_f - f_{t_i})\Delta t + \sigma_f \Delta Z_t^f, \end{aligned}$$

Notice that the Implicit Euler method simplifies to an Euler-Maruyama solution at $\theta = 0$.

Model Implementation

Having formulated the theoretical dynamic system, we start with the implementation. As the documentation [Altschäffel et al., 2020] thoroughly discussed the sources of data, the analysis of the data and further outcomes, we shortly summarized and added the most important points concerning the SDE model in the Appendix.

Vacancy as additional factor: In order to determine the additional factor by finding a variable that shows a high correlation with rent and is suitable from an economic point of view, let us choose vacancy as the best fitting factor. The vacancy rate which is determined by the ratio of vacant apartments is used in various papers to analyze and predict rent and is often described as one of the most important factors in this context [McDonald, 2000]. In addition, a multi-linear regression analysis using normalized housing prices and vacancy data to predict rent displayed in Figure 19 shows that both factors are significant on a $\alpha = 0.05$-level. Due to difficulties in availability of data, the vacancy data is taken from the database Empirica that provides the vacancy rate only for 66% of the cities to be analyzed. Viewing the data, Figure 8 illustrates

the Vacancy rate over time for different cities and Figure 20 in the Appendix displays a histogram of the relative growth of Vacancy over all cities and timesteps. By analyzing the data, one can determine a negative trend on the vacancy rate, which can be set in the context of supply and demand: As the vacancy rate and therefore the supply of available apartments decreases, the price for housing and especially rent prices increases, which follows our displayed data of housing and rent prices in Figures 15 and 16 .

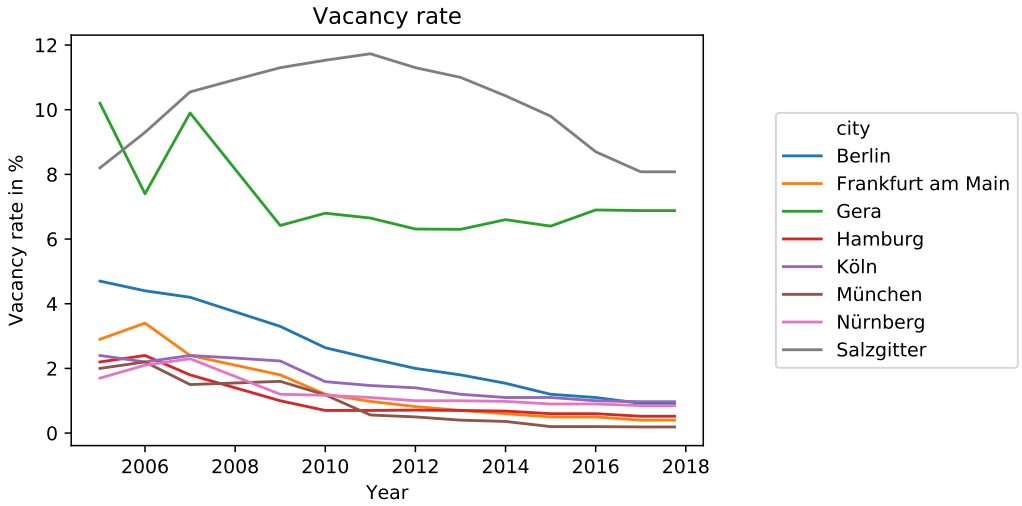


Figure 8. Vacancy data for specific cities

Pipeline for implementation: After the preparation of the data, the simulations of housing and rent price indices for the next ten years are derived with the following procedure: First, we implement a numeric solution (see discrete model) and analytically calculate the parameters for the interest rate equation on country-level using the data of the available time series. Taking the interest rate estimation as input for the House Price Index equation, the parameters of the HPI equation are calibrated for each city: The diffusion parameters can be derived analytically, whereas the parameters of the drift part are determined by optimization. Separately, we implement a numeric solution, as well as the analytical parameter estimation of the additional factor in the same way as calculated for the interest rate equation. Feeding the solution of the housing prices and the vacancy rate in the rent equation, the Rent Index can be estimated by a similar parameter estimation and optimization as for the HPI equation. Having derived all parameters of the dynamic system, the Interest rate and Vacancy rate, as well as the Housing and Rent Prices Indices can be predicted for the next 10 years on quaterly level. Before feeding these predictions in the Business Plan, a cubic interpolation transforms the data into monthly timesteps.

Parameter Estimation

For the parameter estimation, we decided to use two different methods to estimate the parameters of the dynamic system.

Interest rate and additional factor: As the interest rate, as well as the additional factor are defined by a generalized Ornstein-Uhlenbeck process, we can analytically calculate the parameters of the drift of these two equations using Maximum-Likelihood. (In order to derive suitable starting values for the HPI equation, we exclude the jump process from the housing price equation and are able to calculate the parameters for this modified equation analytically, as well.) The Maximum-Likelihood can be formulated as follows:

$$(\hat{k}_r, \hat{\mu}_r) = argmin \sum_{i=1}^{N-1} \left(r_{i+1} - r_i - k_r \left(\mu_r - r_i\right) \Delta t\right)^2, \tag{27}$$

$$(\hat{k}_f, \hat{\mu}_f) = argmin \sum_{i=1}^{N-1} \left(f_{i+1} - f_i - k_f \left(\mu_f - f_i\right) \Delta t\right)^2, \tag{28}$$

$$(\tilde{\lambda}_h, \tilde{\mu}_h) = argmin \sum_{i=1}^{N-1} \left(\left(\frac{h_{i+1} - h_i}{h_i}\right) - \lambda_h(\mu_h - r_i)\Delta t\right)^2. \tag{29}$$

The derivations leading to explicit solutions of this optimization problem are calculated in Appendix of [Altschäffel et al., 2020]. The resulting parameters can be obtained as follows:

$$\begin{aligned}
\tilde{\lambda}_h &= \frac{1}{\Delta t} \frac{\sum_i r_{t_i} \sum_i \frac{h_{i+1}-h_i}{h_i} - (N-1) \sum_i r_{t_i} \sum_i \frac{h_{i+1}-h_i}{h_i}}{(N-1) \sum_i r_{t_i}^2 - \left(\sum_i r_{t_i}\right)^2}, \\
\tilde{\mu}_h &= \frac{\sum_i r_{t_i}^2 \sum_i \frac{h_{i+1}-h_i}{h_i} - \sum_i r_{t_i} \sum_i \frac{h_{i+1}-h_i}{h_i}}{\sum_i r_{t_i} \sum_i \frac{h_{i+1}-h_i}{h_i} - (N-1) \sum_i r_{t_i} \sum_i \frac{h_{i+1}-h_i}{h_i}}, \\
\hat{k}_r &= \frac{1}{\Delta t} \frac{\sum_i r_{t_i} \sum_i (r_{i+1} - r_i) - (N-1) \sum_i r_{t_i} \sum_i r_{i+1} - r_i}{(N-1) \sum_i r_{t_i}^2 - \left(\sum_i r_{t_i}\right)^2}, \\
\hat{\mu}_r &= \frac{\sum_i r_{t_i}^2 \sum_i (r_{i+1} - r_i) - \sum_i r_{t_i} \sum_i r_{i+1} - r_i}{\sum_i r_{t_i} \sum_i (r_{i+1} - r_i) - (N-1) \sum_i r_{t_i} \sum_i r_{i+1} - r_i}, \\
\hat{k}_f &= \frac{1}{\Delta t} \frac{\sum_i f_{t_i} \sum_i (f_{i+1} - f_i) - (N-1) \sum_i f_{t_i} \sum_i f_{i+1} - f_i}{(N-1) \sum_i f_{t_i}^2 - \left(\sum_i f_{t_i}\right)^2}, \\
\hat{\mu}_f &= \frac{\sum_i f_{t_i}^2 \sum_i (f_{i+1} - f_i) - \sum_i f_{t_i} \sum_i f_{i+1} - f_i}{\sum_i f_{t_i} \sum_i (f_{i+1} - f_i) - (N-1) \sum_i f_{t_i} \sum_i f_{i+1} - f_i}.
\end{aligned}$$

Housing and Rent Price Index: 1. Lambda of Jump process

In order to estimate the parameters of the Housing and Rent Price Index, we discard the jumps from the respective time series as described in the Appendix section Jump size. By using the discarded jump data for each city, the parameters of the compound Poisson processes, the jump intensity λ and jump size N

defined as a distribution with jump mean and jump variance, are estimated. We consider the distribution associated with the time of occurrences of the jumps to derive the intensity of the Poisson process λ, described by an exponential distribution. Therefore, we use a maximum likelihood estimation to find the parameter λ of the exponential distribution.
First, we define the likelihood function as

$$L(\lambda) = \prod_{1}^{n} \lambda exp\left(-\lambda x_i\right) = \lambda^n \exp\left(-\lambda n \bar{x}\right), \tag{30}$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the mean of the sample and $x = (x_1, ...x_n)$ an independent and identical distributed sample. The maximum likelihood estimate for the parameters λ can be calculated as follows:

$$\hat{\lambda} = \frac{1}{\bar{x}} = \frac{n}{\sum_i^n x_i}. \tag{31}$$

Due to the limited availability of data, we noticed that it is not possible to accurately estimate different jumps based on the time series, see Figure 21. For this reason, we decide to consider the economic assumption that the market follows cycles, usually lasting 8 years. Incorporating this economic view, we estimate the λ of the Poisson process as the constant probability that one jump occurs in a cycle independent of cities, considering quarterly time steps. Figure 22 in the Appendix displays a barplot counting the number of jumps per simulation for this fixed λ.
Further explanations and the calibration of the Jump size distribution can be derived from the Appendix.

2. Optimization of drift parameters:

   Having estimated suitable starting values of the drift parameters for the House Price Index, an optimization algorithm is used to find the right drift parameter values for each city. The parameters to be optimized, $\tilde{\lambda}_h$ the rate at which the interest rate reverts to $\tilde{\mu}_h$ considering a long-term equilibrium level, are evaluated by minimizing a loss function separately for each city. This loss function calculates the differences of the observed median value over tiles at every time-step Y^{obs}, and the mean value over 1000 simulations at every time-step $\hat{Y^{\theta}}$. The vector θ^* includes the optimal parameters for $\hat{\lambda}_h$ and $\hat{\mu}_h$.

$$\theta^* = argin f \left\| Y^{obs} - \hat{Y^{\theta}} \right\|^2 = argin f \sum_i (Y_i^{obs} - \hat{Y}_i^{\theta})^2, \tag{32}$$

   The same method is used for the Rent Price by defining θ^* as the vector of the parameters $\hat{\mu}_m$, k_{m1} and k_{m2}.

Analysis of the SDE model

In the following section, we present the estimations of our calculations and analyze the output regarding its validity.

First, we want to display our estimations and set it in relation with the actual data. Figure 9 compares 100 simulations of the House Price Index calibrated for Munich with the respective time series available from 21st Data. We notice an increase in the volatility of the estimated data as we go further in time. In detail, the plot displays the median of all estimations at each time step (dotted line) which follows the upward trend of the observed House Price Index between the years 2011-2018 pretty closely. The orange marked area indicates the region between the median and $i * std(housing\ price\ estimations)$, where $i \in \{1, 2, 3, 4\}$. The fact that the orange area surrounds the observed median tells us, that this model is a good choice for the prediction of the House Price Index and that the prediction is reliable.

Similarly for the Rent Price Index, Figure 10 shows well fitting estimations over the years 2013-2018, when taking the median of 100 estimations into consideration. Note that the Rent Price Index starts at year 2013, as we assume that rent prices follow housing prices by two years, see Appendix.

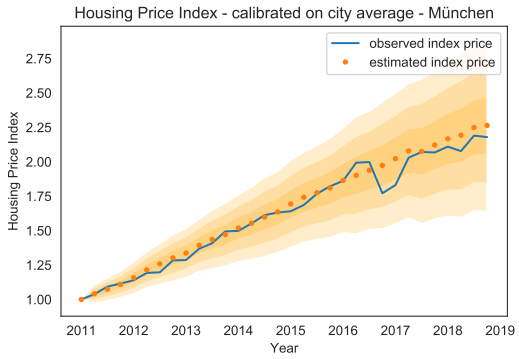


Figure 9. Observed and estimated Housing Price Index over 2011-2018 calibrated on Munich

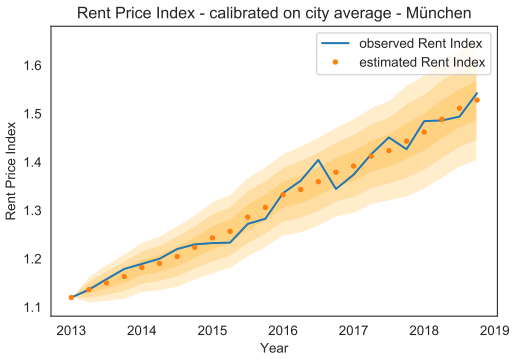


Figure 10. Observed and estimated Rent Price Index over 2013-2018 calibrated on Munich

Further detailed analysis can be found in the Appendix.

Business Plan Model

Rent and sale prices are handy for investments in single real estate objects. Large scale investors with thousands of real estate objects (portfolios) are interested in this portfolio performance as well, i.e. monitoring risk and return for the combination of all objects. For this, CBs business plan tool is integrated in our pipeline. In- and outputs are:

Input Real estate portfolio For every city we define an artificial portfolio of 19 real estate objects, hence a collection of different apartments. Then, the net purchase price is computed by multiplying the number of square meters with the respective price per square meter. Factors like notary fee, real estate agency fee and transfer tax then return the gross purchase price of the portfolio.

Input rent and sale price indices For every city the 100 generated rent and sale price scenarios, as returned by the SDE model, are normalized to the first month in 2020. This means that both variables take the value 100.0 for the first month. They are then used to compute future cash flows and sale prices for each real estate objects.

Output IRR Money looses value over time, €100 today are worth less in 20 years from now. The rate of decay is denoted by inflation. Given yearly cash flows C_n and the inflation rate (the European Central Bank (ECB) targets 2%) one can compute the net present value (NPV), hence the value of future cash flows as of now. Now, if we don't use the inflation rate but an imaginary interest rate $r*$, we can choose $r*$ such that for given C_n the NPV equals zero. This internal rate of return (IRR) is our chosen KPI for return, the higher $r*$, the better the investment (at least relative, not absolute). Intuitively, it is the rate of which our initial investment is interest-paying on average [Reniers, Talarico, and Paltrinieri, 2016].

$$NPV(r*) = \Sigma_{n=0}^{N} \frac{C_n}{(1+r*)^n} \stackrel{!}{=} 0 \tag{33}$$

Output IRR Levered Levered and unlevered denote investments with and without debt. Investing with debt leverages the IRR computed on an equity-baseline both ways. That means that the IRR on invested equity is larger in magnitude when levered. Thus, investors can benefit from potentially bigger upsides but also the negative return (loss) becomes larger for the case of bad investments.
Simply put, $IRR_{levered} \propto Lever \times IRR_{unlevered}$

To sum it up, CBs business plan takes the portfolio information with the predicted rent and sale scenarios and computes an IRR distribution (We have 100 scenarios so the business plan computes a distribution over 100 samples). Vacancy rates, constant rents over a predefined time range (normally, one can't rise rent every year) and other factors influence this return as well. Finally, for simplicity we use the mean of the distribution as final return KPI.

Results

Results Risk Model

The results of the risk model heavily depend on which characteristic, Sharpe Ratio or variance of the profit growth, and which level of u is chosen as the dependent variable for the location factor model and on the value for weight w. Figure 58 shows the outcome of the Logit Model for u=5%. The probability of default all over Germany moves between 20 and 40% and starts from a clear maximum in 2011.

Interestingly, the risk calculated in the location factor model for the five largest cities is quite diverse. Munich and Cologne seem to have a significantly higher risk than the German average and the other large cities.

We then combine the two models with a weight of 0.25 for w.
You can find the aggregated outcome of the risk model with a weight w of 0.25, 5% for u and the Sharpe Ratio as the dependent variable y in the Location Factor Model in Figure 11. Analogously to the Probability of Default of the Logit Model and Risk Score from the Location Factor Model, 1 denotes the highest possible risk, 0 the lowest possible risk. This is the result when setting u in the Logit Model and y in the Location Factor Model solely based on their statistical performance in the logistic and linear regression.

