Abstract

Geographical diversification is an intuitive approach to manage risk in a mortgage portfolio. This paper examines whether high levels of geographical concentration in mortgage debt hinder the ability to diversify risk. To accomplish this, I apply the empirical regularity in regional science known as the rank-size rule, which is a log-linear relationship between city size and rank. With this, I estimate the degree of concentration across various sectors of the mortgage market. These results are improved upon by expanding to a non-linear model, which addresses an important concern regarding the linear fit. These concentration estimates suggest that there may be limits to diversification in heavily concentrated markets. This is particularly true for the non-conforming, jumbo mortgage market where immense concentrations in the largest metropolitan areas dominate the market. The insights from this approach provide a simplification for the relative exposure of mortgage markets to local economic shocks.
1 Introduction

In the context of mortgage portfolios, where loans are secured by the underlying real estate assets, geographical diversification is an attractive approach for risk management. However, the concentration of mortgage debt in major metropolitan areas can result in limitations to the amount of diversification that one can attain. This paper examines the use of rank-size relations to parameterize the geographical concentration of mortgage markets and the potential implications of large concentrations for portfolio diversification.

Rank-size relations provide a means to characterize the geographical concentration in mortgage debt. For example, the Zipf distribution (Zipf, 1949) suggests a linear relationship between the natural logarithms of city sizes and their respective ranks when sorted from largest to smallest. The traditional rank-size rule, or Zipf’s Law, is a special case where the slope of the log-linear relationship is equal to $-1$. This gives the simple result that the size of a city multiplied by its rank is constant, or that the second largest city is $1/2$ the size of the largest, the third is $1/3$ the size of the largest, and so on.

Power law distributions, such as Zipf’s, have been a long-standing empirical regularity with various applications including city sizes, firm sizes, wealth, international trade, and word-use frequencies across many languages. Fitting these distributions effectively produces measurements that describe the degree of concentration in a variable. As a result, the geographical distribution of mortgage debt can be simplified down to just one or two parameters.\footnote{Although originally credited to Auerbach (1913), the rank-size rule was popularized by Zipf (1949), and has since branched into a vast literature. Some more recent papers include: Brakman, Garretsen, Van Marrewijk, and Van Den Berg (1999); Axtell (2001); Reed (2002); Ioannides and Overman (2003); Klass, Biham, Levy, Malcai, and Solomon (2006); Gabaix (1999, 2011); Piantadosi (2014); Chaney (2018).}
For example, when applied to mortgage data from Black Knight Financial Services (BKFS), the estimated slope for the jumbo mortgage market is substantially steeper than that of the conforming loan market. This suggests a greater degree of geographical concentration for jumbo mortgages, which tend to be held as portfolio loans on bank balance sheets since they are ineligible to be purchased by the government sponsored enterprises (GSEs).

With the Zipf distribution, the limiting distribution of city sizes follows a power law distribution where the rate of geometric decline is given by the estimated slope parameter. However, when considering larger and more complete sets of cities, Eeckhout (2004) argues that the law of proportionate effect (city growth is independent of absolute size) results in a limiting distribution that is log-normal, which is known as Gibrat’s Law (Gibrat, 1931). Following some additional debate (Eeckhout, 2009; Levy, 2009), Malevergne, Pisarenko, and Sornette (2011) use the uniformly most powerful unbiased test to compare the two and conclude that the power law hypothesis should be accepted, and that the log-normal hypothesis be rejected.

However, rather than simply testing the hypothesis of city sizes following the traditional rank-size rule, I aim to parameterize the geographic concentration of mortgage debt. One notable drawback of the power law fit of the Zipf distribution is that it is most accurate in the tail and overestimates the size of the largest cities for the U.S. I address this concern by adapting the parabolic fractal distribution, which extends the log-linear form of the Zipf distribution to include a quadratic term for ln(rank). This greatly improves upon the linear fit, particularly for the largest cities.
The interpretation of the estimated parameters becomes more complicated with the parabolic fractal distribution. To address this, I orthogonalize the quadratic term from the linear term, which isolates the effects from each part. Thus, the additional fit provided by curvature can be compared directly to the portion that is explained by the linear term from the Zipf distribution. This can also be interpreted in the sense of comparing the power law fit from the linear component with a correction term that remedies the functional misspecification. I find that the relative effect of these two components is fairly constant in explaining the geographical concentration across both populations and the various sectors of the mortgage market.

Although the market shares of each city can be directly measured from the data, these rank-size relations provide a way to parameterize the distributions into just a single measure of concentration. The fitted rank-size curves are then scaled to produce a probability mass function or implied weights based on these estimates. These effectively represent theoretical weights for portfolios constructed by randomly sampling from the implied distributions. Alternatively, one can think of these weights as capturing the relative exposure of a specific market or variable to local shocks in each individual city.

When considering a portfolio of mortgages, the aggregation from the risk of individual assets to portfolio-level risk involves both a correlation and concentration component. In this manuscript, the focus is on modeling the concentration component using the aforementioned rank-size relations. The correlation across mortgage returns can be decomposed into weak spatial dependence and strong macroeconomic dependence (Dombrowski, Pace, and
Narayanan, 2019), which describe the potential for diversification. Given the
dependence structure of returns, the concentration component can limit the
ability for investors to achieve the suggested diversification potential.

In addition to the vast literature on the rank-size rule, and more generally,
the application of power laws in economics, this paper also contributes to
prior research studying the costs and benefits of geographical diversification.
Previous work in banking has examined the role of geographic expansion on
operational efficiency (Berger and DeYoung, 2001) as well as diversification
benefits for risk reduction in bank holding companies (Deng and Elyasiani,
2008) and small community banks (Emmons, Gilbert, and Yeager, 2004).

Within real estate, the role of geographical diversification has been studied
going back to Corgel and Gay (1987), who focus on diversifying across local
economic conditions. More recently, Cheng and Roulac (2007) measure the
effectiveness of geographical diversification in real estate investment, and Cot-
ter, Gabriel, and Roll (2014) find that increased market integration lowers the
diversification potential for housing investment. Market integration in this
case is analogous to the macroeconomic component of housing dependence,
which is shown in Section 3.2 to lower the risk effect of high geographical
concentration. This reduction in the ability to diversify risk follows from the
lesser potential for diversification that results from the larger correlations.

The remainder of the paper is structured as follows: Section 2 describes the
rank-size relationships and their estimation, Section 3 connects these distri-
butions with portfolio variance and implications for diversification, Section 4
applies them to the BKFS mortgage data, and finally, Section 5 concludes.
2 Rank-Size Relations

The rank-size rule, or Zipf’s Law, states that for a ranked set of observations, the size of a given observation \( (c_i) \) is inversely proportional to its rank \( i \).

\[
c_i \propto \frac{1}{i}
\]  

(1)

In the context of city populations, this suggests that the size of the \( i \)th ranked city will be equal to the population of the largest city \( (c_1) \) divided by its rank. For the U.S., New York City is the top ranked (largest) city with nearly 20 million people within its metropolitan area as of 2018. The traditional rank-size rule would then suggest that the second ranked city, Los Angeles, should have a population of roughly 10 million. However, this simple estimate leaves much room for improvement since the actual census estimates suggest more than 13 million in 2018.

