Measuring the effects of geographical distance on stock market correlation

Stefanie Eckel[a], Gunter Löffler[b], Alina Maurer[*,b], Volker Schmidt[a]

[a] *Ulm University, Institute of Stochastics, Helmholtzstr. 18, 89069 Ulm, Germany*
[b] *Ulm University, Institute of Finance, Helmholtzstr. 18, 89069 Ulm, Germany*

Abstract

Recent studies suggest that the correlation of stock returns increases with decreasing geographical distance. However, there is some debate on the appropriate methodology for measuring the effects of distance on correlation. We modify a regression approach suggested in the literature and complement it with an approach from spatial statistics, the mark correlation function. For the stocks contained in the S&P 500 that we examine, both approaches lead to similar results. Contrary to previous studies we find that beyond 50 miles geographical proximity is irrelevant for stock return correlations. For distances below 50 miles, we can show that the magnitude of local correlations varies with investor sentiment.

Key words: stock returns, residual correlation, mark correlation function, spatial correlation, investor sentiment
JEL: R12, G11, G14

1. Introduction

Several studies have documented that investment decisions are affected by geographical location within a country. Both institutional investors (e.g. Coval and Moskowitz, 2001) as well as retail investors (e.g. Grinblatt and Keloharju, 2001; Huberman, 2001; Ivkovic and Weisbenner, 2005) allocate

[☆] This research was supported by the Deutsche Forschungsgemeinschaft (DFG) through the SFB 649 *Economic Risk*. S. E. is supported by a grant of the graduate college 1100 at Ulm University.

[*] Corresponding author. Tel.: +49-731-50-23591; Fax.: +49-731-50-23950
Email address: alina.maurer@uni-ulm.de (Alina Maurer)

a disproportionately large fraction of their portfolios to firms that are close to their offices or homes. Possible reasons for such investment patterns are informational advantages and behavioral preferences for familiarity.[1] The latter can lead to locally correlated trading through the following channel: local information or events – be they value-relevant or not – lead to correlated trading activity of local investors; as investors focus on local stocks, trading patterns in nearby-stocks will be correlated, too. Pirinsky and Wang (2006) and Barker and Loughran (2007) test an implication of this conjecture and conclude that the correlation of stock returns increases with decreasing distance. Another possible explanation for locally correlated stock returns is locally correlated fundamentals, but Pirinsky and Wang (2006) fail to find support for this second explanation.

There is some debate on the appropriate methodology for measuring the effect of distance on correlation. Barker and Loughran (2007), for example, question the approach of Pirinsky and Wang (2006). One contribution of our paper is therefore methodological. We modify the regression analysis suggested by Barker and Loughran (2007) and complement it with an approach from spatial statistics, the mark correlation function. For the stocks contained in the Standard and Poor's 500 index (S&P 500) that we examine, both approaches lead to similar results. Contrary to previous studies we find that beyond 50 miles geographical proximity is irrelevant for stock return correlations. We only document an increase in correlations if headquarters are less than 50 miles apart.

In a second contribution to the literature we study the time-series behavior of local correlations. This is made possible by the mark correlation approach, which only needs one cross-section of returns to generate estimates of geographical correlation. We find that the magnitude of local correlation varies with a proxy of investor sentiment suggested by Baker and Wurgler (2007). The finding supports the behavioral-based explanation for local correlation.

The remainder of the paper is organized as follows. Section 2 describes the data, Section 3 the methodology. Empirical results on local correlation are reported in Section 4. Section 5 concludes.

[1] For experimental evidence on the role of familiarity, see Heath and Tversky (1991).

2. Data

Our analysis is based on firms listed on the S&P 500 on August 15th 2005.[2] Monthly stock returns are obtained from the Center for Research in Security Prices (CRSP). Address information (state, city and five digit zip code) for the location of headquarters along with the Standard Industrial Classification (SIC) code are from the annual COMPUSTAT Database. Information on population numbers is taken from US Census Bureau (2009), which derive from the census 2000.

We analyze monthly stock returns for the period 01/2000 - 12/2008 for firms with headquarters located in the USA.[3] We do include delisting returns (e.g. the delisting return for Lehman Brothers Holdings Inc. in September 2008 is -99.68%). One could be concerned that such extreme observations might have a strong influence on the results. In a sensitivity analysis, we remove firms that get delisted and re-run the analyses. This has virtually no effect on the findings reported in the paper.

We demand a minimum of 60 monthly stock returns. This reduces the S&P 500 sample to a total of 484 firms or 116886 = 484(484-1)/2 firm pairs (due to symmetry reasons each pair only needs to be considered once).