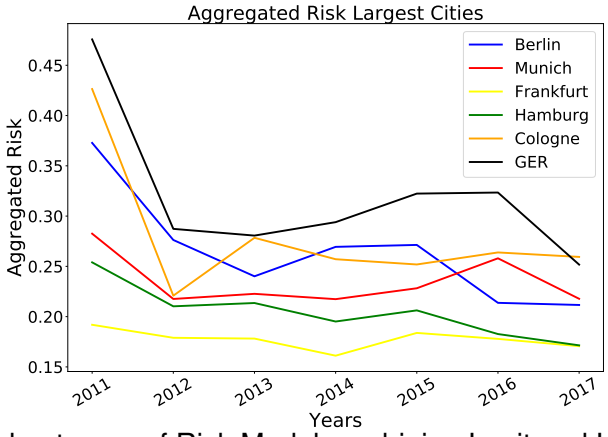


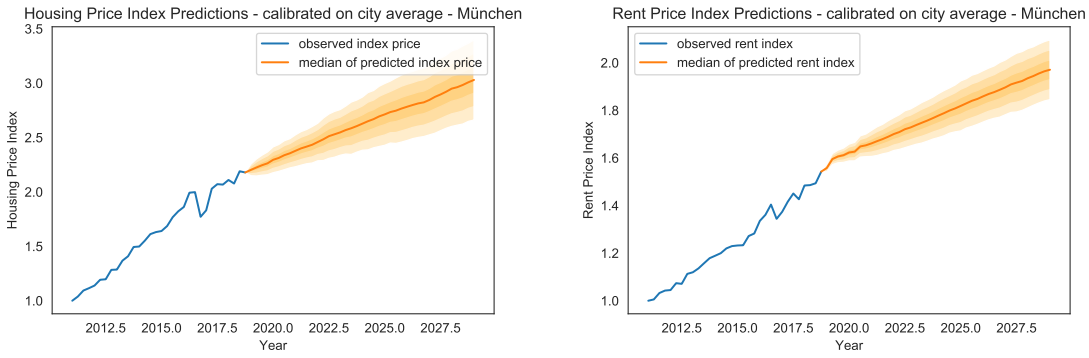
Figure 11. General outcome of Risk Model combining Logit and Location Factor Model

In order to better compare our findings to findings from relevant literature, we also clustered the cities into A,B and C-Cities like in Bulwiengesa, 2019. Similar to Bulwiengesa, 2019 we find that A-cities pose the lowest aggregated risk. Interestingly A-Cities end up being way riskier than B- and C-Cities in the Location Factor Model on average. Find plots on the results, an overview and short explanation on A-,B- and C-Cities in the Appendix .

The results of the Risk Model obviously change significantly when we replace the constant level of 5% for u with a city or time specific level. Please find those additional results in the appendix .

Results SDE Model

The main goal of the SDE model is to predict the trend followed by the Housing Price Index and the Rent price index, taking into consideration macroeconomic factors that can heavily affect the value of the residential market. As an example, we consider the city of Munich keeping in mind that other cities follow a similar behavior. Figure 12a and Figure 12b show the interesting results of our predictions. Both, the Housing Price Index and the Rent Price Index, follow an upward trend with an important increase in the next 8 years. For the Housing Price Index we observe a medium level of standard deviation that increases over time. Comparing these results with the historical city median leads to the assumption of reasonable predictions for future years.
The prediction of the Rent Price Index displays the expected upward trend, but the level of volatility seems to be relatively lower compared to the volatility of the HPI. This observation is in line with the common economic view that historically rents are less volatile than house prices [Gallin, 2008].



(a) Prediction of the Housing Price Index (b) Predictions of Rent Price Index

Figure 12. Prediction from 2018 to 2028 displaying simulations by regions as levels of standard deviation (orange shades) from the median - city of Munich

Therefore, from an historical perspective looking at the time series available, our predictions seem to be accurate and a suitable input for the business plan. Nevertheless, one needs to see the limits of our predictions: It was not possible to include the idea of economic cycles into our model, as well as economic views into the future or incorporate the current situation risen by the outbreak of the corona virus (COVID-19). UBS Global Real Estate [Holzhey, Skoczek, and Hofer, 2019, p. 4] states that “price growth rates have continued to slow in a majority of cities [in Europe]. Average price growth has come to a standstill for the first time since 2012.” Regarding bigger cities, UBS mentions an increase in risk of entering “price-bubble territories” and expects an end to the housing market boom despite the current low mortgage rates [Holzhey et al., 2019, p. 18].

Business Plan Results

As the business plan only transforms the house and rent price indices to IRR distributions we briefly show two sample outputs of it in Figure 59. For the final risk vs. return scatter plot, the mean of each distribution is taken.

Final Results

Our ultimate goal, returning risk versus return is shown in Figure 13. Most cities have similar expected IRRs and expected risk scores. Overall we get positive IRRs for all cities. The median unlevered IRR is $\approx$ 0.5%. Other than expected increasing risk is not associated with higher returns.The R^2 value of 0.006 and a p-value of $\approx$ 0.8 for a linear regression reveal no likely linear dependency of both variables. Return-wise, with highly levered investments the spread of best and worst performers would go up.

Shifting the perspective to geographic point of views (see Figure 14) we see a more risky region around Ruhr area with higher risk variance among these cities as well (Green, yelow and red cities are in near vicinity of each other). Again, linear regressions do not confirm any linear dependencies. Neither risk nor return are coupled to longitude or latitude (R^2 are near zero).

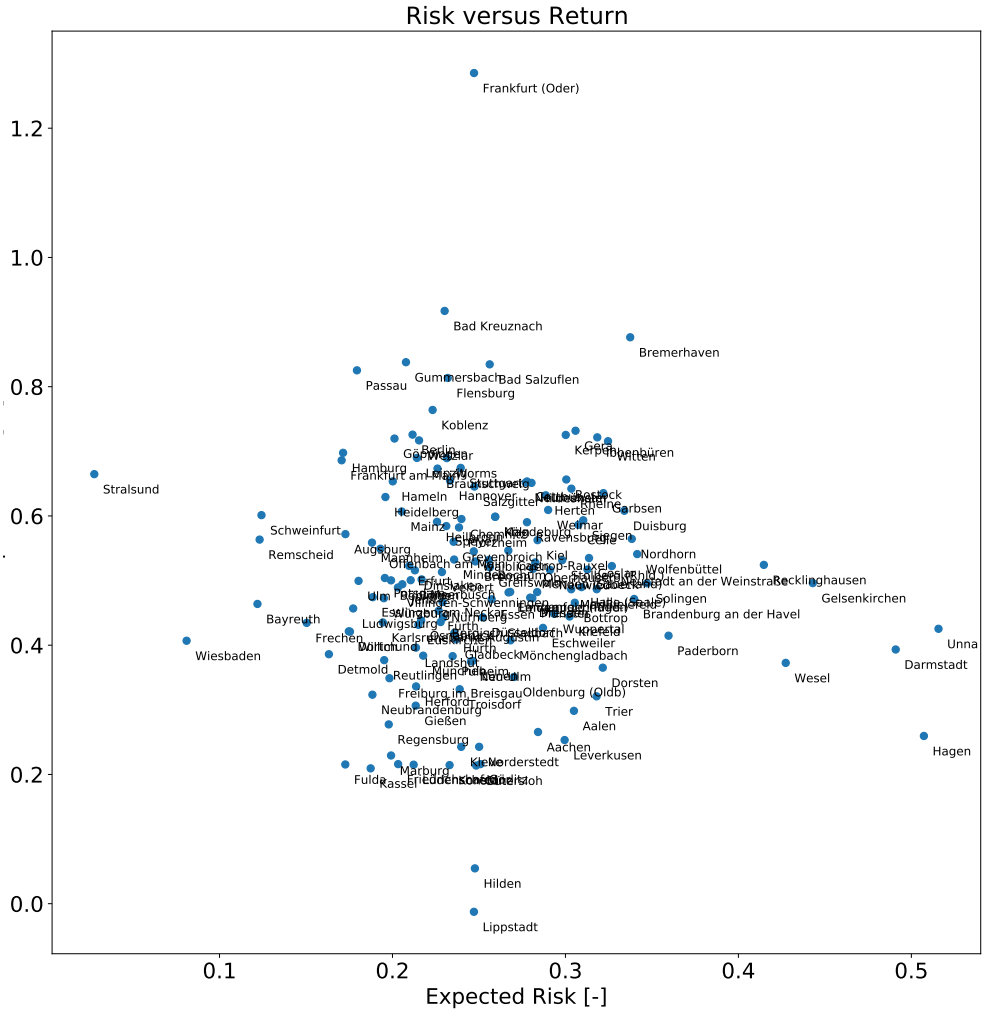


Figure 13. Risk versus unlevered IRR. Most cities have similar risks and returns when looking at the next 10 predicted years. Especially, risk does not increase with higher expected return as one would expect.

Outlook

With one single fixed risk KPI and IRR as return KPI things are clear. But different investors might have different risk definitions or risk appetite. One future enhancement could be the implementation of various risk and return KPIs.

Another enhancement would be incorporating more data in both models and a more sophisticated data and feature extraction pipeline. As with all data driven method, the quality of results is driven by the quality of data.

Third, a analysis not only on city level but on tile or district level could be feasible. That approach would also need more data but would allow for single real estate objects to be evaluated more precise.

Regarding the SDE model, it has to be mentioned that there are various parameters in the beginning of the SDE pipeline that can be set to different values or ranges: For instance, it is possible to vary the lag at which the rent prices follow the housing prices, to calibrate the rent market's sigma on tile or median level of the data or to change the jump size distribution from a lognormal distribution to a normal distribution. Considering further improvements, one could expand the model by a non-constant volatility as seen in the theoretical section given that a longer time-series is available. In that way, the possibility arises to estimate the volatility over time for each city, and therefore could increase the prediction power of the model.

It would be also possible to incorporate other additional factors to the system describing the rent price, in order to reflect a more realistic dynamic. Furthermore having full access to the location information on tile level would give the opportunity to predict the housing and rent prices on tile level to receive an even more accurate local prediction.

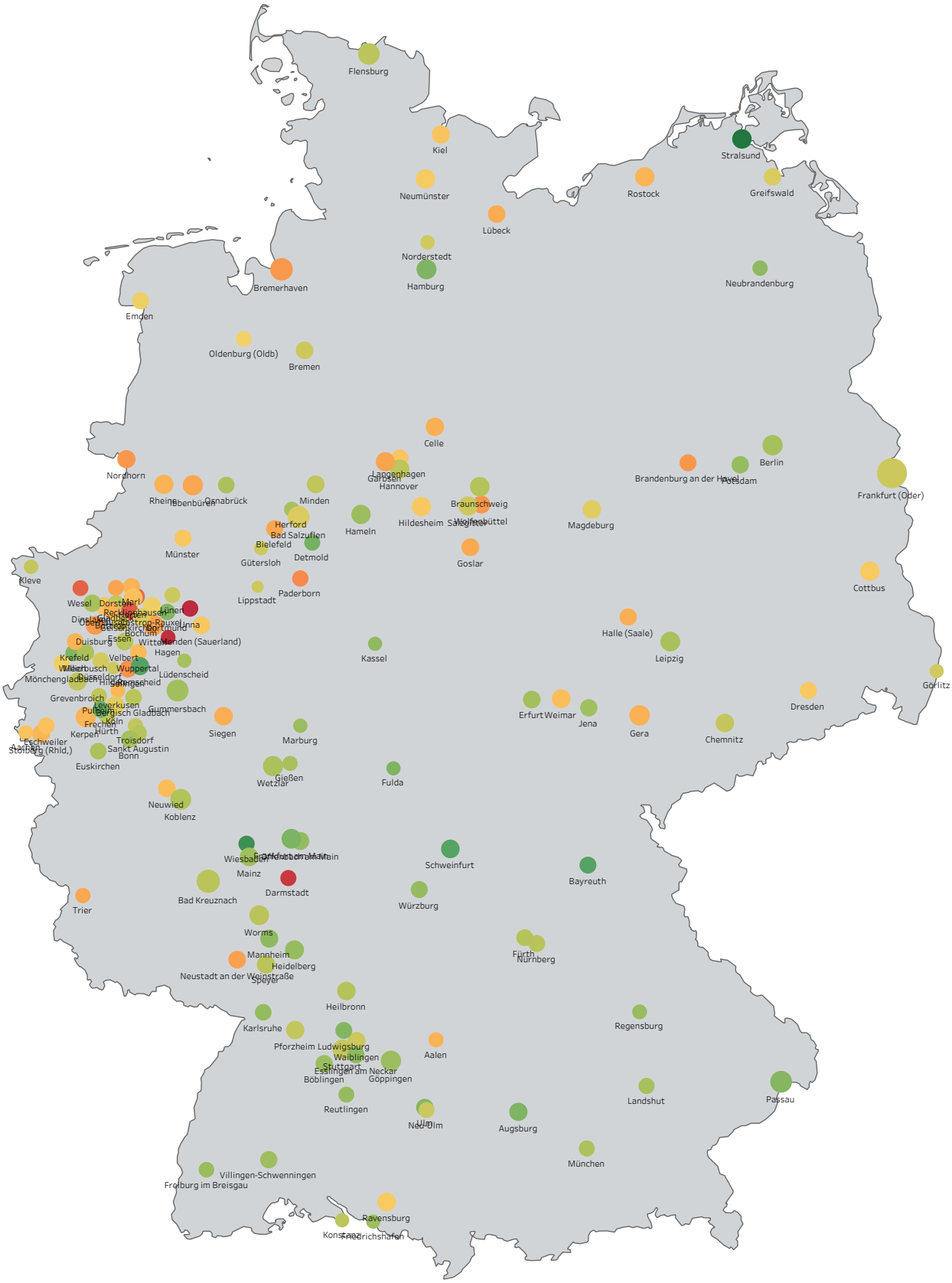


Figure 14. Risk versus unlevered IRR, bigger circles correspond to higher expected IRR, green denotes low risk, red high risk. The only thing that pops out is a higher risk variance in the Ruhr area.

References

Altschäffel, L., Dragoi, O., Mendez, L., Tamada, T., & Wiyoga, G. (2020). Revolution of real estate valuation.

Asaftei, G., Doshi, S., Means, J., & Sanghvi, A. (2018). Getting ahead of the market: How big data is transforming real estate. Retrieved from https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate#

BBR. (2020a). Inkar. Retrieved from https://www.inkar.de/

BBR. (2020b). Inkar. Retrieved from https://www.inkar.de/documents/Indikatoren%20Raum-%20und%20Zeitbezuege.pdf

Belke, A., & Keil, J. (2018). Fundamental determinants of real estate prices: A panel study of german regions. *International Advances in Economic Research*, *24*(1), 25–45.

Bulwiengesa, A. (2019). Rendite-risiko-verhältnis für wohninvestments, 33.

Capital Bay. (2020). Capital bay about us. Retrieved from https://capitalbay.de/ueber-uns

Gallin, J. (2008). The long-run relationship between house prices and rents. *Real Estate Economics*, *36*(4), 635–658.

Higham, D. J., & Kloeden, P. E. (2006). Convergence and stability of implicit methods for jump-diffusion systems. *Int. J. Numer. Anal. Model*, *3*(2), 125–140.

Holzhey, M., Skoczek, M., & Hofer, K. (2019). Ubs global real estate bubble index 2019. *Chief Investment Office GWM Investment Research*.

Kaplan, S., & Garrick, B. J. (1981). On the quantitative definition of risk. *Risk analysis*, *1*(1), 11–27.

Manganelli, B. (2014). *Real estate investing: Market analysis, valuation techniques, and risk management*. Springer.

McDonald, J. F. (2000). Rent, vacancy and equilibrium in real estate markets. *JOURNAL OF REAL ESTATE PRACTICE AND EDUCATION*, *3*(1), 55–69.

Reniers, G., Talarico, L., & Paltrinieri, N. (2016). Chapter 16 - cost-benefit analysis of safety measures. In N. Paltrinieri & F. Khan (Eds.), *Dynamic risk analysis in the chemical and petroleum industry* (pp. 195–205). doi:https://doi.org/10.1016/B978-0-12-803765-2.00016-0

Sharpe, W. F. (1994). The sharpe ratio. *Journal of portfolio management*, *21*(1), 49–58.

Yilmaz, B., & Selcuk-Kestel, A. S. (2018). A stochastic approach to model housing markets: The us housing market case. *Numerical Algebra, Control & Optimization*, *8*(4), 481.

Appendices

Data description - SDE model

Data source: The data of housing and rent prices on tile level is provided by the 21st Dataset, which is further discussed in the Section Data. The effective interest rate of German banks concerning new business on housing loans provided by Deutsche Bundesbank on a quarterly basis is used as input for the interest rate calculations of the SDE pipeline.