One way to generalize this rank-size relation is to allow for different rates of geometric decline in the sequence of ranked sizes. This is done by introducing a shape parameter, \( \alpha \), which governs how quickly the sequence declines.

\[
c_i \propto \frac{1}{i^\alpha}
\]  

(2)

The shape parameter, \( \alpha \), is effectively a measure of the degree of concentration in a variable. For example, the special case of \( \alpha = 0 \) produces a result where every observation has equal size. Since \( i^0 = 1 \), the ranked observations are proportional to a constant, and thus \( c_1 = c_2 = \cdots = c_n \). As a result of the ranking procedure, this case acts as a lower bound and any amount of
geographical concentration will produce larger estimates of $\alpha$.

The traditional rank-size rule refers to the scenario where $\alpha = 1$. In this case, the second ranked city is half the size of the top ranked city, the third ranked city is one-third the size of the top ranked city, and so on. If one factors out the size of the top ranked city, what remains is the following sequence: $1/1, 1/2, 1/3, \ldots, 1/n$. If one views this sequence as a set of weights that sum to one, this normalized set of weights $r(x)$ can define a statistical distribution where (3) is the probability mass function.

$$r(x) = \frac{1/x}{1 + 1/2 + 1/3, \ldots, 1/n}$$  \hfill (3)

In its generalized form, this is known as the Zipf distribution, which has a probability mass function given by (4), where the scaling factor $H_{n,\alpha}$ is a generalized harmonic sum, as in (5).

$$f(x; n, \alpha) = x^{-\alpha}H_{n,\alpha}^{-1}$$  \hfill (4)

$$H_{n,\alpha} = \sum_{i=1}^{n} i^{-\alpha}$$  \hfill (5)

The Zipf distribution has some connections to a few other statistical distributions. For example, the Pareto distribution has the same general notion; however, it is defined for continuous variables as opposed to the discrete case of the Zipf distribution. Another related distribution is the zeta distribution, which is the limit of the Zipf distribution as $n \to \infty$. In this case, the sum in (5) becomes infinite, which is known as the Riemann zeta function. An interesting property of this function is the convergence of the Riemann zeta
function when $\alpha > 1$. This property will be explored further in Section 3; however, the general consequence of this convergence is the existence of an asymptotic bound on the amount of diversification that one can obtain. In other words, at a certain point, including an additional city in the portfolio does not provide any further benefit in regards to lowering portfolio risk.

To provide some additional intuition regarding the concentration parameter, $\alpha$, and the implied weights from the Zipf distribution, consider a scenario where $n = 400$ and $\alpha = \{0, 1, 2\}$. These cases of $\alpha$ respectively refer to the equal-weighted case, traditional rank-size rule, and a convergent case of the Zipf distribution.

In the equal-weighted case ($\alpha = 0$), the weight assigned to the top ranked site is simply $1/n$ or 0.25%, which is the same for all cities. Alternatively, with the traditional rank-size rule ($\alpha = 1$), the top ranked site receives a weight of 15.2%, the second rank is 7.6%, and by rank 60, the weight has decayed down to the equal-weighted level of 0.25%. In the highly concentrated, convergent case ($\alpha = 2$), the top ranked site has a 60.9% weight, the second rank has 15.2%, and the remaining 398 sites make up the remaining 23.9%.

If we examine the median ranks that indicate how many of the top sites are needed to account for 50% of the weight, the equal-weighted case simply yields 200.5, which suggests that the top 200 sites have half the weight and the bottom 200 have the remaining half. In the $\alpha = 1$ case, the larger concentration gives the top 15 sites a 50% share and the bottom 385 the other half. With $\alpha = 2$, over 50% weight is given to just the top ranked site and the remaining 399 produce less than half of the concentration (39.1%).
2.1 Estimation

For an empirical cross-section of data, estimation of the $\alpha$ parameter effectively produces a measurement of the degree of concentration in the variable. This is achieved by fitting the log-linear regression model in (6) where the ordinary least squares coefficient for $\beta_1$ is an estimate of $-\alpha$. The variable of interest, $x$, is ranked from largest to smallest and shifted by 1/2, $r = \text{rank} - 1/2$. This rank shift follows from Gabaix and Ibragimov (2011) to reduce small sample bias. Thus, in regards to city populations, the top rank is 0.5 (New York City), followed by 1.5 (Los Angeles), and so on.

$$\ln(x) = \beta_0 + \beta_1 \ln(r) + \varepsilon$$  \hspace{1cm} (6)

As an example of the fit provided by the Zipf distribution, Figure 1 depicts the fitted curve for the populations across the 50 largest CBSAs from the 2018 intercensal estimates from U.S. Census. With an adjusted $R^2$ of 0.953 and estimated slope of $-0.662$, this demonstrates a relatively close fit for a mild degree of concentration among the top 50 metropolitan areas. Since this estimate of $\alpha$ is less than one, there does not appear to be any major concerns regarding diversification within this subset of the top 50 CBSAs.

An observation that one may have regarding this linear fit is the autocorrelation in the residuals. Gabaix (2009) notes that this positive autocorrelation follows from the ranking procedure, and that as a result, the typical OLS standard errors are incorrect. To address this, standard errors that are presented throughout the paper are computed across 10,000 bootstrap iterations.
Figure 1: Fitted linear rank-size relation for the populations across the 50 largest CBSAs in the 2018 intercensal estimates from the U.S. Census. The estimated slope parameter suggests $\alpha = 0.662$ for the Zipf distribution in (4), and produces an adjusted $R^2$ of 0.953.
The estimation of the rank-size regression model has been studied for quite some time with a variety of approaches. For example, Nishiyama et al. (2008) recommend a trimmed OLS procedure, which removes an optimal number of the top ranked sites to reduce bias when testing for the traditional rank-size rule. However, rather than testing the hypothesis of $\alpha = 1$, this paper aims to simplify the concentration risk of mortgage debt and compare across different section, which would be incomplete without including the largest markets.

Another approach to the issue of autocorrelation and bias is to expand the functional form of the regression to allow for non-linearity in the fitted curve. The motivation for this is emphasized when expanding to the full set of 945 U.S. CBSAs, which are plotted along with their linear fit in Figure 2. From this figure, the non-linearity of this relationship is apparent with the top ranked metros being vastly overestimated by the linear approximation.

This result appears to contrast with the case of country-level populations, in which Laherrère and Sornette (1998) find that China and India appear as outliers while the remaining countries fit a straight line with an $R^2$ of 0.995. This phenomenon, which they term as a “king effect,” also appears in cases such as the populations of French cities, where Paris is the “king” or underestimated outlier in the rank-size regression. Unlike these cases, the scenario with U.S. city populations shows gradual overestimation when moving from the well-fitting mid-section of the curve to the top ranked cities. This pattern suggests that the linear relationship can be improved upon by introducing an additional parameter to correct for the non-linearity apparent in the empirical data.
Figure 2: Fitted linear rank-size relation for the populations across all 945 U.S. CBSAs in the 2018 intercensal estimates from the U.S. Census. The estimated slope parameter suggests $\alpha = 1.244$ for the Zipf distribution in (4), and produces an adjusted $R^2$ of 0.968.
2.2 Parabolic Fractal Distribution

One such non-linear model involves fitting the parabolic fractal distribution (Laherrère, 1996). This distribution expands the log-linear relationship from the Zipf distribution to include a quadratic term for ln(rank), as in (7).

\[
\ln(x) = \beta_0 + \beta_1 \ln(r) + \beta_2 \ln^2(r) + \varepsilon
\]  

(7)

This resembles the translog extension of the Cobb-Douglas production function by fitting a second-order polynomial instead of a straight line to the log-log relationship. With this additional term, the model corrects for the non-linearity that arises when examining a more complete set of locations. As seen in the fitted curve for 2018 populations in Figure 3, this parabolic fit greatly improves upon the linear fit of the Zipf distribution, particularly around the top ranked sites.