For each firm pair we compute the distance between headquarter locations based on the geographical coordinates of the five digit zip code. Applying the correction for the curvature of the earth, the distance $d(i,j)$ between two firms i and j is given by

$$d(i,j) = \rho \arccos(\cos(lat_i)\cos(lat_j)\cos(long_i - long_j) + \sin(lat_i)\sin(lat_j)),$$

where ρ is the earth circumference ($\rho = 3959.871$ miles) and $(lat_i, long_i)$ denote the geographical coordinates, i.e. the latitude and longitude in radian (e.g. Zwillinger, 1995). Geographical coordinates for US zip codes can be obtained from US Census Bureau (2009). Figure 1 shows the distribution of distances between firm pairs in our sample. The graph displays the proportion of firms within a certain distance class. The maximum distance is 2714.6 miles with a mean of 1059.1 miles.

[2] We chose this date to make our results comparable to those of Barker and Loughran (2007).

[3] We check the robustness of our results for the periods 01/2000 - 12/2004 (the sample used by Barker and Loughran, 2007) and 01/2005 - 12/2008.

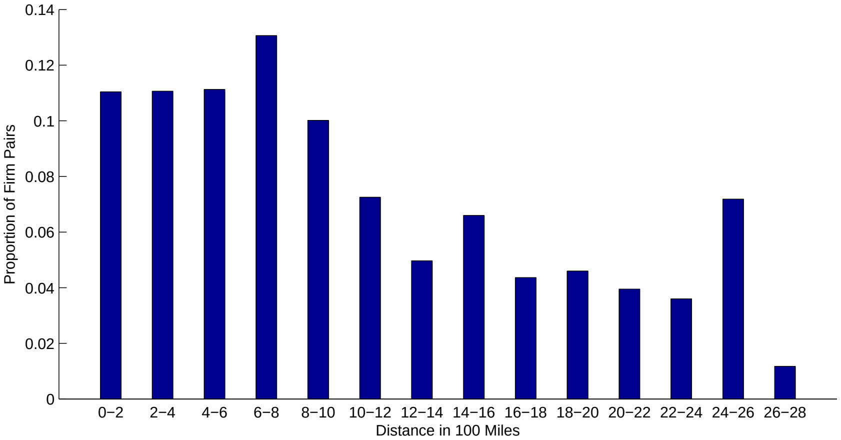


Figure 1: **Sample characteristics.** Proportion of firm pairs at a given distance (measured in 100 miles).

The analysis of comovements of stock returns is based on the analysis of residual stock returns in order to control for the possibility that geographical clustering of firms with similar characteristics leads to higher correlations of nearby firms. Residual returns are obtained from a five-factor regression with excess return being the dependent variable:

$$\begin{aligned} R_{j,t} - RF_t = a_j + b_j(RM_t - RF_t) + s_j SMB_t + h_j HML_t \\ + m_j MOM_t + c_j(RI_{j,t} - RM_t) + \varepsilon_{j,t} \end{aligned} \tag{1}$$

Here, $R_{j,t}$ is the stock return for firm j in month t and RF_t denotes the risk-free return in month t. Based on the results of Fama and French (1993) we control for common variation in stock returns by including the market excess return $RM_t - RF_t$, the return of small minus big stocks (SMB_t), and the return of high book-to-market minus low book-to-market stocks (HML_t) in the set of independent variables.[4] Following Carhart (1997) we also add a momentum factor (MOM_t). In order to control for industrial clustering, we include the difference between the mean industry return and the market return $RI_{j,t} - RM_t$. Here, $RI_{j,t}$ denotes the mean industry return at time t for industry class of firm j. We consider the 48 industry classes as defined in Fama and French (1997), which are based on the four digit SIC code. The

[4] We use the data provided by Kenneth French (see French, 2009).

corresponding monthly industry returns are taken from Kenneth French's data library (see French, 2009), who computes the mean industry return over all stocks traded on the AMEX, NYSE and NASDAQ. Regression (1) is run for each firm separately.

3. Methodology

We introduce a new approach for the analysis of spatial stock market correlations: the mark correlation function. The analysis is complemented by a dummy regression approach.

The main difference between the mark correlation function and a regression analysis relates to the order in which the data is analyzed. Our data consists of a number of measurement locations (locations of firms' headquarters) and a number of measurement times (months). Computing the mark correlation function, we get a functional correlation estimate for each month which is then averaged over time. In the regression approach, we first average over time by computing the time series correlation for each pair individually. Then we average over space by analyzing the drivers of these correlations.