Data display: The plots below display the housing and rent prices for specific cities over time and show a general trend to increasing prices for small and big cities throughout Germany.

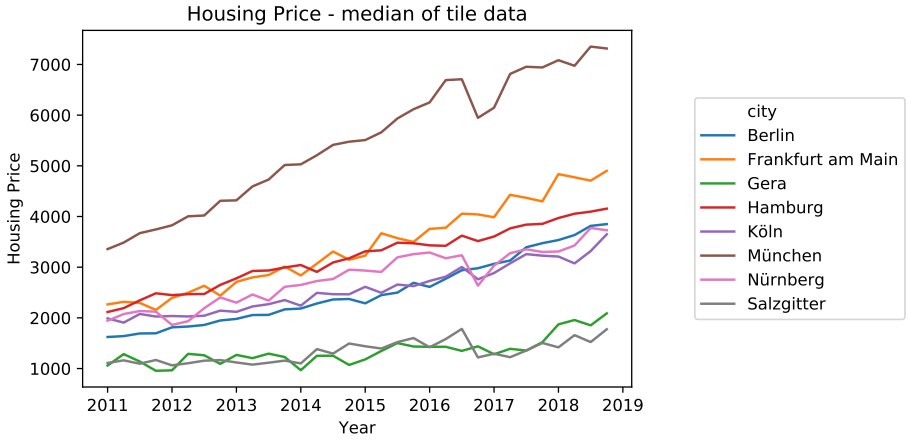


Figure 15. Housing Prices for specific cities

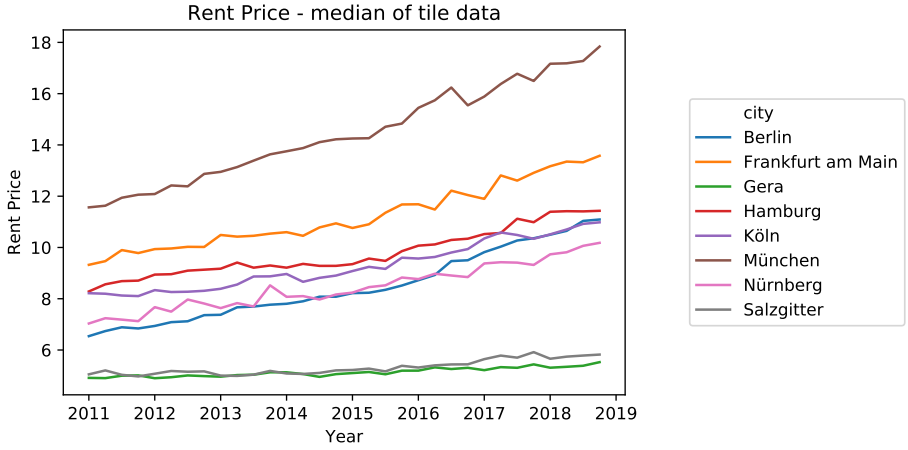


Figure 16. Rent Prices for specific cities

We also want to show a comparison of the tile data and the median for Housing Price and Rent Indices for Munich.

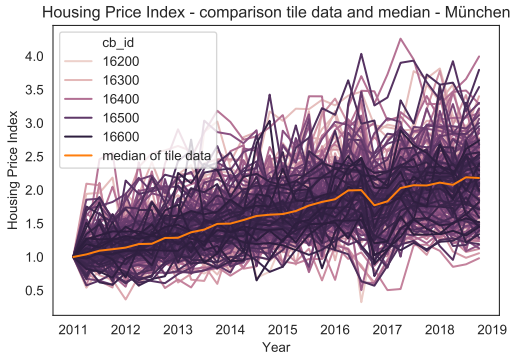


Figure 17. Observed Housing Price Index on tile level and median level - Munich

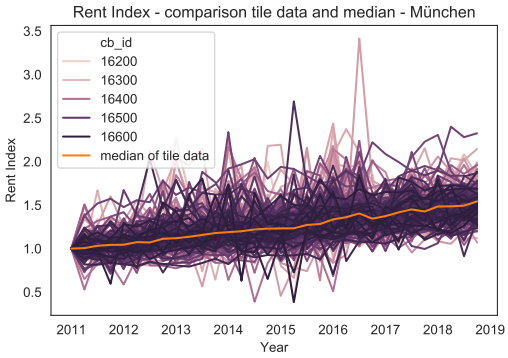


Figure 18. Observed Rent Price Index on tile level and median level - Munich

Vacancy - Additional data analysis In the following, the plots mentioned in Section Model Implementation regarding further analysis of the vacancy data are displayed.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        rent_cell_INDEX   R-squared:                       0.620
Model:                            OLS   Adj. R-squared:                  0.619
Method:                 Least Squares   F-statistic:                     2257.
Date:                Sun, 19 Jul 2020   Prob (F-statistic):               0.00
Time:                        19:51:10   Log-Likelihood:                2401.6
No. Observations:                2772   AIC:                            -4797.
Df Residuals:                    2769   BIC:                            -4779.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.0906      0.006     14.305      0.000       0.078       0.103
sale_cell_INDEX    0.7051      0.012     56.544      0.000       0.681       0.730
Vacancy           -0.0981      0.012     -8.244      0.000      -0.121      -0.075
==============================================================================
Omnibus:                      253.787   Durbin-Watson:                   0.748
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              423.574
Skew:                           0.657   Prob(JB):                     1.05e-92
Kurtosis:                       4.393   Cond. No.                         8.86
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 19. Multi-linear regression analysis for rent prices using housing prices and vacancy rate as independent factors

Figure 20. Relative growth of vacancy over all cities and timesteps

Parameter estimation

Lambda of jump process - Plots: The plots mentioned in the main section are displayed in the following. The first plot shows a histogram displaying the mean jumps over 8 years for each city assuming a calibrated lambda for each city.

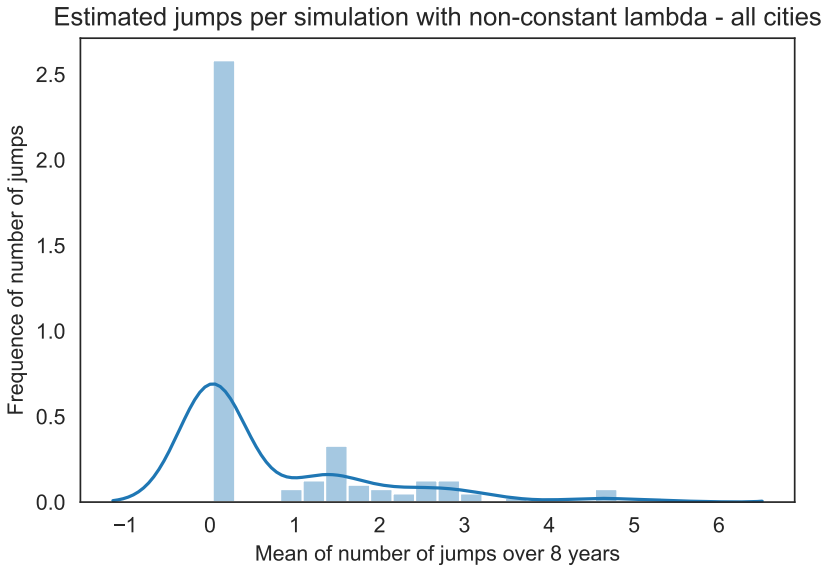


Figure 21. Histogram displaying the mean jumps over 8 years assuming a calibrated lambda of each city

The second plot 22 counts the jumps for each simulation (in total 200 simulations) over 8 years by assuming the same lambda for all cities. Incorporating an economic view, we decided to estimate the λ of the Poisson process as the constant probability that one jump occurs in a cycle of eight years independent of cities, considering

quarterly time steps.

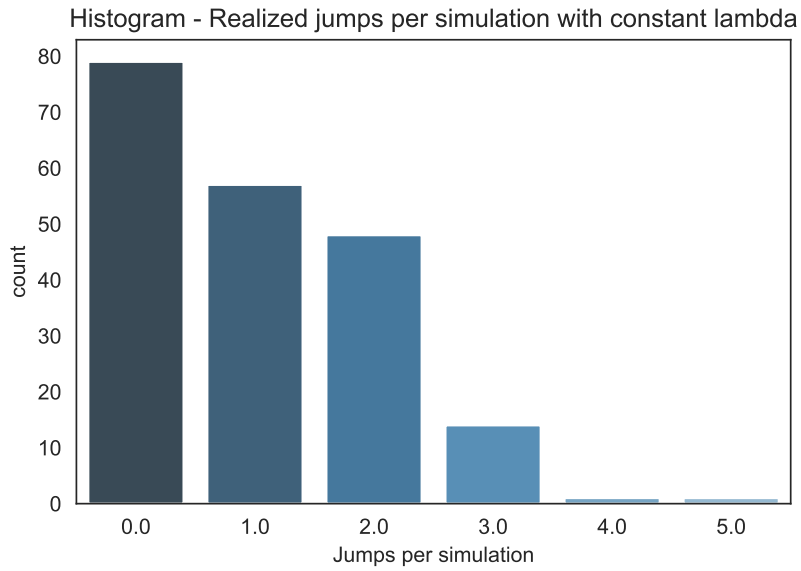


Figure 22. Number of realized jumps per simulation over 8 years assuming a constant lambda for all cities

Size of jumps: The main goal is to fit a distribution to the actual jumps as closely as possible. Due to the short time series, data on tile level is considered and analyzed in the following. To discard the jumps from the respective time series, a jump is defined when the relative growth of one timesteps exceeds the Threshold set at $3 * std(relative\ growth\ of\ data)$. An analysis of the jumps shows that there happen to be exclusively positive jumps in 77% of the cities.

By looking into tile data and displaying the specific histogram of jumps for Munich in Figure 23, one observes mostly positive jumps and a few negative jumps. Considering either the positive or negative jumps, they can be fitted to a shifted lognormal distribution. Incorporating both sides, one adds a bernoulli distribution with probability p that a positive jump happens. Considering that the jumps are fitted to the tile data, the jumps need to be rescaled to the median of relative growth data. Figure 24 shows the histogram of relative growth data of the median including the adjusted fitted sample jumps.

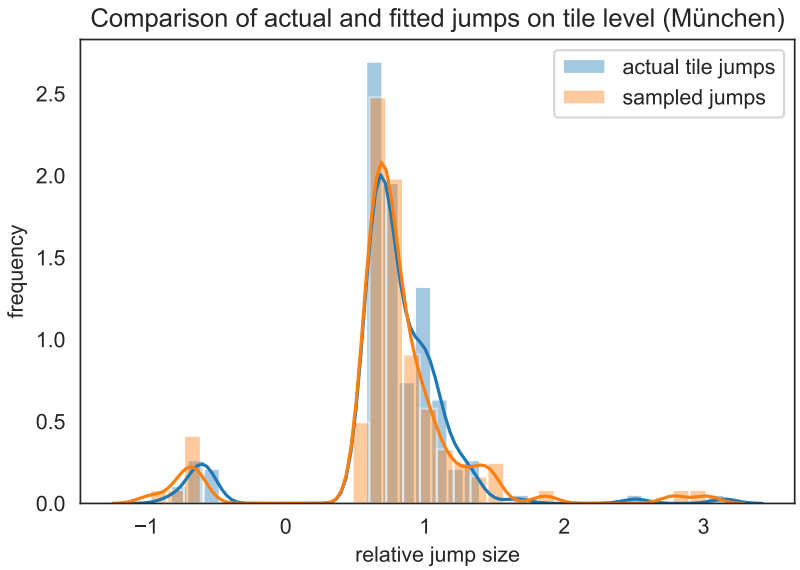


Figure 23. Jumps for Munich on tile level

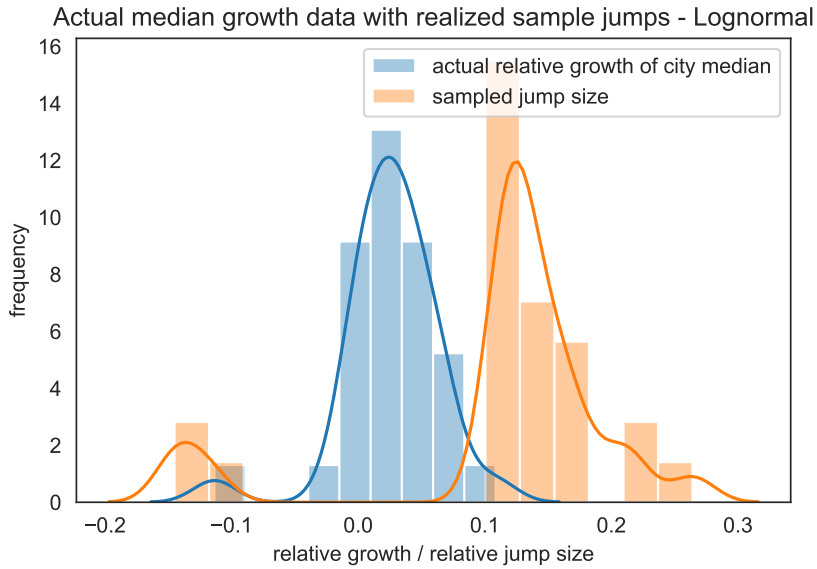


Figure 24. Adjusted jumps for Munich compared to the level of median growth data

Further SDE Model Analysis

Violin plots: We can plot a distribution of predictions for each city. In detail, this means that we can compare the distribution of predicted housing and rent prices with the observed distribution over tile data at different time-steps for each city, and see whether our model generates a similar set of predictions. By looking into the plots describing Munich, one finds a similar distribution for Housing Price Indexes of real and

estimated data, as well as for Rent Price Indexes . One can observe the increasing volatility by the estimated data when going further in time. One observation that needs to be mentioned is that the SDE model is not able to catch the peaks of the observed Rent Price Index in 2016 and 2018 in Munich leading to lower estimated values at these time steps.

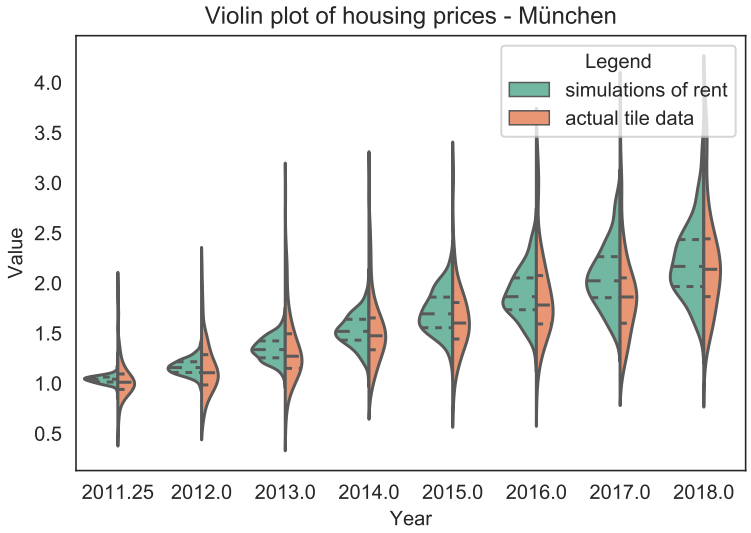


Figure 25. Violin distribution plot for housing price on tile level from 2011 to 2018 for the city of Munich

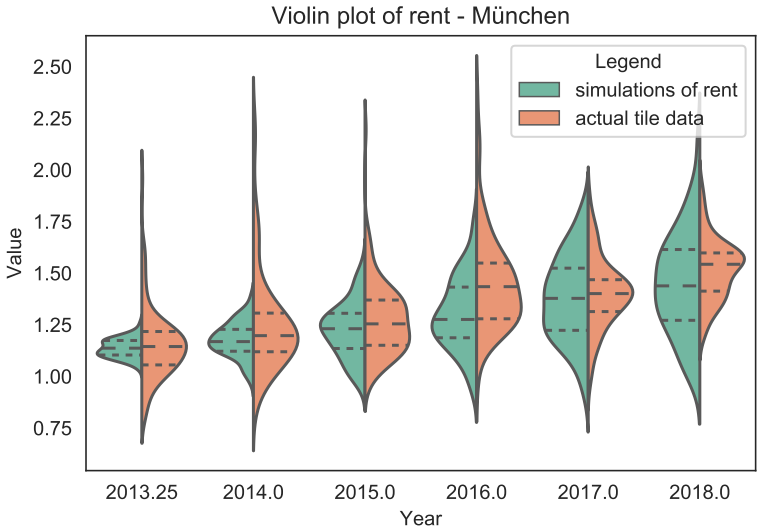


Figure 26. Violin distribution plot for rent price on tile level from 2013 to 2018 for the city of Munich

Lag analysis: Looking at the lag analysis between Rent and Sales, we observe that there is a tendency that the Rent Price Index follows the behavior of the House Price Index rather than the opposite for most of the cities. One can determine by Figure 27 that the best fitting lags occur at $lag = 0$ and $lag = 8\, quaters$. This can be interpreted for most of the cities that rent and housing prices do not tend to lead each other or rent prices follow housing prices by 2 years.