With this additional term, the regression $R^2$ improves to 0.995 from the 0.968 of the linear Zipf fit from Figure 2. However, one drawback of this expanded functional form is the difficulty in comparing the estimates across variables. For example, when comparing the estimates of Zipf’s $\alpha$ across the 2010 population counts and 2018 intercensal estimates from Table 1, the increase in magnitude of the slope estimate from 1.219 to 1.244 suggests that populations are becoming more concentrated in major metropolitan areas. On the other hand, the parabolic fractal estimates are less clear. The magnitude of the linear coefficient gets smaller (−0.174 to −0.155), and the quadratic coefficient becomes larger in magnitude (−0.107 to −0.112).
Figure 3: Fitted quadratic rank-size relation for the populations across all 945 U.S. CBSAs in the 2018 intercensal estimates from the U.S. Census. This functional form follows from the parabolic fractal distribution and produces an adjusted $R^2$ of 0.995.
<table>
<thead>
<tr>
<th></th>
<th>2010 Counts</th>
<th>2018 Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIPF</td>
<td>PFO</td>
</tr>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>18.626</td>
<td>15.071</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.155</td>
<td>0.028</td>
</tr>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>-1.219</td>
<td>-0.984</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.026</td>
<td>0.005</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>-</td>
<td>-0.162</td>
</tr>
<tr>
<td>s.e.</td>
<td>-</td>
<td>0.005</td>
</tr>
<tr>
<td>( \bar{R}^2 )</td>
<td>0.969</td>
<td>0.995</td>
</tr>
<tr>
<td>( n )</td>
<td>945</td>
<td>945</td>
</tr>
<tr>
<td>( p_{50} )</td>
<td>2.698</td>
<td>1.995</td>
</tr>
<tr>
<td>( p_{75} )</td>
<td>8.047</td>
<td>3.991</td>
</tr>
<tr>
<td>( p_{90} )</td>
<td>40.459</td>
<td>10.188</td>
</tr>
</tbody>
</table>

Table 1: Rank-size regressions for the population estimates across all 945 U.S. CBSAs from the 2010 census counts and 2018 intercensal estimates. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 2. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, \( p_{50}, p_{75}, \) and \( p_{90} \) refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.
One way to address the interpretation issue is to orthogonalize the quadratic term in (7) with the linear term. This is accomplished by regressing $\ln^2(r)$ on $\ln(r)$, as in (8), and using the residuals (9) in place of the quadratic term. At this point, the estimated slope coefficient $\hat{\beta}_1$ will be equal to that of the linear Zipf regression and $\hat{\beta}_2$ will capture the impact of the curvature provided by the quadratic component. To make comparisons regarding the relative effects of each of these components, all of the variables are normalized to have unit variance in the regression. The result, denoted by PFO, is given in (10) where $\sigma_x$, $\sigma_r$, and $\sigma_u$ are the standard deviations of $\ln(x)$, $\ln(r)$, and $\hat{u}$, respectively.

$$\ln^2(r) = \gamma_0 + \gamma_1 \ln(r) + u$$  \hspace{1cm} (8)

$$\hat{u} = \ln^2(r) - \hat{\gamma}_0 - \hat{\gamma}_1 \ln(r)$$  \hspace{1cm} (9)

$$\frac{\ln(x)}{\sigma_x} = \beta_0 + \beta_1 \frac{\ln(r)}{\sigma_r} + \beta_2 \frac{\hat{u}}{\sigma_u} + \varepsilon$$  \hspace{1cm} (10)

Since these transformations simply orthogonalize and scale the variables, this PFO regression results in the exact same set of predictions as the standard parabolic fractal regression. However, now the coefficients capture the relative impact of each of the independent components. For example, with the 2018 population estimates in Table 1, the linear coefficient of $-0.984$ is identical in effect to the Zipf slope coefficient of $-1.244$ after the scaling. However, now the $\hat{\beta}_2$ coefficient has some interpretive value.

Since each coefficient represents the standard deviation of the respective component in the fitted model, the coefficient of $-0.165$ relates to a variance of $0.027$ (or $0.165^2$). Meanwhile, the standard deviation of $0.984$ from the
linear component relates to a variance of 0.968, which is also the $R^2$ from the linear, Zipf model. Mathematically, since the dependent variable is also normalized to have unit variance, the sum of these variances, $0.968 + 0.027 = 0.995$, is equal to the PFO model $R^2$.

An additional consideration when fitting a parabola is the monotonicity implied by the ranking procedure. For the Zipf distribution, the slope parameter is guaranteed to be non-positive.\(^2\) However, with the parabolic fit, it is possible that the estimates produce a parabola that is increasing for some portion of the domain. This is resolved by imposing two constraints, (12) and (13), which require that the first derivative, (11), be non-positive at both bounds of the domain, $r \in [0.5, n - 0.5]$. Since the derivative is linear in $\ln$ (rank), this forces the fitted curve to be non-increasing over the entire domain.

\[
d(r) = \frac{d \ln(x)}{d \ln(r)} = \beta_1 + 2\beta_2 \ln(r) \tag{11}
\]

\[
d(0.5) \leq 0 \implies \beta_1 \leq -2\beta_2 \ln(0.5) \tag{12}
\]

\[
d(n - 0.5) \leq 0 \implies \beta_2 \leq \frac{-\beta_1}{2\ln(n - 0.5)} \tag{13}
\]

For the orthogonalized variant of the parabolic fractal distribution, the two monotonicity constraints are similar; however, due to the orthogonalization and scaling, a few minor tweaks must be made. The details for the transformed constraints are included in Appendix A.

\(^2\)A slope of zero is possible in the equal-weighted case; however, any variation in size will result in a negative slope due to the ranking procedure.
3 Implications for Portfolio Risk

The analysis of risk is an important consideration for both portfolio selection and optimization. In the traditional mean-variance setup of Markowitz (1952), risk is proxied with the variance of the portfolio returns as in (14). The components of this are the portfolio weights, $w$, and the covariance matrix for asset returns, $\Sigma$.

$$\sigma_p^2 = w'\Sigma w$$

(14)

In regards to geographical concentration, the focus is on modeling the distribution of the $w$ component of this equation, given the return distribution characterized by $\Sigma$. The rank-size relations described in the previous section provide a way to simplify this concentration into just a few estimated parameters. To the degree that a portfolio follows similar weights, the fitted relations can be scaled to produce implied portfolio weights. Alternatively, these implied weights can be thought of as the proportional importance of each CBSA to the national market. This section demonstrates how to obtain the implied weights from the fitted rank-size regressions and the implications for the ability to diversify risk when the weights are distributed as such.