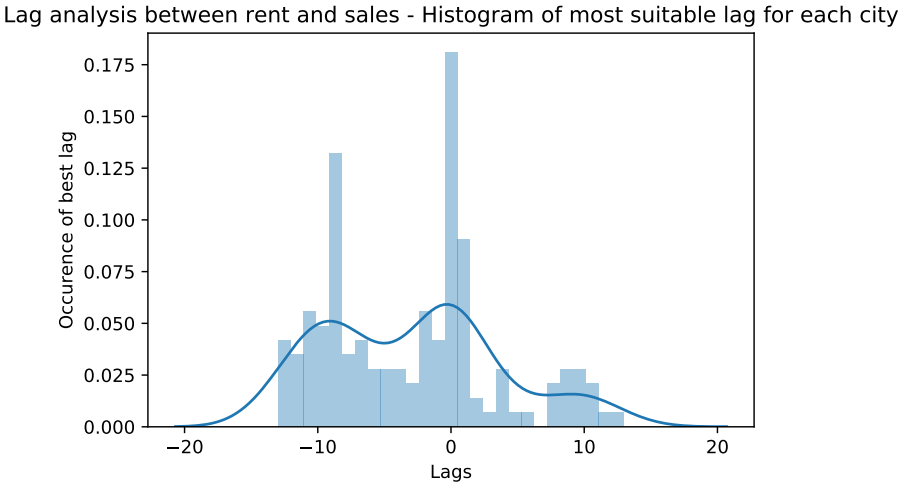


Figure 27. Lag analysis between the Rent price and Housing price for each city. For cities on the left side of the graph, rent prices follow the house prices trend, and on the right side of the histogram house prices follow the rent price trend.

Predicted Simulations

In order to entirely display our results, we want to include the plots showing 100 predicted simulations for all of the factors. The plots show predictions 10 years into the future calibrated for the city of Munich.

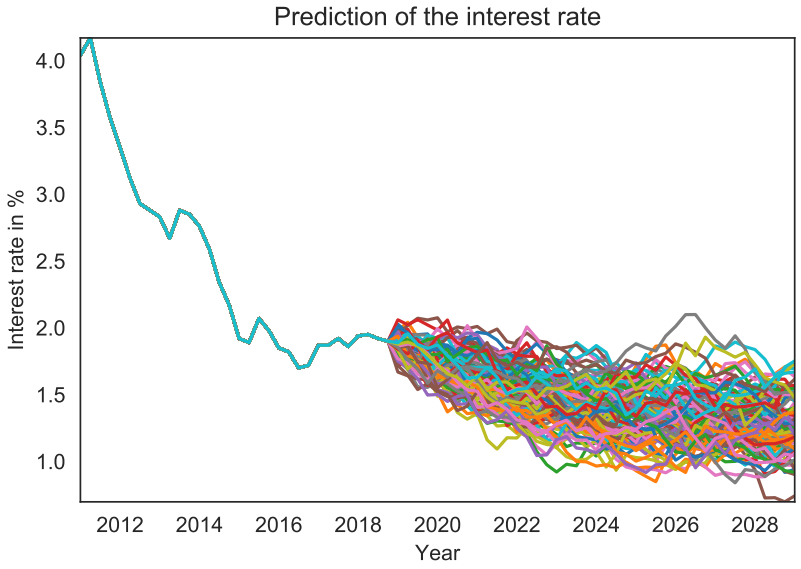


Figure 28. Interest rate - 100 predictions for the next 10 years - calibrated for Munich

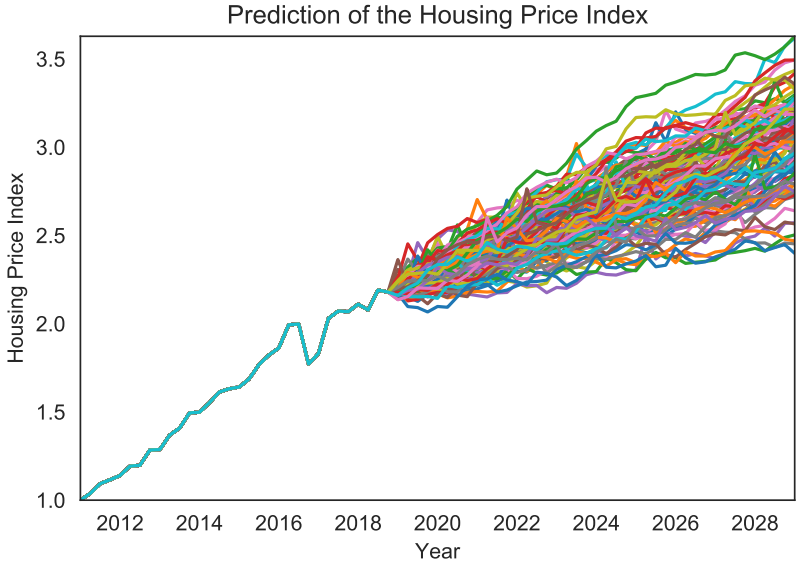


Figure 29. Housing Price Index - 100 predictions for the next 10 years - calibrated for Munich

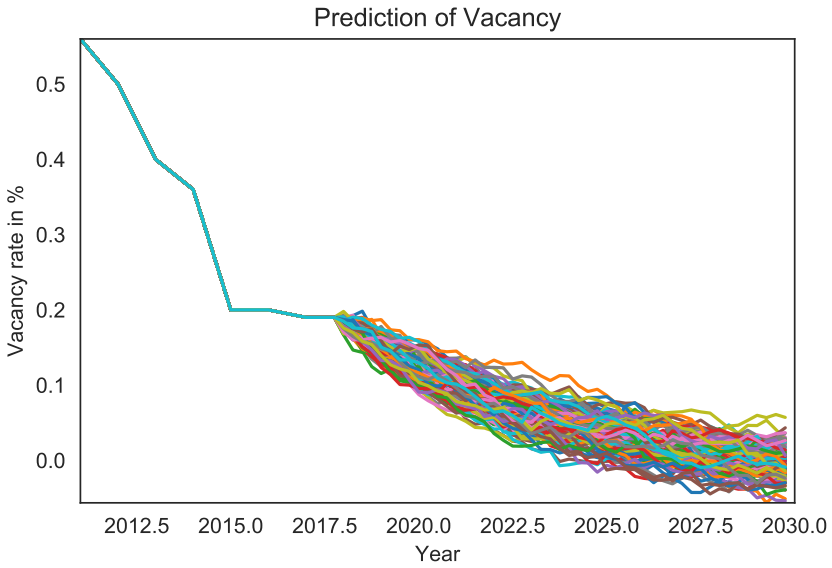


Figure 30. Vacancy rate - 100 predictions for the next 10 years - calibrated for Munich

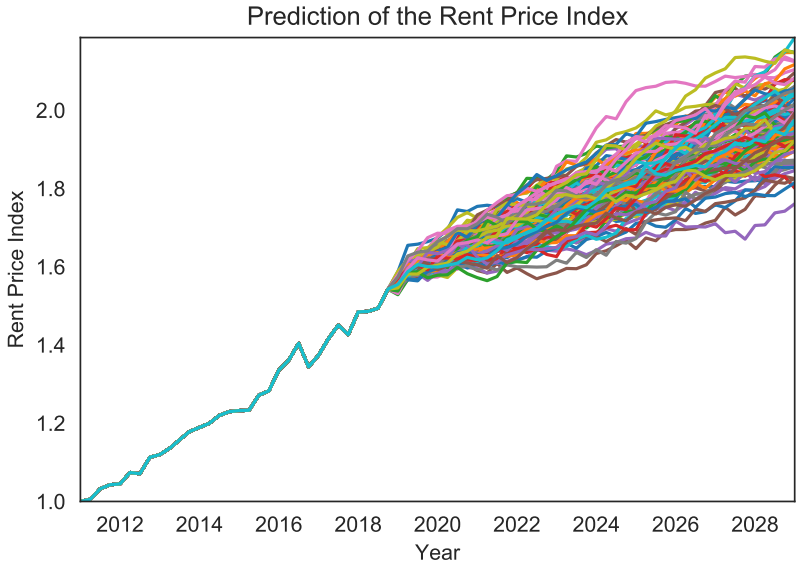


Figure 31. Rent Price Index - 100 predictions for the next 10 years - calibrated for Munich

Research

The main inspiration for the Risk Model comes from [Bulwiengesa, 2019], which categorizes 43 German cities according to different risk scores and tries to answer questions like *which variables are stabilizers?* and *which are risk factors?*, *where do these factors apply?*, *what is the impact of them and can we measure stability or risk?*
The paper *Fundamental Determinants of Real Estate Prices: A Panel Study of German Regions* [Belke and Keil, 2018] summarizes variables that could drive real estate prices and explains different estimation methods. We systematically went through the related literature and listed the variables that have been analysed in the past. These variables ranged from interest rates and mortgage rates to macroeconomic factors like unemployment rate, GDP and public policy regulations. This review helped us understand how we could start classifying variables, for example in supply- and demand-side factors. With this background knowledge we continued to gather more data.
We looked at all variables in the INKAR database and included all data points in our pipeline that could be related to our risk factor. Furthermore we tried to request a longer time series for house and rent prices at the *Bundesinstitut für Bau-, Stadt- und Raumforschung (BSSR)* for a possible extreme values analysis and better regression results. Furthermore we asked the institution *Gutachterausschuss für Grundstückswerte*, which has copies of all real estate purchase contracts, for more datasets. Unfortunately there is no aggregated database from the *Gutachterausschüsse* across Germany. Rather every city or county has its own database and it was not feasible to obtain this data.
However we were able to get the market rent prices (Angebotsmieten) from the *BBSR-Wohnungsmarktbeobachtung* and *IDN ImmoDaten GmbH* for 109 German cities from 2004-2019, where 77 cities match the 149 cities we had available. The datapoints are based on newspaper and internet ads for flats and describe the net cold rent prices. Since the data from 21st is aggregated the same way and therefore a direct comparison is possible. We calculated the mean relative error which is 4,16% and the maximum relative error is 16,18% for the city of Rostock in the year 2016. Figure 32 (appendix) shows a histogram where the mean of the relative error over the whole available timeframe is displayed in the x-axis and the number of corresponding cities is displayed in the y-axis.

Obviously this comparison cannot be considered a validation of the 21st dataset since we can only compare half of the cities. But for these cities the comparison shows that the data is fairly similar to the data collected by public institutions.

Data Pipeline

Location Factor Model

Find our approach to identify the best set of features X in the Location Factor Model below.

Drop Missing Values: Some of the INKAR data was not available for a few cities. We dropped every column completely that had at least one missing value to be able to end up with a Location Factor risk score for all cities.

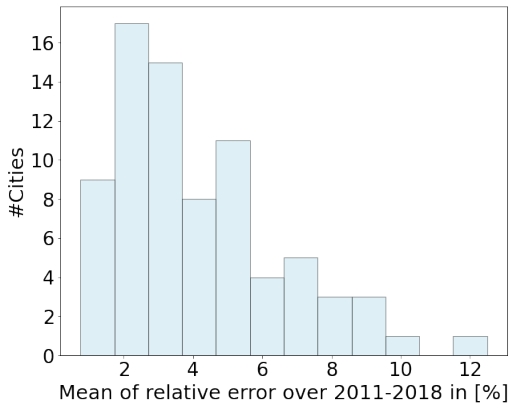


Figure 32. Relative error of rent prices between 21st and BSSR

Year	INKAR Variable X1	INKAR Variable X2
2011	NaN -> 123	111
2012	123	...
...	...	...
2016	...	321
2017	...	NaN -> 321

Backward Fill

Forward Fill

Figure 33. Backward- and forward-fill for NaNs

Include Interactions: We then include interactions and add a variable for the product of every 2-tupel combination of variables. We call those new features according to following schema: $mult\_ < column1 > \_x\_ < column2 >$.

Remove correlated features: In a first step we remove one of two features with a Pearson Correlation Index higher than 0.8.

Scale Features: In order to increase comparability of the different features we then scale all features to be between 0 and 1 using a MinMaxScaler.

Recursive Feature Elimination: In a last step we run a Recursive Feature Elimination Algorithm to identify the 15 factors that are most significant.

General Results

Find the regression results for the approach described in the main part using a level of 5% for u in the Logit Model and for the Sharpe Ratio and variance of the profit growth in the Location Factor Model.s

Results: Logit

Model:	Logit	Pseudo R-squared:	0.210
Dependent Variable:	y	AIC:	936.4231
Date:	2020-07-17 14:36	BIC:	1040.3701
No. Observations:	1043	Log-Likelihood:	-447.21
Df Model:	20	LL-Null:	-566.20
Df Residuals:	1022	LLR p-value:	3.0117e-39
Converged:	1.0000	Scale:	1.0000
No. Iterations:	8.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Einkommensteuer	-2.0572	0.9829	-2.0930	0.0363	-3.9837	-0.1308
Population	-4.9122	2.5345	-1.9381	0.0526	-9.8797	0.0553
growth_Anteil Schutzsuchender an Bevölkerung	2.1385	1.5746	1.3581	0.1744	-0.9477	5.2248
growth_Population	-6.2201	2.2120	-2.8120	0.0049	-10.5556	-1.8847
mult_Income_x_Pendlersaldo	-2.8226	1.0901	-2.5892	0.0096	-4.9592	-0.6859
mult_Studierende_x_growth_app_licence_residential	4.2318	1.2072	3.5055	0.0005	1.8658	6.5979
mult_Einkommensteuer_x_Growth_Shrink_Ratio	-2.0190	1.6854	-1.1979	0.2310	-5.3223	1.2844
mult_Verhältnis junge zu alten Erwerbsfähigen_x_Beschäftigte Tertiärer Sektor	-1.4476	0.7644	-1.8939	0.0582	-2.9458	0.0505
mult_growth_Einkommensteuer_x_growth_Population	-0.6167	1.6642	-0.3706	0.7110	-3.8784	2.6451
mult_growth_Anteil Schutzsuchender an Bevölkerung_x_growth_Bruttowertschöpfung	6.2480	2.8407	2.1995	0.0278	0.6804	11.8157
mult_growth_Empfänger von Grundsicherung im Alter (Altersarmut)_x_growth_Growth_Shrink_Ratio	-3.7728	1.4599	-2.5844	0.0098	-6.6341	-0.9115
growth_Schuldnerquote_lagged_1	1.7541	0.7536	2.3276	0.0199	0.2770	3.2312
growth_Existenzgründungen_x_lagged_2	2.3993	1.0993	2.1826	0.0291	0.2447	4.5539
growth_Anteil Teilzeitbeschäftigte_lagged_2	-2.1682	0.8031	-2.6999	0.0069	-3.7421	-0.5942
growth_Beschäftigte am Wohnort mit akademischem Abschluss_lagged_2	1.9486	0.8279	2.3536	0.0186	0.3259	3.5713
growth_SGB II - Quote_lagged_1	1.6386	0.9479	1.7286	0.0839	-0.2193	3.4965
growth_Personen in Bedarfsgemeinschaften_lagged_2	2.7145	0.8769	3.0954	0.0020	0.9957	4.4332
growth_birth_death_lagged_2	-1.7249	0.7221	-2.3888	0.0169	-3.1401	-0.3097
Growth_Shrink_Ratio_lagged_1	2.3353	1.5280	1.5283	0.1264	-0.6596	5.3301
growth_Growth_Shrink_Ratio_lagged_1	2.9805	1.4196	2.0996	0.0358	0.1982	5.7627
const	0.0958	1.2975	0.0739	0.9411	-2.4472	2.6389

Figure 34. Regression results for u=5%

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.477
Dependent Variable:	SharpeRatio_Absolut	AIC:	-40.8291
Date:	2020-07-20 11:21	BIC:	7.2341
No. Observations:	149	Log-Likelihood:	36.415
Df Model:	15	F-statistic:	10.00
Df Residuals:	133	Prob (F-statistic):	1.46e-15
R-squared:	0.530	Scale:	0.040233

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
Leistungen für Wohngeld_mean	-0.1817	0.5132	-0.3541	0.7238	-1.1967	0.8333
Anteil Erholungsfläche_mean	0.6042	0.1676	3.6052	0.0004	0.2727	0.9357
Ärzte je Einwohner_mean	0.0144	0.7646	0.0189	0.9850	-1.4980	1.5268
Ein- und Zweifamilienhäuser_mean	-0.6722	0.7080	-0.9494	0.3441	-2.0727	0.7282
Großunternehmen_mean	-0.3073	0.1263	-2.4336	0.0163	-0.5571	-0.0575
mult_Erreichbarkeit von Flughäfen_x_Anteil Erholungsfläche_mean	-0.5965	0.1693	-3.5236	0.0006	-0.9314	-0.2617
mult_Leistungen für Wohngeld_mean_x_Ein- und Zweifamilienhäuser_mean	0.6850	0.6656	1.0292	0.3053	-0.6315	2.0016
mult_Nahversorgung Grundschulen Durchschnittsdistanz_x_Erreichbarkeit von Oberzentren	0.0301	0.1100	0.2739	0.7846	-0.1874	0.2476
mult_Nahversorgung Supermärkte Durchschnittsdistanz_x_Erreichbarkeit von Autobahnen	0.2633	0.1658	1.5875	0.1148	-0.0648	0.5913
mult_Anteil Erholungsfläche_mean_x_Erreichbarkeit von Autobahnen	0.4851	0.1780	2.7252	0.0073	0.1330	0.8372
mult_Anteil Erholungsfläche_mean_x_Erreichbarkeit von IC/EC/ICE-Bahnhöfen	-0.5054	0.2481	-2.0365	0.0437	-0.9962	-0.0145
mult_Anteil Erholungsfläche_mean_x_Erreichbarkeit von Oberzentren	-0.3560	0.1275	-2.7912	0.0060	-0.6082	-0.1037
mult_Erreichbarkeit von Autobahnen_x_Nahversorgung Apotheken Durchschnittsdistanz	-0.5366	0.2240	-2.3953	0.0180	-0.9797	-0.0935
mult_Erreichbarkeit von IC/EC/ICE-Bahnhöfen_x_Nahversorgung Apotheken Durchschnittsdistanz	0.3353	0.1592	2.1059	0.0371	0.0204	0.6502
mult_Ärzte je Einwohner_mean_x_Ein- und Zweifamilienhäuser_mean	0.2848	0.8441	0.3374	0.7363	-1.3849	1.9545
const	0.7101	0.2951	2.4061	0.0175	0.1264	1.2939

Omnibus:	49.134	Durbin-Watson:	1.731
Prob(Omnibus):	0.000	Jarque-Bera (JB):	266.930
Skew:	1.024	Prob(JB):	0.000
Kurtosis:	9.229	Condition No.:	132

Figure 35. Regression result using the Sharpe Ratio as the dependent variable y.

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.341
Dependent Variable:	growth_profit_var	AIC:	-643.5940
Date:	2020-07-17 14:45	BIC:	-595.5309
No. Observations:	149	Log-Likelihood:	337.80
Df Model:	15	F-statistic:	6.095
Df Residuals:	133	Prob (F-statistic):	1.41e-09
R-squared:	0.407	Scale:	0.00070419

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
Leistungen für Wohngeld_mean	0.1343	0.0679	1.9777	0.0500	-0.0000	0.2686
Anteil Erholungsfläche_mean	-0.0229	0.0222	-1.0330	0.3035	-0.0668	0.0210
Ärzte je Einwohner_mean	0.1713	0.1012	1.6939	0.0926	-0.0287	0.3714
Ein- und Zweifamilienhäuser_mean	0.3190	0.0937	3.4058	0.0009	0.1337	0.5043
Großunternehmen_mean	0.0460	0.0167	2.7547	0.0067	0.0130	0.0791
mult_Erreichbarkeit von Flughäfen_x_Anteil Erholungsfläche_mean	0.0165	0.0224	0.7349	0.4637	-0.0278	0.0608
mult_Leistungen für Wohngeld_mean_x_Ein- und Zweifamilienhäuser_mean	-0.2591	0.0881	-2.9429	0.0038	-0.4333	-0.0850
mult_Nahversorgung Grundschulen Durchschnittsdistanz_x_Erreichbarkeit von Oberzentren	0.0154	0.0145	1.0617	0.2903	-0.0133	0.0442
mult_Nahversorgung Supermärkte Durchschnittsdistanz_x_Erreichbarkeit von Autobahnen	-0.0497	0.0219	-2.2635	0.0252	-0.0931	-0.0063
mult_Anteil Erholungsfläche_mean_x_Erreichbarkeit von Autobahnen	-0.0174	0.0236	-0.7407	0.4602	-0.0640	0.0291
mult_Anteil Erholungsfläche_mean_x_Erreichbarkeit von IC/EC/ICE-Bahnhöfen	0.0695	0.0328	2.1158	0.0362	0.0045	0.1344
mult_Anteil Erholungsfläche_mean_x_Erreichbarkeit von Oberzentren	0.0141	0.0169	0.8341	0.4057	-0.0193	0.0474
mult_Erreichbarkeit von Autobahnen_x_Nahversorgung Apotheken Durchschnittsdistanz	0.0644	0.0296	2.1716	0.0317	0.0057	0.1230
mult_Erreichbarkeit von IC/EC/ICE-Bahnhöfen_x_Nahversorgung Apotheken Durchschnittsdistanz	-0.0525	0.0211	-2.4929	0.0139	-0.0942	-0.0108
mult_Ärzte je Einwohner_mean_x_Ein- und Zweifamilienhäuser_mean	-0.2068	0.1117	-1.8516	0.0663	-0.4277	0.0141
const	-0.0692	0.0390	-1.7723	0.0786	-0.1464	0.0080

Omnibus:	133.511	Durbin-Watson:	2.020
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2617.671
Skew:	3.030	Prob(JB):	0.000
Kurtosis:	22.619	Condition No.:	132

Figure 36. Regression result using the variance of the growth of the profit as the dependent variable y.

Quadrants Backtesting

During data discovery in the beginning of this project, we found that the price to rent ratio varies significantly among the tiles of the same city at one point in time. Based on our argumentation in the chapter we should be able to find empirically that prices of the Growth and Shrink Quadrants are converging. As stated earlier we tested the Growth and Shrink Quadrant idea for the five largest German cities and for the German average. Furthermore we investigated how often the effect of shrinking purchasing prices in tiles form the Shrink Quadrant and growing purchasing prices in the tiles of the Growth Quadrant is actually observable and how long it lasts on average per city. We define the end of the effect whenever the mean purchasing price of the tiles that were part of the Shrink Quadrant in a given quarter reaches the mean purchasing price of the tiles tiles that were part of the Growth Quadrant in a given quarter times 1.05. We set the border to 1.05 times the purchasing price of the Growth Quadrant, because we want to show that Growth and Shrink Quadrant prices are converging to a certain extend. One could also set different levels or measure if and how long it takes until the prices of Growth and Shrink Quadrants will be equal. According to this rule for only *359* cases out of *4768* [1] observations the effect wasn't observable. When we remove 2018, which is the last year of the available timeframe, out of our sample of now *4172* observations only for *85* we cannot show the effect. Figure 37 shows that prices from Growth and Shrink prices are not converging towards the end of the available timeframe. Probably our available timeframe ended before the purchasing prices converged.

[1] $149 cities * 8 years * 4 quarters$

Due to time constraints we just performed an exploratory analysis for a few cities. Those exploratory results seem to support our idea and the features we built based on that idea were significant in many of the different Logit Models. It is important to keep in mind that the Quadrant idea is based on the simple but major assumption that a real estate object in a given tile is only defined by its purchasing and rent price. This is obviously not the case in reality. In reality real estate objects are defined by various features: When was the object built? When was it renovated the last time? Where exactly is it located? In an efficient market all those features should be represented in the price. If one had more detailed data for the different tiles, they could investigate whether there are other reasons that lead to the documented differences in the price to rent ratio in a given city or if those differences are signs of an inefficient real estate market, that poses possibilities for arbitrage profits to a certain extend. With the data we were provided, the different tiles are just defined by purchasing and rent prices. All other interesting tile specific features are unknown to us.

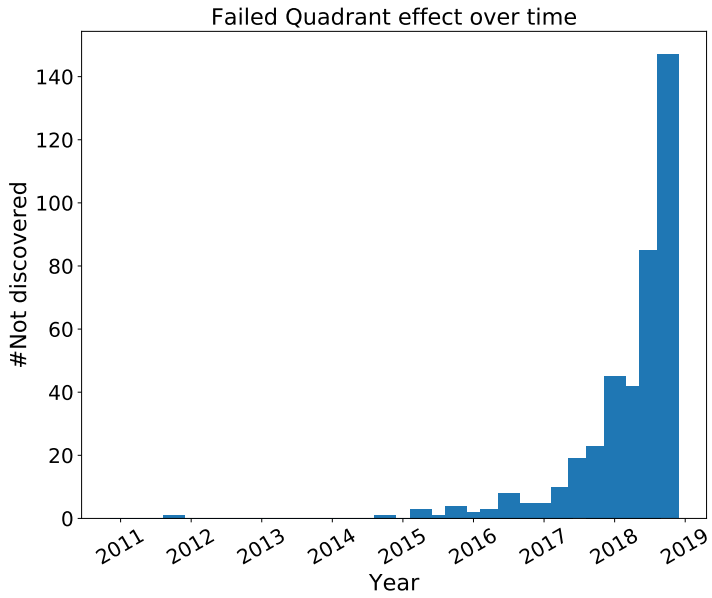


Figure 37. Quarters for which purchasing prices of the Shrink Quadrant didn't reach 1,05 * purchasing prices of the Growth Quadrant.

Figure 38 shows that for the majority of all observations the prices converge during a year. On average there seems to be a clear trend that it takes roughly half a year for prices to converge, nevertheless for some cities it seems to take significantly longer **??**.

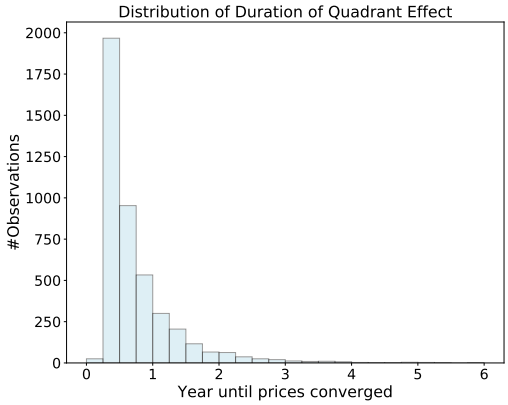


Figure 38. Distribution of how long it takes for prices from Growth and Shrink Quadrants to converge whenever the effect is observable.

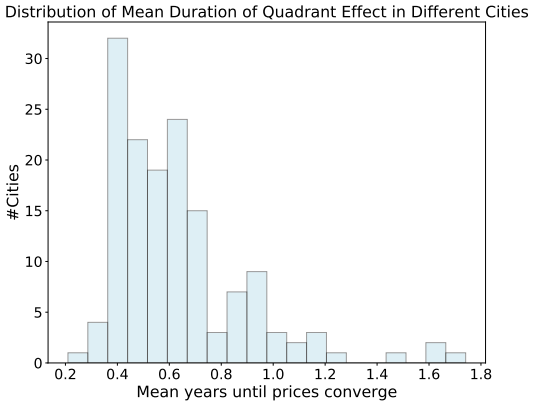


Figure 39. Distribution of how long it takes on average per city for prices from Growth and Shrink Quadrants to converge whenever the effect is observable.

More Regression Results

As described earlier we tested many different levels for u in the Logit Model. Find the most important regression results here.

City specific regression results. We tried city specific levels for u to identify city specific risk drivers. We calculated the average, 0,2% and 0,3% quantile of profit for every city and set u correspondingly. Find the regression results for the 0,2% quantile and the 0,3% quantile in figures 43 and 44.

Country and time specific regression results. Furthermore we calculated time specific levels for u, that changed from year to year. We tried to set u to the country wide average of the profit of a certain year and the country wide 0,2% and 0,3% quantiles of every year. We tried to identify risk drivers that made a city perform better or worse than a yearly changing country wide level. Find the regression results for the 0,2% quantile and the 0,3% quantile in figures 45 and 46.

Growth of profit for each city. Additionally we also tried to set the dependent variable y in the Logit Model based on whether or not the profit for a certain city c grows from one point in time qy to the next.

$$growthprofit = \frac{profit_{c,qy} - profit_{c,qy-1}}{profit_{c,qy-1}} \tag{34}$$

$$y = \begin{cases} 0, \text{if growthprofit} > 0 \\ 1, \text{if growthprofit} < 0 \end{cases} \tag{35}$$

Find the regression results for this approach in figure 47.

Additionally we tried different constant levels for u ranging from 5% to 15% and tested a different profit calculation

$$profit_{c,t} = \frac{rent_{c,t-a} + purchase_{c,t+1}}{purchase_{c,t}} - 1 \tag{36}$$

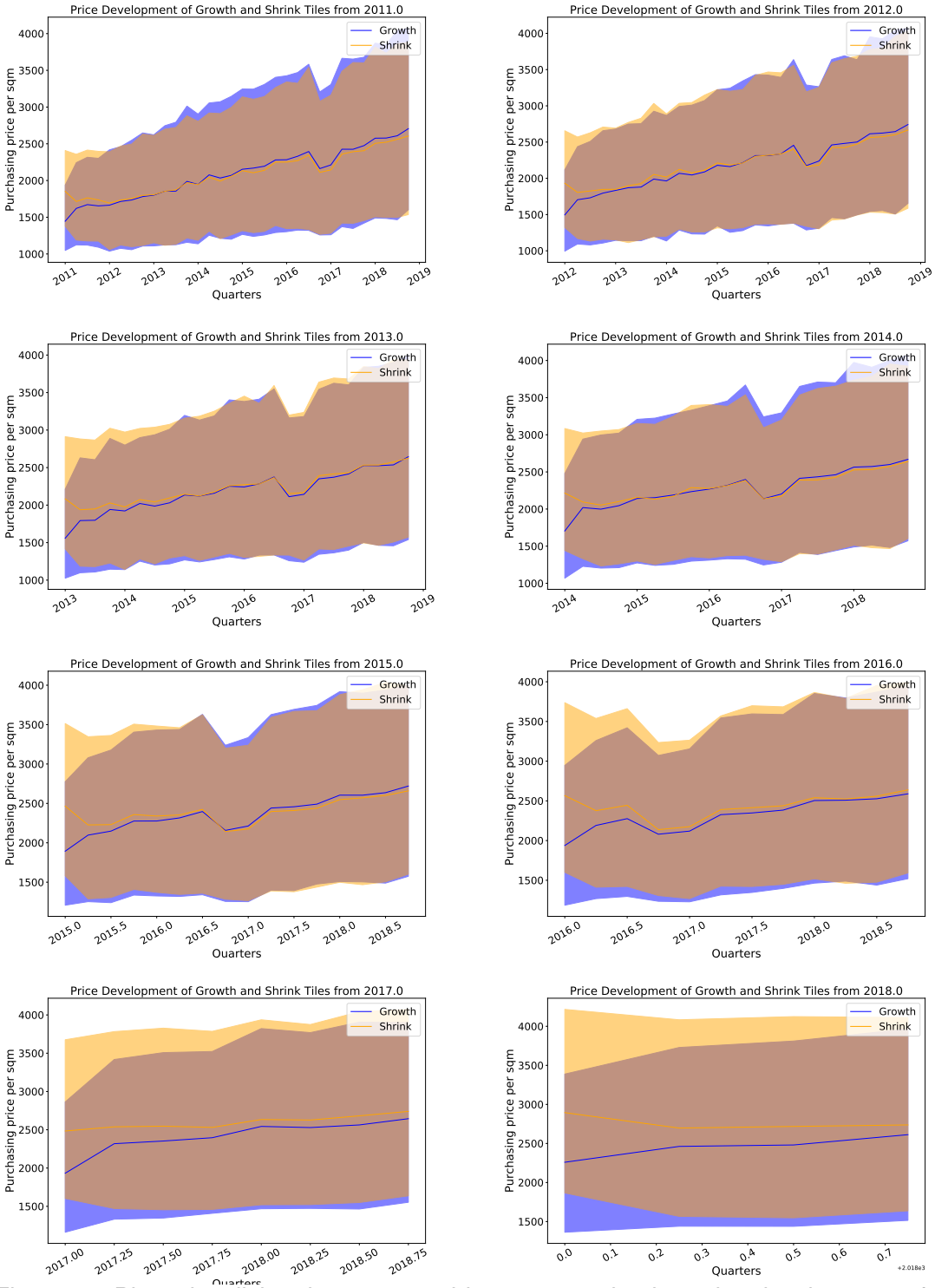


Figure 40. Plots visualising the country wide mean purchasing price development of city specific growth and shrink tiles. Updated every quarter, removed values that are higher (lower) than the 90% (10%) quantile.