3.1 Implied Weights

Similar to the derivation of the Zipf distribution in Section 2, the fitted sizes from the estimated rank-size regressions must be scaled by their sum to produce the implied weights. From the fitted values of each regression model, the logarithmic functional form typically requires an adjustment to account
for the log-normality of the transformed residuals. Using the Zipf distribution as an example, the natural exponential of (6) produces (15), which simplifies to (16) following from the properties of exponentials.

\[ x = \exp(\beta_0 + \beta_1 \ln(r) + \varepsilon) \]  
\[ = \exp(\beta_0) \cdot r^{\beta_1} \cdot \exp(\varepsilon) \]  

(15)  

(16)

Thus, when taking the expectation of \( x \) given \( \beta \), the first two terms are fixed and factor out, as in (17). However, if \( \varepsilon \) is normally distributed, then \( \mathbb{E}[\exp(\varepsilon)] = \exp(\sigma^2/2) \), and scales the expected size from the regression.

\[ \mathbb{E}[x|\beta] = \exp(\beta_0) \cdot r^{\beta_1} \cdot \mathbb{E}[\exp(\varepsilon)] \]  

(17)

Although this is important for calculating the expected values for the size of each CBSA, this scaling is equal for each site. As a result, its effect is offset when dividing by the sum to obtain implied weights, as in (18)–(20).

\[ w = \frac{\mathbb{E}[x|\beta]}{\sum \mathbb{E}[x|\beta]} \]  
\[ = \frac{\exp(\beta_0) \cdot r^{\beta_1} \cdot \mathbb{E}[\exp(\varepsilon)]}{\sum_{r=1}^{n} \exp(\beta_0) \cdot r^{\beta_1} \cdot \mathbb{E}[\exp(\varepsilon)]} \]  
\[ = \frac{r^{\beta_1}}{\sum_{r=1}^{n} r^{\beta_1}} \]  

(18)  

(19)  

(20)

These implied weights in (20) are identical to the probability mass function in (4) defining the Zipf distribution. This same process can be repeated for the implied weights from the parabolic fractal distribution; however, the solution does not simplify as neatly and is left for Appendix B.
3.2 Portfolio Variance

With these implied weights from the rank-size regressions, these simplifications can be substituted into the portfolio variance equation in (14). In this framework, the implications of the asset concentration can be examined and compared across the two rank-size models. As a way to compare the fit in terms of portfolio risk, these can also be compared with the empirical weights given by the original data. As will be demonstrated, the parabolic fractal distribution outperforms the linear Zipf regression.

To begin, consider the simple case of independent and identically distributed returns. In this case, the covariance matrix is \( \sigma^2 I_n \) where \( I_n \) is an \( n \) by \( n \) identity matrix with ones on the diagonal and zeros for the off-diagonal elements. This leads to a portfolio variance equal to the individual asset variance multiplied by the sum of squared weights, as in (23).

\[
\sigma_{iid}^2 = w' \sigma^2 I w \\
= \sigma^2 w' w \\
= \sigma^2 (w_1^2 + w_2^2 \ldots w_n^2)
\]

For the case of Zipf-distributed weights, the sum of squared weights can be simplified using the harmonic number notation from (5). For each asset \( i \), the implied weight can be written as (24) and its squared value in (25).

\[
w(n, \alpha)_i = \frac{i^{-\alpha}}{H_{n,\alpha}} \\
w(n, \alpha)^2_i = \frac{i^{-2\alpha}}{H_{n,\alpha}^2}
\]
Substituting this into the i.i.d. portfolio variance from (23), the numerator in (26) is simply another harmonic number, which yields the solution in (27).

\[
\sigma_{iid}^2 = \sigma^2 \sum_{i=1}^{n} \frac{i^{-2\alpha}}{H_{n,\alpha}^2} \quad (26)
\]

\[
= \sigma^2 \frac{H_{n,2\alpha}}{H_{n,\alpha}^2} \quad (27)
\]

In this scenario, the asymptotic properties of harmonic sums demonstrates a potential limiting effect in regards to diversification. As noted back in Section 2, when the \(\alpha\) parameter is greater than one, the harmonic sum converges as \(n \to \infty\). This implies that the denominator in (27) approaches a finite value. Similarly, the numerator also converges and the i.i.d. portfolio variance has an asymptotic bound greater than zero. This suggests a limit to the amount of diversification that one can obtain.

On the other hand, if the concentration is low (\(\alpha \leq 1\)), the harmonic sum diverges and increases to infinity along with \(n\). Since the denominator continues to increase, the i.i.d. portfolio variance will approach zero for sufficiently large \(n\). This rate of decline is faster for less concentrated asset classes.

For example, Figure 4 demonstrates the decline in the i.i.d. portfolio variance as \(n\) increases with different concentrations. The equal-weighted portfolio (\(\alpha = 0\)) diversifies at a rate of \(1/n\). The traditional rank-size rule of \(\alpha = 1\) declines at a slower rate, but still approaches zero as \(n\) increases. The convergent case of \(\alpha = 2\) exhibits a much slower rate of decline as well as an asymptotic lower bound of 0.4, which demonstrates the limit on diversification that can be imposed by a highly concentrated market.
Figure 4: Portfolio variance across i.i.d. assets with unit variances and Zipf distributed weights.
Building off of this idea of an asymptotic limit to the level of diversification that one can obtain, another way to compare the implied weights from the rank-size regressions is to examine the number of sites required to reach a certain percentage of the potential diversification. For example, since the $\alpha = 2$ case has a lower bound of 0.4 for the i.i.d. variance, 50% diversification would be obtained at the rank where the i.i.d. variance is equal to 0.7.

This involves re-weighting each subset of weights such that they sum one. Using the earlier example of $n = 400$ and $\alpha = 2$, a portfolio with only the top ranked site will re-weight the 0.609 to 1, which produces a variance of 1. Similarly, for a portfolio of the top two sites, the weights of 0.609 and 0.152 re-weight to 0.8 and 0.2. This produces a variance of $0.8^2 + 0.2^2 = 0.68$. Thus, more than 50% of the potential risk reduction is obtained just by including the second ranked site in the portfolio. This rank for this 50% reduction in risk is denoted as $p_{50}$ and is presented along with $p_{75}$ and $p_{90}$ in the bottom three rows of each table.

As a visual aid to understanding these percentile ranks, Figure 5 presents a comparison between the linear Zipf weights, the parabolic PFO weights, and the empirical weights. For each curve, the $p_{50}$ rank indicates the number of cities required to lower the i.i.d. portfolio variance 50% of the way to its fully diversified variance. Unlike the theoretical case where $n \rightarrow \infty$, this fully diversified variance is simply based off of the full weight vector. As can be seen in the figure, the Zipf distribution’s overestimation of the top ranked sites tends to overestimate the limiting factor; however, the PFO model greatly improves on this and traces the empirical weights quite well.
Figure 5: Comparison between linear Zipf weights, PFO weights, and empirical weights for i.i.d. portfolio variances as portfolio size increases.
The correlation across asset returns is another critical component for portfolio variance and the ability to diversify risk. For mortgages, this component can be represented as a function of the correlation across the underlying property returns and the default rate of the loans (Dombrowski et al., 2019). Real estate assets demonstrate high levels of both spatial (weak) and macroeconomic (strong) cross-sectional dependence. In regards to geographical diversification, risk can be reduced by diversifying away the weak dependence; however, the strong dependence remains regardless of portfolio size.