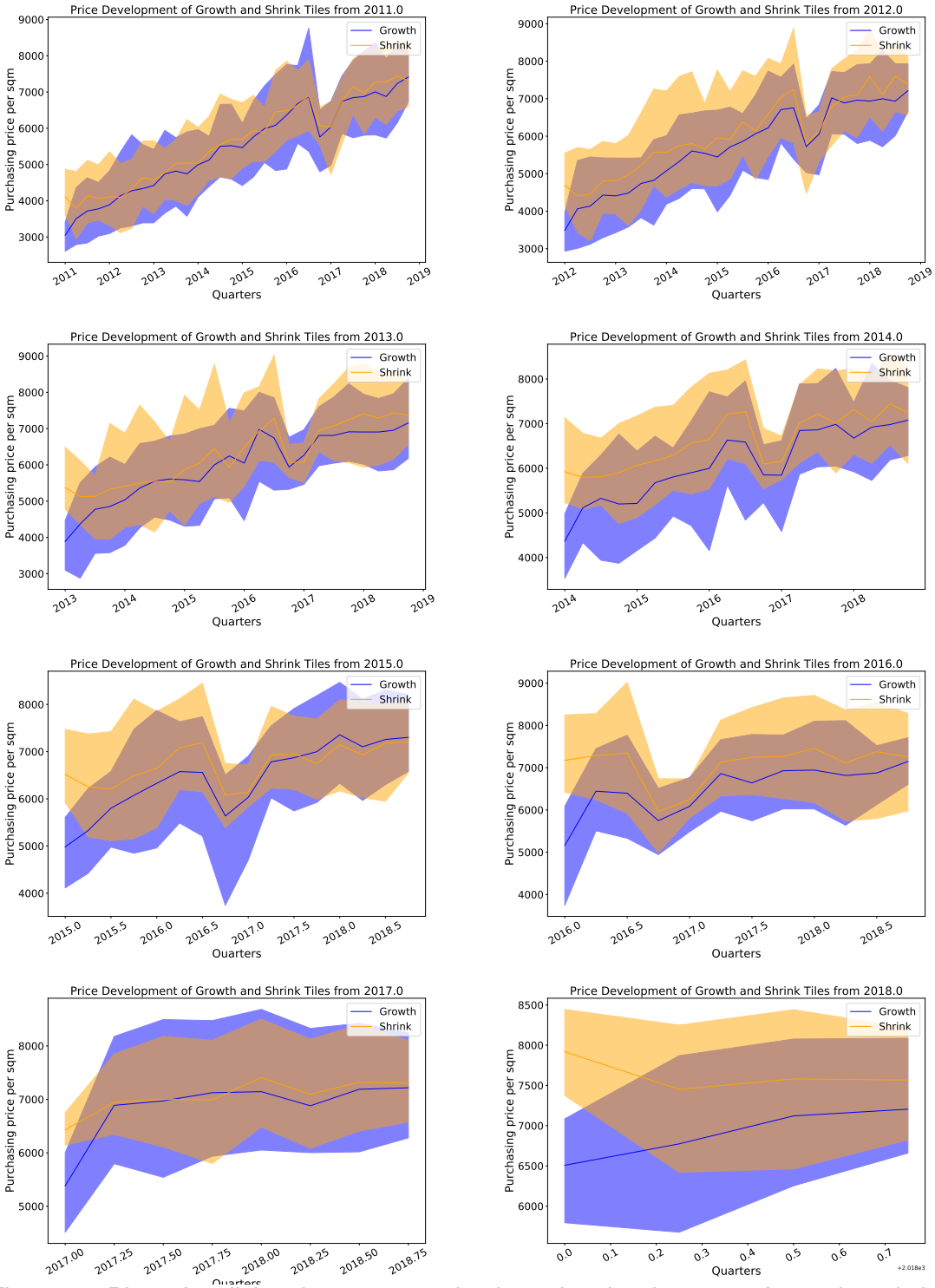


Figure 41. Plots visualising the mean purchasing price development of growth and shrink tiles in Munich. Updated every quarter, removed values that are higher (lower) than the 90% (10%) quantile.

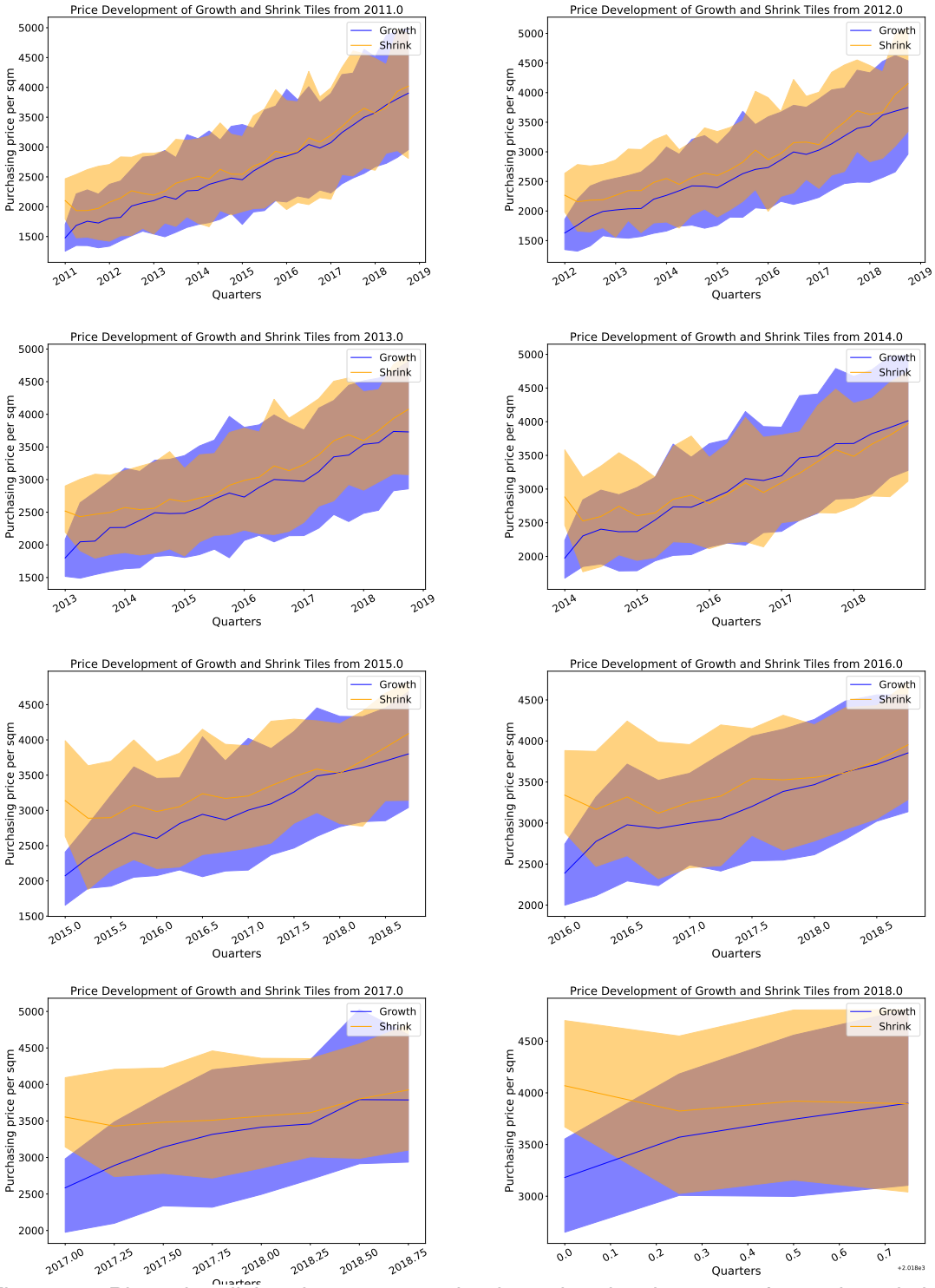


Figure 42. Plots visualising the mean purchasing price development of growth and shrink tiles in Berlin. Updated every quarter, removed values that are higher (lower) than the 90% (10%) quantile.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.165
Dependent Variable:	y	AIC:	1048.8658
Date:	2020-07-16 10:47	BIC:	1152.8128
No. Observations:	1043	Log-Likelihood:	-503.43
Df Model:	20	LL-Null:	-602.68
Df Residuals:	1022	LLR p-value:	2.2292e-31
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
growth_Schuldnerquote	1.9116	0.6896	2.7719	0.0056	0.5599	3.2632
growth_birth_death	2.2272	0.7066	3.1522	0.0016	0.8424	3.6121
mult_Studierende_x_growth_app_licence_residential	2.7004	1.1025	2.4494	0.0143	0.5396	4.8613
mult_Pendlersaldo_x_growth_Growth_Shrink_Ratio	4.0816	1.5464	2.6395	0.0083	1.0508	7.1124
mult_growth_Einkommensteuer_x_Growth_Shrink_Ratio	-3.7731	1.8074	-2.0875	0.0368	-7.3156	-0.2306
mult_growth_Schuldnerquote_x_growth_Growth_Shrink_Ratio	-4.3626	1.9347	-2.2550	0.0241	-8.1545	-0.5707
mult_growth_Anteil Teilzeitbeschäftigte_x_growth_Ehescheidungen	3.7017	1.7103	2.1644	0.0304	0.3497	7.0538
mult_growth_Empfänger von Grundsicherung im Alter (Altersarmut)_x_growth_Betten in FV-Betrieben	-2.5402	1.2630	-2.0112	0.0443	-5.0157	-0.0647
mult_growth_Empfänger von Grundsicherung im Alter (Altersarmut)_x_growth_Growth_Shrink_Ratio	-2.7814	1.4921	-1.8641	0.0623	-5.7058	0.1430
mult_growth_Lebenserwartung_x_growth_SGB II - Quote	2.4333	1.5141	1.6072	0.1080	-0.5342	5.4008
mult_growth_Betten in FV-Betrieben_x_growth_app_stock	-2.0152	1.1296	-1.7840	0.0744	-4.2291	0.1988
growth_Schuldnerquote_lagged_1	1.4059	0.6534	2.1518	0.0314	0.1253	2.6865
growth_Existenzgründungen_x_lagged_1	-1.5870	0.9726	-1.6318	0.1027	-3.4932	0.3192
growth_Anteil Teilzeitbeschäftigte_lagged_2	-2.6069	0.7631	-3.4162	0.0006	-4.1026	-1.1113
growth_Beschäftigte am Wohnort mit akademischem Abschluss_lagged_1	-1.6878	0.7931	-2.1282	0.0333	-3.2422	-0.1334
growth_Beschäftigte am Wohnort mit akademischem Abschluss_lagged_2	1.4277	0.8347	1.7104	0.0872	-0.2084	3.0637
growth_Personen in Bedarfsgemeinschaften_lagged_2	2.3138	0.7500	3.0851	0.0020	0.8438	3.7838
growth_birth_death_lagged_2	-2.3581	0.7068	-3.3364	0.0008	-3.7433	-0.9729
Growth_Shrink_Ratio_lagged_1	2.3826	1.1157	2.1356	0.0327	0.1959	4.5694
growth_Growth_Shrink_Ratio_lagged_1	2.8393	1.0358	2.7411	0.0061	0.8091	4.8695
const	0.6792	1.8919	0.3590	0.7196	-3.0288	4.3871

Figure 43. Regression results for u=0.2% quantile of profit for each city.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.147
Dependent Variable:	y	AIC:	1242.2585
Date:	2020-07-16 10:47	BIC:	1346.2055
No. Observations:	1043	Log-Likelihood:	-600.13
Df Model:	20	LL-Null:	-703.85
Df Residuals:	1022	LLR p-value:	3.7680e-33
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
growth_Anteil Schutzsuchender an Bevölkerung	2.4021	1.4402	1.6678	0.0954	-0.4208	5.2249
growth_Betten in FV-Betrieben	-1.4077	0.3137	-4.4875	0.0000	-2.0225	-0.7929
growth_birth_death	1.4397	0.8157	1.7649	0.0776	-0.1591	3.0385
Growth_Shrink_Ratio	-2.3552	1.3908	-1.6934	0.0904	-5.0811	0.3707
mult_Studierende_x_growth_app_licence_residential	1.9324	1.1009	1.7553	0.0792	-0.2253	4.0900
mult_Pendlersaldo_x_growth_migrbal	3.4868	1.4647	2.3806	0.0173	0.6161	6.3575
mult_Population_x_growth_migrbal	2.7298	1.5762	1.7319	0.0833	-0.3595	5.8192
mult_growth_Einkommensteuer_x_growth_Population	-2.8770	0.8641	-3.3297	0.0009	-4.5705	-1.1835
mult_growth_Anteil Schutzsuchender an Bevölkerung_x_growth_birth_death	1.7744	1.6211	1.0946	0.2737	-1.4029	4.9517
mult_growth_Empfänger von Grundsicherung im Alter (Altersarmut)_x_growth_Growth_Shrink_Ratio	-3.3626	1.2257	-2.7433	0.0061	-5.7650	-0.9602
mult_growth_Lebenserwartung_x_growth_Personen in Bedarfsgemeinschaften	3.0591	1.1988	2.5518	0.0107	0.7095	5.4087
mult_growth_Personen in Bedarfsgemeinschaften_x_growth_svb_living	-2.7218	0.9642	-2.8229	0.0048	-4.6115	-0.8320
growth_Existenzgründungen_x_lagged_2	1.3121	0.9248	1.4187	0.1560	-0.5006	3.1247
growth_Anteil Teilzeitbeschäftigte_lagged_2	-2.3958	0.6770	-3.5386	0.0004	-3.7228	-1.0688
growth_Beschäftigte am Wohnort mit akademischem Abschluss_lagged_2	1.6569	0.6695	2.4749	0.0133	0.3447	2.9691
growth_Ehescheidungen_lagged_2	-2.3019	0.9702	-2.3725	0.0177	-4.2035	-0.4003
growth_birth_death_lagged_2	-1.3555	0.6034	-2.2465	0.0247	-2.5381	-0.1729
Shrink_Quadrant_Sum_lagged_1	-2.0931	1.0937	-1.9138	0.0556	-4.2368	0.0505
Growth_Shrink_Ratio_lagged_1	2.5304	1.3661	1.8523	0.0640	-0.1471	5.2080
growth_Growth_Shrink_Ratio_lagged_1	2.0444	1.3621	1.5009	0.1334	-0.6253	4.7140
const	-1.0508	1.5012	-0.7000	0.4839	-3.9931	1.8914

Figure 44. Regression results for u=0.2% quantile of profit for each city.