To examine the effect of strong cross-sectional dependence in this context, a correlation of \( \rho \) is introduced to the i.i.d. returns and a unit variance is assumed, as in (28). This can be rewritten into matrix form as in (29), where \( I_n \) is an \( n \) by \( n \) identity matrix and \( \iota_n \) is a vector of ones.

\[
V_{ij} = \begin{cases} 
1, & \text{if } i = j \\
\rho, & \text{if } i \neq j 
\end{cases} 
\]  

(28)

\[
V = (1 - \rho)I_n + \rho \iota_n \iota_n' 
\]  

(29)

Substituting this correlation matrix into (14), (30) simplifies to (31).

\[
\sigma_p^2 = w'Vw 
\]  

(30)

\[
= (1 - \rho)w'w + \rho w' \iota_n \iota_n' w 
\]  

(31)

Then (32) follows since both \( w' \iota_n \) and \( \iota_n' w \) are equal to one since the weight vectors sum to one. Thus, (33) provides the portfolio variance as a function of the macroeconomic risk (\( \rho \)) and i.i.d. portfolio variance described previously.
\begin{align*}
  &= (1 - \rho)w'w + \rho \\
  &= (1 - \rho)\sigma_{iid}^2 + \rho
\end{align*}

The effect of this correlation on portfolio variance is a reduction in the potential for diversification. This subsequently lowers the impact of any concentration risk coming from the portfolio weights. For example, if \( \rho = 0.3 \), this portion of the variance is not diversifiable. The sum of squared weights (or \( \sigma_{iid}^2 \)) can still limit the ability to achieve the potential diversification; however, this term is scaled by \( (1 - \rho) \) or 0.7. More strong dependence (ex. specialized portfolios) diminishes the diversification potential in general and lowers the impact of high concentrations. However, as will be demonstrated in the next section, some sectors of the mortgage market (such as jumbo mortgages) exhibit substantially larger degrees of concentration, which may offset the attenuation from large correlations.

4 Empirical Concentration

Now that the previous examples for the rank-size relations have demonstrated the concentration of population across U.S. CBSAs, the focus shifts to the mortgage market. Using a large set of loan-level mortgage data from BKFS, this section models the geographical concentration of various sectors by partitioning the data across a number of dimensions, including GSE-eligibility, lien priority, interest rate type, loan purpose, occupancy status, and docu-
mentation level.

The raw BKFS dataset includes a loan table, which provides origination characteristics for more than 173 million loans. Additionally, monthly remittance tables begin in January 1989 and provide updates on loan statuses, balances, and interest rates. After a relatively mild cleaning process,\textsuperscript{3} we are left with just over 150 million loans originating between January 1990 and November 2016, which is the most recent month of observation.

4.1 Results

In the context of financial variables such as mortgage debt, the concentration of population only reflects part of the equation. In addition to large population centers originating more loans, there is also a pricing differential between major cities and smaller, less urban locations. This suggests that while measurements of population concentration may act as a reasonable baseline for the quantity of loans, the value of the debt is likely to be even more highly concentrated.

This distinction between concentration in quantity vs. concentration of balances is demonstrated in Table 2. This table presents the rank-size estimates for both the total number of loans originated over the sample period along with the estimates for the concentration of the aggregate balances. From the Zipf slope estimates, the increase from $\hat{\alpha} = 1.56$ for quantities to $\hat{\alpha} = 1.73$ for balances shows a materially larger degree of concentration for the aggregate debt compared to simply the number of loans. When compared

\textsuperscript{3}See Appendix C for a detailed account of the data cleaning process.
to the 2018 population estimate of $\hat{\alpha} = 1.24$, both mortgage quantities and balances appear to be far more heavily concentrated than is suggested by measures of population concentration. Thus, the geographical concentration of mortgage debt is larger than that of the number of loans, which are both larger than the concentration present in populations across the country.

In regards to the orthogonalized parabolic fractal estimates, the additional fit provided by the curvature explains a similar proportion of the loan quantity concentration ($0.966/0.163 = 5.93$) as with the aggregate debt balances ($0.968/0.163 = 5.94$). This suggests that the comparison of the linear components captures roughly the same variation in the rank-size relationship.

For the diversification percentile ranks, the Zipf estimates suggest that 90% of the potential diversification is obtained by diversifying across the top 21 ranked cities. However, since the overestimation of top ranked sites leads to an overestimated lower bound for the diversification potential (as in Figure 5), these ranks are less reliable than those for the PFO model, which more closely tracks the true empirical weights.

For the PFO estimates, the percentile ranks suggest that roughly 50% of the potential diversification is obtained just by including the two largest cities in the portfolio. To attain 75% of the diversification potential, $p_{75}$ suggests this is accomplished with the top four cities. Then for 90% of the diversification potential, the top ten cities reduce the i.i.d. portfolio variance 90% of the way to its fully diversified lower bound.
Table 2: All Mortgage Originations

<table>
<thead>
<tr>
<th></th>
<th>Loan Quantities</th>
<th>Balances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIPF</td>
<td>PFO</td>
</tr>
<tr>
<td>$\hat{\beta}_0$</td>
<td>19.375</td>
<td>11.992</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.270</td>
<td>0.091</td>
</tr>
<tr>
<td>$\hat{\beta}_1$</td>
<td>-1.555</td>
<td>-0.966</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.045</td>
<td>0.016</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>-</td>
<td>-0.163</td>
</tr>
<tr>
<td>s.e.</td>
<td>-</td>
<td>0.024</td>
</tr>
<tr>
<td>$\bar{R}^2$</td>
<td>0.912</td>
<td>0.960</td>
</tr>
<tr>
<td>n</td>
<td>929</td>
<td>929</td>
</tr>
<tr>
<td>$p_{50}$</td>
<td>2.353</td>
<td>1.985</td>
</tr>
<tr>
<td>$p_{75}$</td>
<td>5.690</td>
<td>3.916</td>
</tr>
<tr>
<td>$p_{90}$</td>
<td>20.825</td>
<td>9.705</td>
</tr>
</tbody>
</table>

Table 2: Rank-size regressions for the total quantities and balances of mortgage originations across 929 U.S. CBSAs over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 2. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, $p_{50}$, $p_{75}$, and $p_{90}$ refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.
To narrow the focus and provide some comparisons across different segments of the mortgage market, Tables 3–9 present the rank-size estimates for various subsets of loans. The first partition that is examined is the conforming vs. jumbo loan markets. One characteristic of the mortgage market that distinguishes it from other debt markets is the prominence of the GSEs, which stimulate demand in the secondary mortgage market by purchasing and securitizing any loans that meet their criteria. One such criterion is the loan amount, which is set by the Federal Housing Finance Agency. Mortgages that fall below this threshold are classified as conforming loans and are eligible for purchase by the GSEs. Loans with balances above this limit are classified jumbo loans and are often held as portfolio loans on bank balance sheets or packaged into private-label mortgage-backed securities (MBS).

Table 3 examines the non-conforming, jumbo loan market, which suggests drastically larger degrees of concentration (\( \hat{\alpha} = 2.53 \) for loan quantities and 2.58 for aggregate debt). On the other hand, conforming loans (Table 4) are far less concentrated with \( \hat{\alpha} = 1.54 \) and 1.69, respectively for quantities and balances. Since the less concentrated, conforming loans are purchased and securitized by the GSEs, this would suggest that the loans held in bank portfolios or in private-label MBS tend to be more highly concentrated than those that are in the GSE securities.

In regards to the diversification percentile ranks, the conforming loan market estimates appear fairly similar to those for the full mortgage market with \( p_{50}, p_{75}, \) and \( p_{90} \) respectively equal to 2, 4, and 10. However, for the jumbo market, \( p_{75} \) reduces to 3.5 and \( p_{90} \) falls to roughly 7.7.
Table 3: Jumbo Mortgages

<table>
<thead>
<tr>
<th></th>
<th>Loan Quantities</th>
<th>Balances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIPF</td>
<td>PFO</td>
</tr>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>19.954</td>
<td>7.598</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.461</td>
<td>0.150</td>
</tr>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>-2.534</td>
<td>-0.967</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.077</td>
<td>0.026</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>-</td>
<td>-0.163</td>
</tr>
<tr>
<td>s.e.</td>
<td>-</td>
<td>0.040</td>
</tr>
<tr>
<td>( \bar{R}^2 )</td>
<td>0.901</td>
<td>0.962</td>
</tr>
<tr>
<td>( n )</td>
<td>913</td>
<td>913</td>
</tr>
<tr>
<td>( p_{50} )</td>
<td>1.770</td>
<td>1.939</td>
</tr>
<tr>
<td>( p_{75} )</td>
<td>2.621</td>
<td>3.564</td>
</tr>
<tr>
<td>( p_{90} )</td>
<td>4.595</td>
<td>7.724</td>
</tr>
</tbody>
</table>

Table 3: Rank-size regressions for the total quantities and balances of jumbo mortgage originations across U.S. CBSAs over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 2. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, \( p_{50} \), \( p_{75} \), and \( p_{90} \) refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.
Table 4: Conforming Mortgages

<table>
<thead>
<tr>
<th>Loan Quantities</th>
<th>Balances</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZIPF</td>
<td>PFO</td>
</tr>
<tr>
<td>$\hat{\beta}_0$</td>
<td>19.276</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.272</td>
</tr>
<tr>
<td>$\hat{\beta}_1$</td>
<td>-1.541</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.046</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>-</td>
</tr>
<tr>
<td>s.e.</td>
<td>-</td>
</tr>
<tr>
<td>$\bar{R}^2$</td>
<td>0.909</td>
</tr>
<tr>
<td>n</td>
<td>929</td>
</tr>
<tr>
<td>$p_{50}$</td>
<td>2.372</td>
</tr>
<tr>
<td>$p_{75}$</td>
<td>5.783</td>
</tr>
<tr>
<td>$p_{90}$</td>
<td>21.544</td>
</tr>
</tbody>
</table>

Table 4: Rank-size regressions for the total quantities and balances of conforming mortgage originations across U.S. CBSAs over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 2. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, $p_{50}$, $p_{75}$, and $p_{90}$ refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.
Another separation for the mortgage market is the lien priority. In event of foreclosure, the priority of the mortgage lien indicates the riskiness for the debtholder to recover some of the losses. After foreclosure sales, which tend to be at a substantial discount (Clauretie and Daneshvary, 2009), first-lien debtholders are paid down prior to any recovery for junior liens.

In Table 5, the left columns model the concentration of first-lien mortgages. These are the vast majority of the loans with roughly 96% of all loans and 99% of the aggregate debt. The first-lien debt demonstrates similar results to the full set of mortgages in Table 2. This is contrasted with junior-lien mortgages (right columns of Table 5) that have only a residual claim to recovery in event of a default. The concentration estimate of 1.83 suggests a mildly larger degree of concentration for this smaller market; however, this does not appear to be statistically significant given the standard errors.

As with the previous results for the PFO model, both the first-lien and junior-lien partitions exhibit similar relative effects between the linear portion of the fit and the correction for the non-linearity. For the first-lien subset, the relative effect is nearly identical to the full market (0.968/0.163 = 5.94). Then for the smaller junior-lien market, the explanatory power is slightly lower; however, the relative effect of the two components (0.958/0.161 = 5.95) is still fairly constant.

In regards to the percentile ranks, the results are both quantitatively similar to those for the full mortgage market. The number of cities needed to obtain 50%, 75%, and 90% shares of the diversification potential remain in a similar range around 2, 4, and 9, respectively in the PFO model.
<table>
<thead>
<tr>
<th></th>
<th>First-Lien</th>
<th></th>
<th>Junior-Lien</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIPF</td>
<td>PFO</td>
<td>ZIPF</td>
<td>PFO</td>
</tr>
<tr>
<td>(\hat{\beta}_0)</td>
<td>32.075</td>
<td>17.805</td>
<td>27.999</td>
<td>14.609</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.288</td>
<td>0.101</td>
<td>0.313</td>
<td>0.117</td>
</tr>
<tr>
<td>(\hat{\beta}_1)</td>
<td>-1.733</td>
<td>-0.968</td>
<td>-1.825</td>
<td>-0.958</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.048</td>
<td>0.018</td>
<td>0.053</td>
<td>0.021</td>
</tr>
<tr>
<td>(\hat{\beta}_2)</td>
<td>-</td>
<td>-0.163</td>
<td>-</td>
<td>-0.161</td>
</tr>
<tr>
<td>s.e.</td>
<td>-</td>
<td>0.027</td>
<td>-</td>
<td>0.028</td>
</tr>
<tr>
<td>(\bar{R}^2)</td>
<td>0.914</td>
<td>0.963</td>
<td>0.893</td>
<td>0.943</td>
</tr>
<tr>
<td>n</td>
<td>929</td>
<td>929</td>
<td>929</td>
<td>929</td>
</tr>
<tr>
<td>(p_{50})</td>
<td>2.172</td>
<td>1.978</td>
<td>1.997</td>
<td>1.974</td>
</tr>
<tr>
<td>(p_{75})</td>
<td>4.830</td>
<td>3.869</td>
<td>4.170</td>
<td>3.840</td>
</tr>
<tr>
<td>(p_{90})</td>
<td>15.186</td>
<td>9.399</td>
<td>11.466</td>
<td>9.211</td>
</tr>
</tbody>
</table>

Table 5: Rank-size regressions for the balances of mortgage originations partitioned by lien priority over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 2. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, \(p_{50}\), \(p_{75}\), and \(p_{90}\) refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.
The type of interest rate for a mortgage is another loan characteristic that appears to suggest some variation around the degree of geographical concentration. In Table 6, the estimated concentration for adjustable rate mortgages ($\alpha = 1.82$) appears to be slightly larger than that of the fixed rate mortgage market ($\alpha = 1.69$). Similar to previous results, the PFO model suggests consistent improvements to the linear fit of the Zipf distribution and the estimated percentile ranks are also unchanged.

When comparing mortgages that are originated for new purchases and refinancing activity (Table 7), both subsets suggest similar degrees of geographical concentration, which are also similar to the full mortgage market.

In Table 8, the concentration of investment properties ($\alpha = 2.08$) appears to be larger than for owner-occupied properties ($\alpha = 1.71$). The estimates for the investment property subset do tend to carry slightly less explanatory power in the rank-size regressions; however, as with the junior-lien market, if the relative effect of the two components are compared (0.955/0.161), then the proportional importance of each component is shown to be constant.