,where $rent_{c,t-a}$ denotes the previous rent, lagged by a years. We did this for $a \in \{1, 2\}$. This approach was chosen to proxy the not observable contractual rent. Rent prices from 21st are market rent prices, which are usually higher than the contractual rents, which investors would really earn when buying a real estate object. Since the rents are constantly rising, approximating the contractual rents by the lagged market rents is a viable solution. In total we tested 40 different values for y. You can find regression results for all levels in our gitlab documentation.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.165
Dependent Variable:	y	AIC:	917.0805
Date:	2020-07-17 15:55	BIC:	1021.0275
No. Observations:	1043	Log-Likelihood:	-437.54
Df Model:	20	LL-Null:	-523.85
Df Residuals:	1022	LLR p-value:	2.6687e-26
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Lebenserwartung	-1.9370	0.7028	-2.7562	0.0058	-3.3145	-0.5596
Einwohner-Arbeitsplatz-Dichte	-1.7562	0.7618	-2.3054	0.0211	-3.2492	-0.2631
growth_Population	-4.6622	2.4293	-1.9191	0.0550	-9.4236	0.0992
growth_birth_death	1.5492	0.7047	2.1984	0.0279	0.1680	2.9304
Growth_Shrink_Ratio	-3.6372	2.4434	-1.4886	0.1366	-8.4261	1.1517
mult_Income_x_growth_app_completion_residential	2.2533	0.9049	2.4902	0.0128	0.4798	4.0269
mult_Studierende_x_growth_Growth_Shrink_Ratio	4.5541	1.6297	2.7945	0.0052	1.3600	7.7481
mult_Einkommensteuer_x_Growth_Shrink_Ratio	-0.6949	1.7336	-0.4008	0.6885	-4.0926	2.7028
mult_Schuldnerquote_x_birth_death	-2.5359	0.9287	-2.7306	0.0063	-4.3561	-0.7157
mult_Pendlersaldo_x_growth_app_licence_residential	5.0537	1.7995	2.8084	0.0050	1.5268	8.5806
mult_growth_Einkommensteuer_x_growth_Population	0.5859	1.8542	0.3160	0.7520	-3.0482	4.2201
mult_growth_Durchschnittsalter der Bevölkerung _x_growth_Growth_Shrink_Ratio	-2.6840	1.2199	-2.2002	0.0278	-5.0749	-0.2931
mult_growth_Arbeitslosenquote_x_growth_Beschäftigte Tertiärer Sektor	-3.1148	1.4051	-2.2169	0.0266	-5.8687	-0.3610
mult_growth_Lebenserwartung_x_growth_SGB II - Quote	2.8529	1.5629	1.8254	0.0679	-0.2104	5.9162
mult_growth_SGB II - Quote_x_growth_birth_death	4.7329	1.9797	2.3907	0.0168	0.8528	8.6129
Income_lagged_1	-0.8868	0.4959	-1.7882	0.0737	-1.8587	0.0852
growth_Einwohner-Arbeitsplatz-Dichte_lagged_2	-1.6509	1.0234	-1.6132	0.1067	-3.6566	0.3549
growth_birth_death_lagged_2	-1.6357	0.7030	-2.3268	0.0200	-3.0136	-0.2579
Growth_Shrink_Ratio_lagged_1	2.6757	1.4700	1.8202	0.0687	-0.2055	5.5570
growth_Growth_Shrink_Ratio_lagged_1	2.7310	1.3874	1.9685	0.0490	0.0119	5.4502
const	-1.9665	1.7571	-1.1191	0.2631	-5.4103	1.4774

Figure 45. Regression results for u=0.2% quantile of profit for Germany each year.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.135
Dependent Variable:	y	AIC:	1147.2434
Date:	2020-07-17 15:56	BIC:	1251.1904
No. Observations:	1043	Log-Likelihood:	-552.62
Df Model:	20	LL-Null:	-638.90
Df Residuals:	1022	LLR p-value:	2.7548e-26
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Lebenserwartung	-1.1205	0.4903	-2.2854	0.0223	-2.0814	-0.1596
Population	-3.3977	1.8790	-1.8083	0.0706	-7.0805	0.2850
growth_Population	-4.7360	1.9473	-2.4321	0.0150	-8.5527	-0.9193
Growth_Shrink_Ratio	-2.8440	1.4960	-1.9010	0.0573	-5.7762	0.0882
mult_Income_x_Pendlersaldo	-1.5351	0.9748	-1.5748	0.1153	-3.4456	0.3754
mult_Income_x_growth_Ehescheidungen	2.6519	1.1926	2.2236	0.0262	0.3144	4.9894
mult_Studierende_x_growth_app_licence_residential	2.0901	1.3054	1.6012	0.1093	-0.4684	4.6485
mult_Pendlersaldo_x_growth_app_licence_residential	3.5792	1.8076	1.9801	0.0477	0.0364	7.1220
mult_growth_Einkommensteuer_x_growth_Population	0.1398	1.4675	0.0952	0.9241	-2.7365	3.0161
mult_growth_Durchschnittsalter der Bevölkerung _x_growth_Steuerkraft	2.8019	1.2144	2.3071	0.0210	0.4216	5.1821
mult_growth_Existenzgründungen_x_x_growth_app_stock	2.4294	1.0005	2.4281	0.0152	0.4684	4.3903
mult_growth_Steuerkraft_x_growth_Bruttowertschöpfung	2.7859	1.3660	2.0394	0.0414	0.1085	5.4632
mult_growth_Empfänger von Grundsicherung im Alter (Altersarmut)_x_growth_Growth_Shrink_Ratio	-2.4207	1.2194	-1.9852	0.0471	-4.8106	-0.0308
mult_growth_Personen in Bedarfsgemeinschaften_x_growth_svb_living	-2.7542	0.9961	-2.7651	0.0057	-4.7065	-0.8020
mult_growth_Ehescheidungen_x_growth_birth_death	-4.4506	1.9144	-2.3249	0.0201	-8.2027	-0.6985
growth_SGB II - Quote_lagged_1	2.0082	0.7537	2.6643	0.0077	0.5309	3.4855
growth_Einwohner-Arbeitsplatz-Dichte_lagged_2	-2.0813	0.8892	-2.3406	0.0193	-3.8241	-0.3385
Shrink_Quadrant_Sum_lagged_1	-1.1920	1.7077	-0.6980	0.4852	-4.5389	2.1550
Growth_Shrink_Ratio_lagged_1	2.7980	1.3794	2.0285	0.0425	0.0945	5.5016
growth_Growth_Shrink_Ratio_lagged_1	1.9986	1.2766	1.5656	0.1174	-0.5035	4.5006
const	-0.6366	2.6716	-0.2383	0.8117	-5.8730	4.5997

Figure 46. Regression results for u=0.2% quantile of profit for Germany each year.

Alternative Risk Model Results

The outcomes of the risk model heavily depend on which values we choose as the dependent variables for the regressions in Logit and Location Factor Model. To come up with our results in the main part we choose those values that lead to the highest measure of certainty R^2 in the regressions, but since the differences in R^2 are partly quite small it is worth investigating, what the risk model results would look like for different dependent variables y.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.169
Dependent Variable:	y	AIC:	1206.9431
Date:	2020-07-17 15:58	BIC:	1310.8900
No. Observations:	1043	Log-Likelihood:	-582.47
Df Model:	20	LL-Null:	-701.05
Df Residuals:	1022	LLR p-value:	4.3902e-39
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
growth_SGB II - Quote	-2.3052	0.8492	-2.7145	0.0066	-3.9696	-0.6408
Growth_Shrink_Ratio	-1.7337	1.0053	-1.7246	0.0846	-3.7040	0.2366
growth_Growth_Shrink_Ratio	-0.2056	1.3714	-0.1499	0.8808	-2.8936	2.4824
mult_Pendlersaldo_x_growth_app_licence_residential	7.5125	1.7126	4.3866	0.0000	4.1559	10.8692
mult_Pendlersaldo_x_growth_Growth_Shrink_Ratio	4.8113	1.4251	3.3762	0.0007	2.0183	7.6044
mult_Population_x_growth_Growth_Shrink_Ratio	-4.7674	1.9688	-2.4215	0.0155	-8.6260	-0.9087
mult_growth_Durchschnittsalter der Bevölkerung _x_growth_Arbeitslosenquote	-4.0571	1.1810	-3.4353	0.0006	-6.3719	-1.7424
mult_growth_Durchschnittsalter der Bevölkerung _x_growth_Growth_Shrink_Ratio	-2.7789	1.2910	-2.1525	0.0314	-5.3092	-0.2485
mult_growth_Schuldnerquote_x_growth_Arbeitslosenquote	2.7667	1.0010	2.7640	0.0057	0.8048	4.7286
mult_growth_Schuldnerquote_x_growth_migrbal	-4.2528	1.8866	-2.2542	0.0242	-7.9505	-0.5550
mult_growth_Steuerkraft_x_growth_Anteil Schutzsuchender an Bevölkerung	3.5199	1.3181	2.6703	0.0076	0.9363	6.1034
mult_growth_Empfänger von Grundsicherung im Alter (Altersarmut)_x_growth_Growth_Shrink_Ratio	-2.2823	1.3940	-1.6373	0.1016	-5.0144	0.4498
mult_growth_SGB II - Quote_x_growth_Beschäftigte Tertiärer Sektor	-2.1347	1.3699	-1.5583	0.1192	-4.8196	0.5503
mult_growth_Ehescheidungen_x_growth_birth_death	-3.2006	1.7877	-1.7904	0.0734	-6.7043	0.3032
growth_Beschäftigte am Wohnort mit akademischem Abschluss_lagged_1	-2.0631	0.6222	-3.3156	0.0009	-3.2827	-0.8436
growth_Personen in Bedarfsgemeinschaften_lagged_1	2.8734	0.7342	3.9138	0.0001	1.4344	4.3123
growth_svb_living_lagged_1	1.3643	0.5559	2.4544	0.0141	0.2748	2.4537
growth_svb_living_lagged_2	1.6151	0.4714	3.4257	0.0006	0.6910	2.5391
growth_svb_working_lagged_1	1.6047	0.8301	1.9332	0.0532	-0.0222	3.2316
growth_Growth_Shrink_Ratio_lagged_1	1.6070	0.8837	1.8185	0.0690	-0.1250	3.3391
const	0.2257	2.6376	0.0856	0.9318	-4.9440	5.3954

Figure 47. Regression results for growth profit > 0.

Variance of Profit Growth instead of Sharpe Ratio in Location Factor Model. Replacing the Sharpe Ratio by the Variance of profit growth does not change the results significantly (figure 48 and 49).

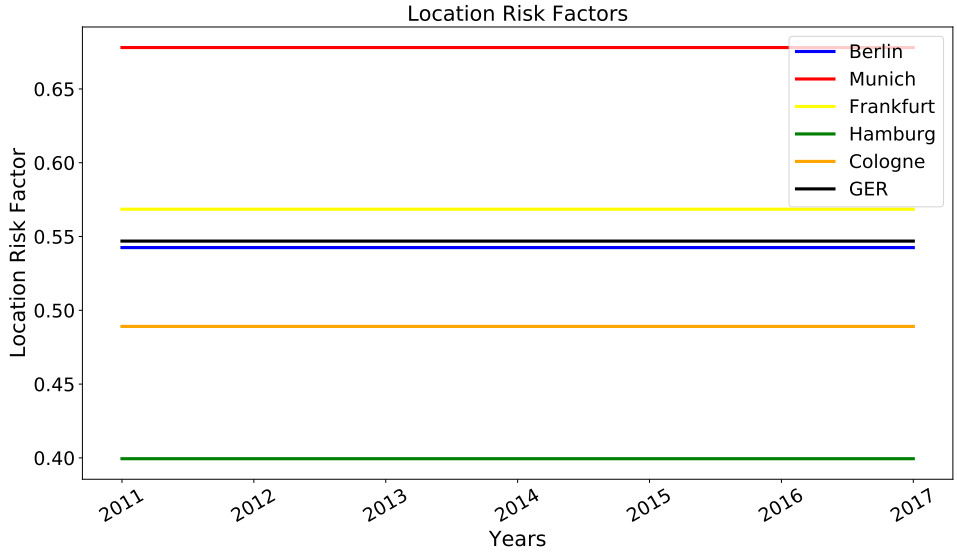


Figure 48. Location Factor risk score using Variance of Profit Growth.

City specific level of u in Logit Model. Another important approach is to calculate a city specific risk score. We can thereby identify city specific risk drivers and get a more detailed picture, but it is also makes it more difficult to compare the risk of different cities to each other. The risk score from the location factor model is based on the Sharpe Ratio

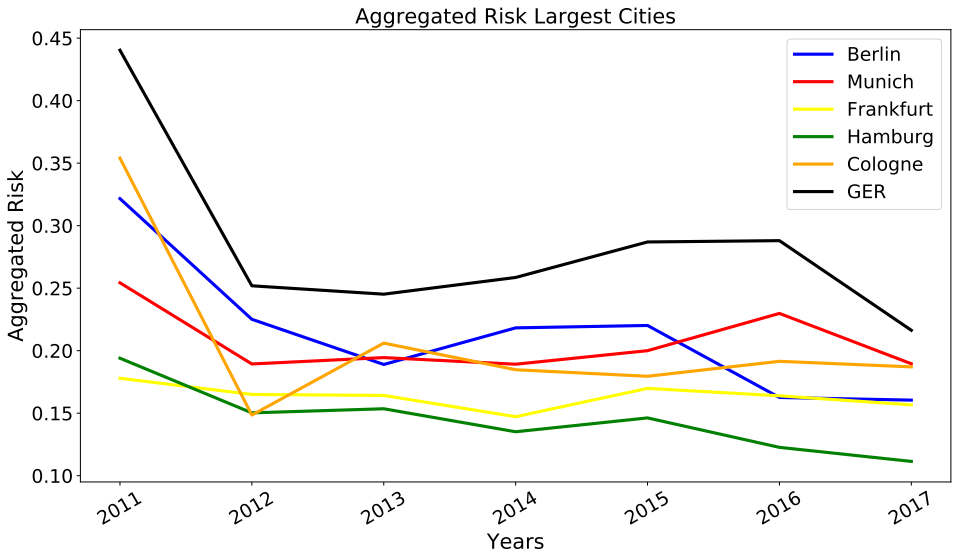


Figure 49. Combined risk score using Variance of Profit Growth.

like in the main part. Interestingly when taking a look the trend over several cities, risk seems to increase more strongly towards the end of the time frame (Figure 50 and 51) in comparison to the risk calculated based on constant level of u.

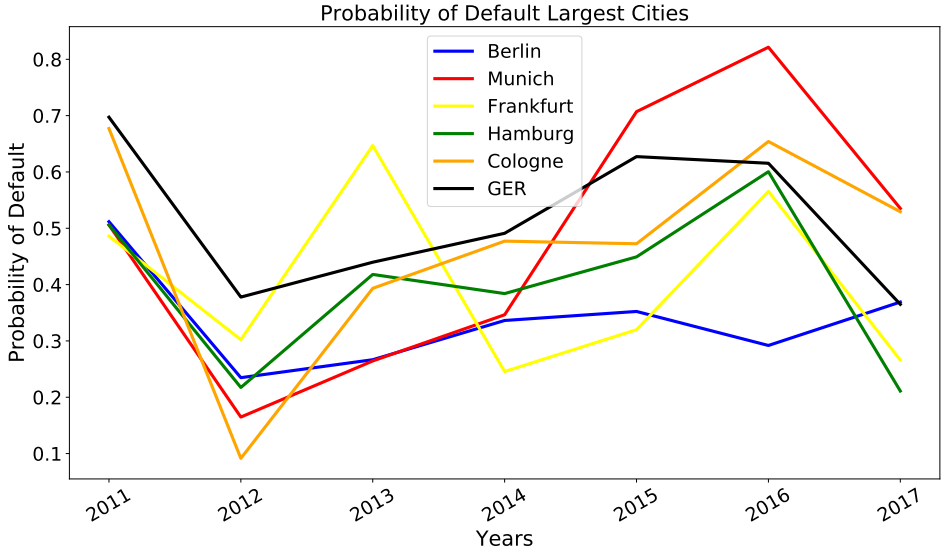


Figure 50. Probability of default using u=0,2 quantile of profits for a city of whole timeframe.

Growth of profit in Logit Model. As stated in we also set y in the Logit Model based on future growth of the profit in a certain city c. Interestingly the profit over the whole timeframe is more or less stable over time, therefore this approach ends up labelling nearly

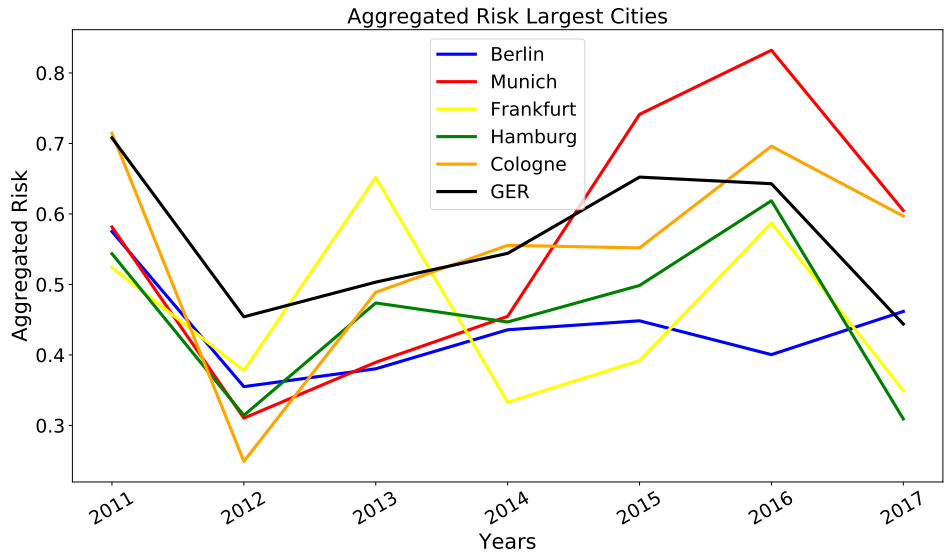


Figure 51. Combined risk score using u=0,2 quantile of profits for a city of whole timeframe.

all cities with an extremely high risk over the whole timeframe (figures 52 and 53).