Lastly, the geographical concentration between loans with full documentation are compared to those with less than full documentation. In Table 9, the estimates for the fully documented mortgage market suggest only slightly less concentration than the non-fully documented loans. On the other dimensions of comparison, such as the relative effects in the PFO model and the percentile ranks, these estimates also produce fairly similar results.
Table 6: Mortgage Debt by Rate Type

<table>
<thead>
<tr>
<th></th>
<th>Fixed Rate</th>
<th>Adjustable Rate</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIPF</td>
<td>PFO</td>
<td>ZIPF</td>
<td>PFO</td>
</tr>
<tr>
<td>$\hat{\beta}_0$</td>
<td>31.699</td>
<td>18.037</td>
<td>31.672</td>
<td>14.783</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.284</td>
<td>0.099</td>
<td>0.350</td>
<td>0.117</td>
</tr>
<tr>
<td>$\hat{\beta}_1$</td>
<td>-1.688</td>
<td>-0.967</td>
<td>-2.072</td>
<td>-0.972</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.048</td>
<td>0.018</td>
<td>0.059</td>
<td>0.020</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>-</td>
<td>-0.163</td>
<td>-</td>
<td>-0.164</td>
</tr>
<tr>
<td>s.e.</td>
<td>-</td>
<td>0.026</td>
<td>-</td>
<td>0.032</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.911</td>
<td>0.961</td>
<td>0.922</td>
<td>0.972</td>
</tr>
<tr>
<td>n</td>
<td>929</td>
<td>929</td>
<td>927</td>
<td>927</td>
</tr>
<tr>
<td>$p_{50}$</td>
<td>2.172</td>
<td>1.980</td>
<td>1.899</td>
<td>1.963</td>
</tr>
<tr>
<td>$p_{75}$</td>
<td>4.830</td>
<td>3.881</td>
<td>3.369</td>
<td>3.756</td>
</tr>
<tr>
<td>$p_{90}$</td>
<td>15.186</td>
<td>9.481</td>
<td>7.512</td>
<td>8.713</td>
</tr>
</tbody>
</table>

Table 6: Rank-size regressions for the balances of mortgage originations partitioned by interest rate type over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 2. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, $p_{50}$, $p_{75}$, and $p_{90}$ refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.
Table 7: Mortgage Debt by Loan Purpose

<table>
<thead>
<tr>
<th></th>
<th>New Purchases</th>
<th>Refinances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIPF</td>
<td>PFO</td>
</tr>
<tr>
<td>$\hat{\beta}_0$</td>
<td>31.074</td>
<td>17.372</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.297</td>
<td>0.098</td>
</tr>
<tr>
<td>$\hat{\beta}_1$</td>
<td>-1.724</td>
<td>-0.969</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.050</td>
<td>0.017</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>–</td>
<td>-0.163</td>
</tr>
<tr>
<td>s.e.</td>
<td>–</td>
<td>0.027</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.916</td>
<td>0.966</td>
</tr>
<tr>
<td>$n$</td>
<td>929</td>
<td>929</td>
</tr>
<tr>
<td>$p_{50}$</td>
<td>2.124</td>
<td>1.979</td>
</tr>
<tr>
<td>$p_{75}$</td>
<td>4.654</td>
<td>3.871</td>
</tr>
</tbody>
</table>

Table 7: Rank-size regressions for the balances of mortgage originations partitioned by loan purpose over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 2. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, $p_{50}$, $p_{75}$, and $p_{90}$ refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.
Table 8: Mortgage Debt by Occupancy

<table>
<thead>
<tr>
<th></th>
<th>Owner-Occupied</th>
<th></th>
<th>Investment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIPF</td>
<td>PFO</td>
<td>ZIPF</td>
<td>PFO</td>
</tr>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>31.577</td>
<td>17.798</td>
<td>28.505</td>
<td>12.915</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.276</td>
<td>0.101</td>
<td>0.414</td>
<td>0.150</td>
</tr>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>-1.713</td>
<td>-0.970</td>
<td>-2.084</td>
<td>-0.955</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.046</td>
<td>0.018</td>
<td>0.069</td>
<td>0.026</td>
</tr>
<tr>
<td>( \hat{\beta}_2 )</td>
<td>-</td>
<td>-0.163</td>
<td>-</td>
<td>-0.161</td>
</tr>
<tr>
<td>s.e.</td>
<td>-</td>
<td>0.027</td>
<td>-</td>
<td>0.033</td>
</tr>
<tr>
<td>( \bar{R}^2 )</td>
<td>0.921</td>
<td>0.967</td>
<td>0.868</td>
<td>0.938</td>
</tr>
<tr>
<td>n</td>
<td>929</td>
<td>929</td>
<td>928</td>
<td>928</td>
</tr>
<tr>
<td>( p_{50} )</td>
<td>2.139</td>
<td>1.979</td>
<td>1.895</td>
<td>1.962</td>
</tr>
<tr>
<td>( p_{75} )</td>
<td>4.708</td>
<td>3.875</td>
<td>3.337</td>
<td>3.743</td>
</tr>
<tr>
<td>( p_{90} )</td>
<td>14.377</td>
<td>9.443</td>
<td>7.388</td>
<td>8.641</td>
</tr>
</tbody>
</table>

Table 8: Rank-size regressions for the balances of mortgage originations partitioned by occupancy status over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 2. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, \( p_{50}, p_{75}, \) and \( p_{90} \) refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.
Table 9: Mortgage Debt by Documentation

<table>
<thead>
<tr>
<th></th>
<th>Full Doc</th>
<th></th>
<th>Non-Full Doc</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZIPF</td>
<td>PFO</td>
<td>ZIPF</td>
<td>PFO</td>
</tr>
<tr>
<td>$\hat{\beta}_0$</td>
<td>30.736</td>
<td>17.258</td>
<td>31.821</td>
<td>17.487</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.283</td>
<td>0.108</td>
<td>0.294</td>
<td>0.101</td>
</tr>
<tr>
<td>$\hat{\beta}_1$</td>
<td>-1.702</td>
<td>-0.961</td>
<td>-1.753</td>
<td>-0.969</td>
</tr>
<tr>
<td>s.e.</td>
<td>0.048</td>
<td>0.020</td>
<td>0.049</td>
<td>0.018</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>-</td>
<td>-0.162</td>
<td>-</td>
<td>-0.163</td>
</tr>
<tr>
<td>s.e.</td>
<td>-</td>
<td>0.026</td>
<td>-</td>
<td>0.027</td>
</tr>
<tr>
<td>$\bar{R}^2$</td>
<td>0.901</td>
<td>0.950</td>
<td>0.916</td>
<td>0.966</td>
</tr>
<tr>
<td>$n$</td>
<td>928</td>
<td>928</td>
<td>929</td>
<td>929</td>
</tr>
<tr>
<td>$p_{50}$</td>
<td>2.154</td>
<td>1.979</td>
<td>2.083</td>
<td>1.977</td>
</tr>
<tr>
<td>$p_{75}$</td>
<td>4.763</td>
<td>3.878</td>
<td>4.506</td>
<td>3.863</td>
</tr>
</tbody>
</table>

Table 9: Rank-size regressions for the balances of mortgage originations partitioned by documentation status over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 2. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, $p_{50}$, $p_{75}$, and $p_{90}$ refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.
5 Conclusion

This paper characterizes the geographical concentration of the mortgage market using the empirical regularity from regional science known as rank-size rule. This allows for the simplification of a set of portfolio weights into just one or two parameters that measure the degree of concentration. Rank-size distributions such as the Zipf or parabolic fractal distribution demonstrate how high levels of concentration can impose limits on the ability to achieve geographical diversification.

The linear relationship between ln(size) and ln(rank) suggested by the Zipf distribution is extended to allow for non-linearity by fitting the parabolic fractal distribution along with an orthogonalized variant, which isolates the effect of the curvature. This extension helps correct for the overestimation of the top ranked cities evident from the linear fit of the Zipf distribution and provides more accurate predictions for the degree of concentration in mortgage debt.

The application of these rank-size relations to empirical mortgage data from BKFS suggests considerable degrees of concentration. When compared to the estimate of Zipf’s $\alpha$ for populations (1.24 in 2018), the quantity of mortgage originations produces a materially larger estimate of 1.56. Taking into account the higher property values of these large metros, this concentration estimate increases to 1.73 for the aggregate balances of the debt originations.

This dataset also allows for examination of specific sectors of the mortgage market. One such sector that exhibits substantially higher concentration is the jumbo market. With $\alpha$ estimates of roughly 2.5, this suggests that these
loans tend to be heavily concentrated in just a few large markets. Since these loans are not eligible for purchase by the government sponsored entities, this may lead to increased risk among mortgage portfolios held by banks.

In addition to the jumbo mortgage market, several other sectors of the mortgage market exhibit relatively large degrees of geographical concentration, albeit to a lesser degree. For example, mortgages secured by investment properties are substantially more concentrated when compared to owner-occupied properties. Similarly, junior-lien mortgages are slightly more concentrated than their first-lien counterparts, and adjustable rate mortgages show slightly more concentration than fixed rate mortgages.

With these large degrees of geographical concentration in mortgage debt, portfolios constructed from these assets are likely to experience limits to the amount of diversification that can be obtained. The Zipf distribution and the convergence of the generalized harmonic sum allow for a rigorous demonstration for how such limits to geographical diversification can lead to a lower bound on portfolio risk. As a result, the relative weights for various cities demonstrates how local economic shocks for some cities remain local, but others can propagate through to global shocks for highly concentrated markets. For example, a local shock to California is effectively a global shock to the jumbo mortgage market.
References


A Orthogonalized PF Constraints

Similar to the two monotonicity constraints for the standard parabolic fractal given by (12) and (13) in Section 2.2, the quadratic functional form requires two constraints to force the fitted curve to be non-increasing. To start, substitute the definition of orthogonalized term from (9) into the regression equation from (10).

\[
\frac{\ln(x)}{\sigma_x} = \beta_0 + \frac{\beta_1}{\sigma_r} \ln(r) + \frac{\beta_2}{\sigma_r} \frac{\ln^2(r) - \hat{\gamma}_0 - \hat{\gamma}_1 \ln(r)}{\sigma_e} + \varepsilon \tag{34}
\]

As with the standard parabolic fractal regression, the constraints require the first derivative (35) to be non-positive over the relevant domain.

\[
d(r) = \frac{d \ln(x)/\sigma_x}{d \ln(r)} = \frac{\beta_1}{\sigma_r} + \frac{\beta_2}{\sigma_e} \frac{2 \ln(r) - \hat{\gamma}_1}{\sigma_e} \tag{35}
\]

Another result from the normalization of \(\ln(r)\) is that the domain shifts as well. Thus, rather than being defined over \(r \in [0.5, n - 0.5]\), the domain is now from \(\ln(0.5)/\sigma_r\) to \(\ln(n - 0.5)/\sigma_r\). This leads to the two following monotonicity constraints:

\[
\frac{\beta_1}{\sigma_r} + \frac{\beta_2}{\sigma_e} \frac{2 \ln(0.5)/\sigma_r - \hat{\gamma}_1}{\sigma_e} \leq 0 \tag{36}
\]

\[
\frac{\beta_1}{\sigma_r} + \frac{\beta_2}{\sigma_e} \frac{2 \ln(n - 0.5)/\sigma_r - \hat{\gamma}_1}{\sigma_e} \leq 0 \tag{37}
\]
B  Implied Weights from PFO Regression

Following the scaling of the fitted, linear curve for the Zipf distribution to obtain implied portfolio weights in Section 3.1, this appendix repeats the process to derive the implied weights from the orthogonalized parabolic fractal regression.

After orthogonalizing and scaling the variables to estimate the fit for the parabolic fractal distribution, the resulting regression from (10) is repeated in (38).

\[
\frac{\ln(x)}{\sigma_x} = \beta_0 + \beta_1 \frac{\ln(r)}{\sigma_r} + \beta_2 \frac{\hat{u}}{\sigma_u} + \varepsilon
\]  

(38)

The first step involves multiplying both sides by \(\sigma_x\).

\[
\ln(x) = \sigma_x \beta_0 + \frac{\sigma_x \beta_1}{\sigma_r} \ln(r) + \frac{\sigma_x \beta_2}{\sigma_u} \hat{u} + \sigma_x \varepsilon
\]  

(39)

After taking the exponential of both sides, we obtain (40), which simplifies to (41) using the properties of exponentials.

\[
x = \exp \left(\sigma_x \beta_0 + \frac{\sigma_x \beta_1}{\sigma_r} \ln(r) + \frac{\sigma_x \beta_2}{\sigma_u} \hat{u} + \sigma_x \varepsilon\right)
\]  

(40)

\[
x = \exp(\sigma_x \beta_0) \cdot r^{\frac{\sigma_x \beta_1}{\sigma_r}} \cdot \exp \left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right) \cdot \exp(\sigma_x \varepsilon)
\]  

(41)

As with the Zipf distribution, the fitted values are scaled by their sum to obtain the implied probability mass function (or weights). The first and last terms are both constants, and thus cancel out, which simplifies (42) to (43).

\[
w = \frac{\exp(\sigma_x \beta_0) \cdot r^{\frac{\sigma_x \beta_1}{\sigma_r}} \cdot \exp \left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right) \cdot \exp(\sigma_x \varepsilon)}{\sum_{r=1}^{n} r^{\sigma_x \beta_1/\sigma_r} \cdot \exp \left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right) \cdot \exp(\sigma_x \varepsilon)}
\]  

(42)

\[
w = \frac{r^{\sigma_x \beta_1/\sigma_r} \cdot \exp \left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right)}{\sum_{r=1}^{n} r^{\sigma_x \beta_1/\sigma_r} \cdot \exp \left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right)}
\]  

(43)
C Data Cleaning

The full loan table from the Black Knight Financial Services (BKFS) dataset contains loan-level information for 173,310,331 loans. Although the dataset includes loans from as early as October 1949, we remove observations with \( \text{ClosingMonth} < 121 \), which indicate loans originating prior to 1990. The BKFS data format is a monthly sequence where \( \text{ClosingMonth} = 0 \) for December 1979, \( \text{ClosingMonth} = 1 \) for January 1980, and so on. Additionally, 35 more observations with \( \text{ClosingMonth} > 443 \) are removed as these suggest loans originating after the November 2016 cutoff of the dataset.

Although the data includes a variable for the ZIP code, there are two primary reasons for conducting the analyses at the Core-Based Statistical Area (CBSA) level. First, the ZIP code variable is only reliable at the three-digit level. Despite the BKFS documentation indicating only the first three digits, there are some observations that contain all five digits. However, many simply have the first three, followed by 00. Also, ZIP codes are constructed such that their population sizes are fairly consistent, and thus, larger cities simply have more ZIP codes. Thus, for the rank-size approach in this paper, the CBSA geography level is most appropriate.

The BKFS variable, \textit{CBSA\_MetroDivId}, provides the identifier for the CBSA of each property securing the respective loans. However, for the 11 CBSAs that are broken into Metropolitan Divisions, the variable provides the division code instead of the CBSA code. These divisions are aggregated into their respective CBSAs using the July 2015 delineation file from the U.S. Census. Although there have been more recent updates to these delineations, the July 2015 update is the most recent prior to our obtainment of the dataset. Thus, any changes do not impact the assignment. For example, the Chicago-Naperville-Evanston, IL, Metropolitan Division code was changed from 16974 to 16984 in the Sept. 2018 update.

Additionally, observations are removed if there are missing or non-positive original loan balances or terms (or terms longer than 480 months). This leaves a final set of 150,468,530 loans spanning 929 CBSAs.