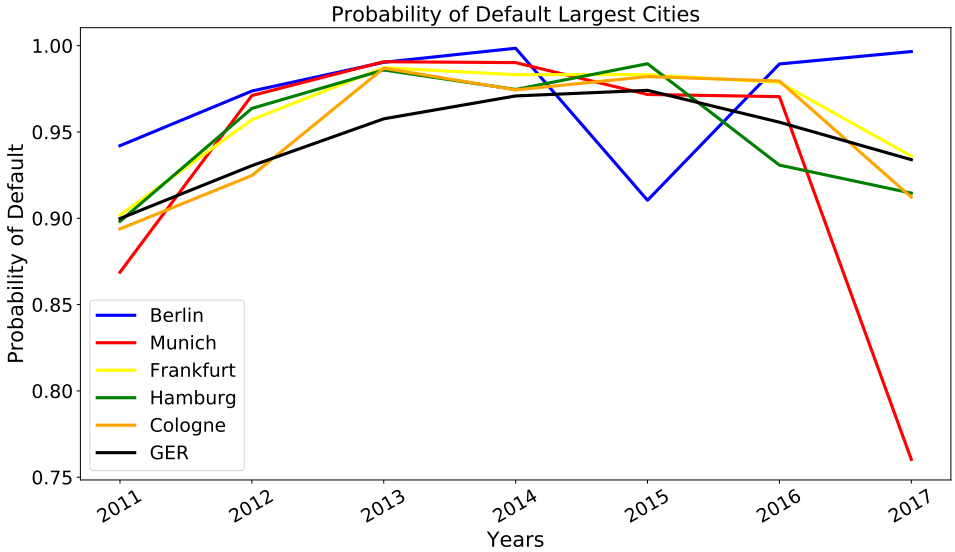


Figure 52. Probability of default basing y on whether the profit for a city is growing or not for a city of whole timeframe.

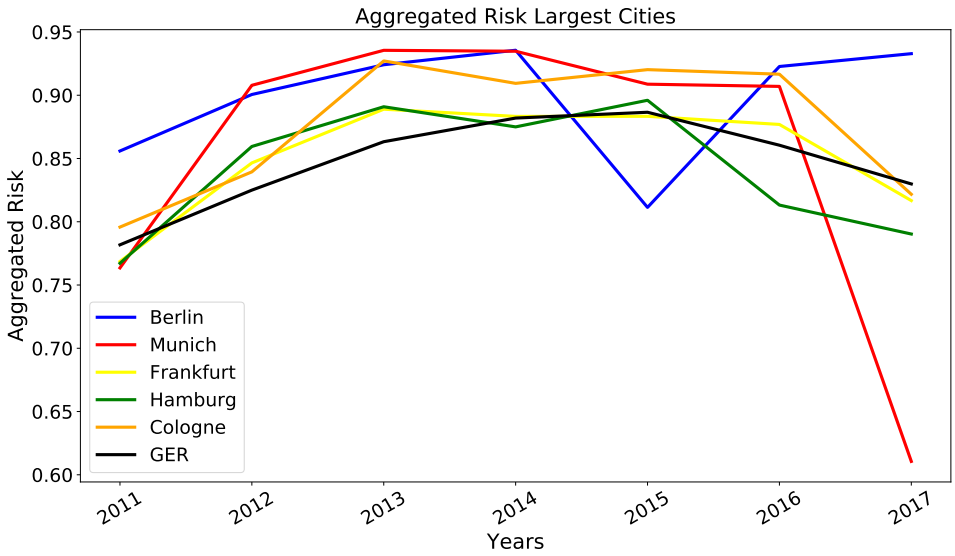


Figure 53. Probability of default basing y on whether the profit for a city is growing or not for a city of whole timeframe.

A, B and C-Cities

We base the categorisation of cities into A, B and C cities on the paper from bulwiengesa [Bulwiengesa, 2019]. They roughly describe the different clusters as follows: A-cities are cities with international and national importance and a fully functioning real estate market in all segments. B-cities are nationally and regionally important, whereas C-cities are only partly nationally important but usually function as regional hubs.

A-Cities	B-Cities	C-Cities
Berlin	Bochum	Aachen
Düsseldorf	Bonn	Augsburg
Frankfurt (Main)	Bremen	Bielefeld
Köln	Dresden	Darmstadt
München	Duisburg	Erfurt
Stuttgart	Essen	Freiburg
	Hannover	Heidelberg
	Karlsruhe	Kiel
	Leipzig	Lübeck
	Mannheim	Magdeburg
	Münster	Mainz
	Nürnberg	Mönchengladbach
	Wiesbaden	Offenbach (Main)
		Osnabrück
		Potsdam
		Regensburg
		Rostock
		Saarbrücken
		Wuppertal

Table 3
City categorization in A-, B- and C-Cities

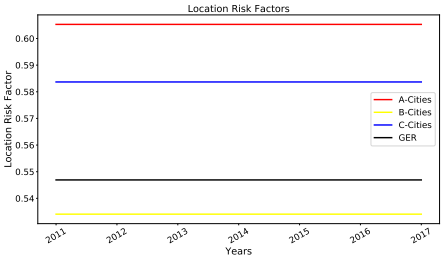


Figure 54. Mean of Normalised Risk Score from Location Factor model using SharpeRatio as dependent variable for A, B and C-Cities.

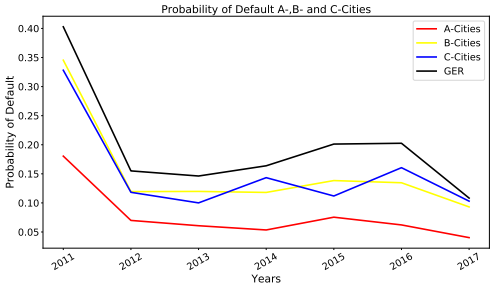


Figure 55. Mean of Probability of Default over time in the five largest German cities and the German average for A, B and C-Cities.

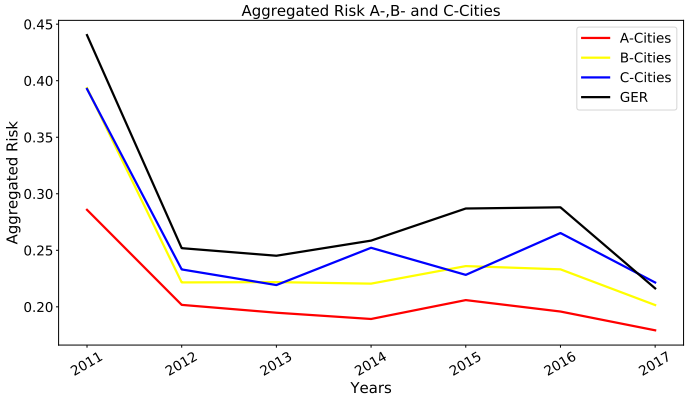


Figure 56. Mean of the general outcome of Risk Model combining Logit and Location Factor Model for A, B and C-Cities.

Variable	Description	Example Value
AGS	21RE unique city identifier	1001000
Year	Observation year	2011
Population	Population	82801
migrbal	Migration balance	-5751
birth	Births	763
death	Deaths	952
birth_death	Birth divided by death	0,801470588
unemp	Unemployed persons	5151
unemp_rate	Unemployment rate	11,8
app_stock	Apartment stock	48109
app_licence_residential	Construction permits for residential buildings	320
app_licence_nonresidential	Construction permits for non-residential buildings	6
app_licence_total	Construction permits in total	326
app_completion_residential	Construction completions for residential buildings	165
app_completion_nonresidential	Construction completions for non-residential buildings	1
app_completion_total	Construction completions in total	166
svb_living	Employees paying social insurance in the given area	27016
svb_working	Employees paying social insurance outside the given area	39028

Table 4
ags_data.csv data: For every AGS basic variables are given on a yearly basis.

Variable	Description	Example Value
AGS	21RE unique city identifier	1001000
cid_id	21RE unique city tile identifier	9
Year	Observation year	2011
Income	Average annual income	57365269461077700.00

Table 5
cid_income.csv data: For every city on tile level average annual income.

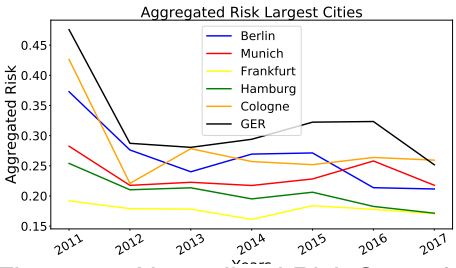


Figure 57. Normalised Risk Score from Location Factor model using SharpeRatio as dependent variable

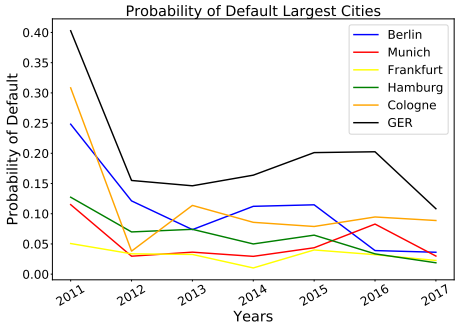


Figure 58. Probability of default over time in the five largest German cities and the German average

Variable	Description	Example Value
AGS	21RE unique city identifier	1001000
Year	Observation year	2011
cb_id	21RE unique city tile identifier	9
rent_cell	Rent per m^2 of comparable residential buildings [€]	5.81857317757772
data_availability_rent	Availability of rent data [not used]	47.8800325456365
sale_cell	House price per m^2 of comparable residential buildings [€]	1368.55592915848
data_availability_sale	Availability of house price data [not used]	54.3473363705241

Table 6
cid_prices.csv data: For every city on tile level average rent and sale prices.

Variables	Comments
Baulandumsatz	Durchschnittlich umgesetzte Fläche baureifen Landes je ha Siedlungsfläche
Baulandpreise	Durchschnittliche Kaufwerte für Bauland in [€] je m^2
Städtebauförderung	Städtebauförderung insgesamt (langfristig) in [€] je Einwohner
GRW gewerbliche Wirtschaft	Zuschüsse 'Verbesserung der Einzelbetriebe' in [€] je Einwohner
GRW Infrastruktur	Zuschüsse 'Verbesserung der Infrastruktur' in [€] je Einwohner
Hochschulförderung	Hochschulförderung in [€] je Einwohner
Erreichbarkeit Autobahnen	Pkw-Fahrtzeit zur nächsten Autobahnanschlussstelle in Minuten
Erreichbarkeit Flughafen	Pkw-Fahrtzeit zum nächsten international Flughafen in Minuten
Erreichbarkeit Oberzentren	Pkw-Fahrtzeit zum nächsten Oberzentrum in Minuten
Distanz Supermarkt	Nahversorgung Supermärkte Durchschnittsdistanz
Distanz Apotheken	Nahversorgung Apotheken Durchschnittsdistanz
Distanz Grundschulen	Einwohnergewichtete Luftliniendistanz zur nächsten Grundschule
Siedlungs- und Verkehrsfläche	Anteil der Fläche in %
Erholungsfläche	Anteil Erholungsfläche an der Fläche in %
Ärtzte	Ärzte je 10.000 Einwohner
Pflegebedürftige	Pflegebedürftige je 10.000 Einwohner
Leistungen für Wohngeld	Durchschnittliche monatliche Leistungen für Wohngeld in [€] je Haushalt

Table 7
INKAR variables without time series data: Only on city level. For a more detailed description refer to the INKAR website, see BBR, 2020b

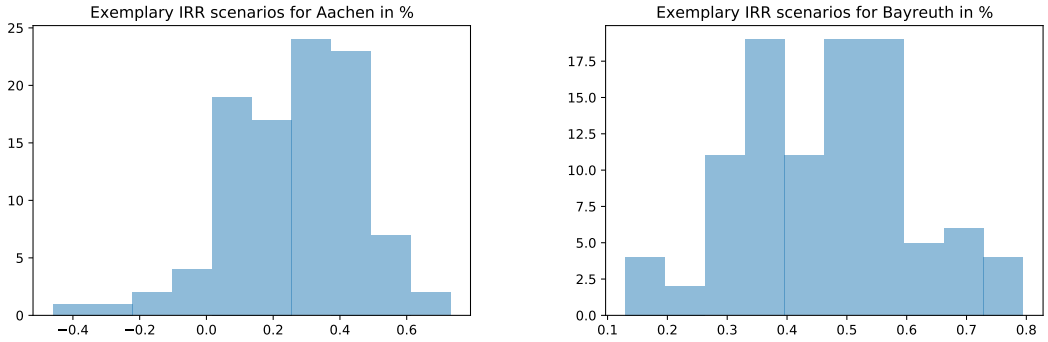


Figure 59. Unlevered IRR distribution for two selected cities. 100 samples were taken each.

Variables	Comment	Years
Pendlersaldo	Pendlersaldo je 100 SV Beschäftigte am Arbeitsort	97-17
Abfallmenge	Entsorgte oder behandelte Abfallmenge je Einwohner in kg	02-15
Krankenhausbetten	Krankenhausbetten je 1000 Einwohner	04-16
Pkw Dichte	Pkw je 1.000 Einwohner	08-17
Wohnfläche	Wohnfläche je Einwohner in m^2	95-17
Ein- und Zweifamilienhäuser	Anteil der Wohngebäude mit 1 und 2 Wohnungen in %	95-17
Anteil Mehrfamilienhäuser	Anteil Wohnungen in Mehrfamilienhäusern an allen Wohnungen in %	95-17
Junge zu Alten Erwerbsfähigen	Verhältnis junge (15-<20J) zu alten (60-<65J) Erwerbsfähigen in %	95-17
Anteil Teilzeitbeschäfitigte	Anteil der SV Beschäftigten (Teilzeit) in %	12-17
Akademiker	Anteil Beschäftigte mit akademischem Berufsabschluss	12-17
Sekundärer Sektor	Beschäftigte sekundärer Sektor	08-17
Tertiärer Sektor	Beschäftigte tertiärer Sektor	08-17
Durchschnittsalter	Durchschnittsalter der Bevölkerung in Jahren	95-17
Ehescheidungen	Ehescheidungen je 1.000 Einwohner 18 Jahre und älter	03-17
Schutzsuchende	Anteil Schutzsuchender an Bevölkerung in %	07-17
Lebenserwartung	Mittlere Lebenserwartung eines Neugeborenen in Jahren	93-15
Studenten	Studierende an Hochschulen je 1.000 Einwohner	95-17
Schuldnerquote	Private Schuldner je 100 Einwohner 18 Jahre und älter	04-17
Betten in FV-Betrieben	Betten in Fremdenverkehrsbetrieben je 1.000 Einwohner	95-17
GroSSunternehmen	Anteil der Betriebe mit mehr als 250 SV-Beschäftigten in permille	06-15
Umsatzsteuer	Umsatzsteuer in 1000 [€] je umsatzsteuerpflichtige Betriebe	09-17
Bruttowertschöpfung	Bruttowertschöpfung insgesamt in 1.000 [€] je Erwerbstätigen	00-17
BIP	Bruttoinlandsprodukt je Einwohner	00-17
Existenzgründungen	Anzahl neuerrichtete Gewerbebetriebe je 1000 Einwohner	06-17
Mietpreise	Duchschnittliche Angebotsmiete je m^2 klassifiziert in Stufen	06-17
Steuerkraft	Gemeindliche Steuerkraft in [€] je Einwohner	95-17
Einkommensteuer	Einkommensteuer in [€] je Einwohner	95-17
Gewerbesteuer	Gewerbesteuer in [€] je Einwohner	01-17
Einwohner-Arbeitsplatz-Dichte	Einwohner und Beschäftigte je km^2	97-17
SGB II - Quote	Anteil der Leistungsberechtigten der unter 65-jährigen in %	10-17
Altersarmut	Empfänger von Grundsicherung im Alter (Altersarmut)	08-17
Bedarfsgemeinschaft	Personen in Bedarfsgemeinschaften je 1.000 Einwohner	10-17

Table 8

INKAR variables with time series data: For every city on a yearly basis. For a more detailed description refer to the INKAR website, see BBR, 2020b