3.1. Mark Correlation Function

We introduce a method from spatial statistics that is based on the so-called mark correlation function for marked point processes (e.g. Cressie, 1993; Stoyan and Stoyan, 1994; Illian et al., 2008). In order to analyze the spatial correlations of stock returns, we consider the locations of firm headquarters as points X_i on the (spherical) surface of the earth and their residual stock returns as marks R_i of these points. The sequence $(X_1, R_1), (X_2, R_2), \ldots$ is then a marked point process on the sphere $S_\rho \subset \mathbb{R}^3$, where S_ρ has its midpoint at the origin and circumference $\rho = 3959.871$ miles. The point process is assumed to be isotropic, which means that its distribution is invariant with respect to arbitrary rotations of the (spherical) coordinate system. Further, the marked point process is called independently marked (or independently labeled) if the marks $R_1, R_2, \ldots$ are independent and identically distributed random variables, which are independent of the sequence $X_1, X_2, \ldots$

Our study includes only firms with headquarters located within the USA. Consequently, we can consider the sequence $(X_1, R_1), (X_2, R_2), \ldots$ as the restriction of a (more comprehensive) marked point process on S_ρ to the territory of the USA.

The mark correlation function $\kappa(r)$ of the marked point process quantifies the stochastic correlation of marks of points that are located at a given distance $r > 0$. Heuristically speaking, positive values of $\kappa(r)$ indicate that pairs of points with distance r have similar marks, while negative values of $\kappa(r)$ indicate different marks. In case of independent marking, it can be shown that $\kappa(r) \equiv 0$ holds for any $r > 0$. The mark correlation function of $(X_1, R_1), (X_2, R_2), \ldots$ can therefore be interpreted as a quantitative characteristic of the spatial interaction between the marks R_i of the points X_i.

A more formal definition of the mark correlation function can be found e.g. in Illian et al. (2008), where numerous applications of this point process characteristic are discussed. Further examples of statistical correlation analysis for spatial marked point patterns are investigated in Eckel et al. (2008), where the temporal trend of the geographical correlations of the purchasing power in Baden-Württemberg, Germany, is analysed, and in Mattfeldt et al. (2009), who present a spatial correlation analysis of labelling patterns for mammary carcinoma cell nuclei.

For any $r \in (0, r_{\max})$, where $r_{\max}$ is a suitably chosen maximum distance, a statistical estimator $\widehat{\kappa}(r)$ for $\kappa(r)$ is given by

$$\widehat{\kappa}(r) = \frac{\displaystyle\sum_{X_i, X_j \in W, i \neq j} k_h(r - |X_i - X_j|)(R_i - \widehat{\mu})(R_j - \widehat{\mu})}{\displaystyle\sum_{X_i, X_j \in W, i \neq j} k_h(r - |X_i - X_j|)} \Big/ \widehat{\sigma^2}\,, \tag{2}$$

where $|X_i - X_j|$ is the spherical distance of X_i and X_j, while k_h is the Epanechnikov kernel with bandwidth $h = 20$ miles, and W denotes the sampling window (in our case, the territory of the USA). Furthermore,

$$\widehat{\mu} = \frac{1}{\#\{n : X_i \in W\}} \sum_{X_i \in W} R_i$$

and

$$\widehat{\sigma^2} = \frac{1}{\#\{i : X_i \in W\} - 1} \sum_{X_i \in W} (R_i - \widehat{\mu})^2$$

are estimators for the mean and variance of the marks, respectively. The estimator $\widehat{\kappa}(r)$ given in formula (2) has been implemented using the Java-based GeoStoch library, which has been developed during the last 10 years at Ulm University (Mayer et al., 2004).

For the definition and estimation of the mark correlation function it is convenient to consider so-called simple point patterns only, i.e. there is at most one mark R_i at any location X_i, which means that $R_i = R_j$ if $X_i = X_j$. Thus, we aggregate the residual returns of firms with the same zip code to a single value, where we use the mean residual return as joint mark. This leaves us with 355 locations, i.e. 355 points X_i in the point pattern.

Given the estimates $\widehat{\kappa}^{(1)}(r), \ldots, \widehat{\kappa}^{(T)}(r)$ for a total of T months, the mark correlation at a distance r is measured by $\overline{\kappa}(r) = \big(\widehat{\kappa}^{(1)}(r) + \ldots + \widehat{\kappa}^{(T)}(r)\big)/T$ and an approximate (pointwise) 95% confidence interval is given by

$$\big(\overline{\kappa}(r) - z_{0.975}\mathrm{SE}(\overline{\kappa}(r)),\, \overline{\kappa}(r) + z_{0.975}\mathrm{SE}(\overline{\kappa}(r))\big),$$

where

$$\mathrm{SE}(\overline{\kappa}(r)) = \sqrt{\frac{1}{T(T-1)} \sum_{t=1}^{T} \big(\widehat{\kappa}^{(t)}(r) - \overline{\kappa}(r)\big)^2}\,, \tag{3}$$

and $z_{0.975}$ denotes the 0.975 quantile of the standard normal distribution.[5]

3.2. Regression Model

We complement the mark correlation approach by analyzing the effect of distances between headquarter locations on the correlation of stock returns by means of a least squares (OLS) regression. Pairwise correlation is regressed on a set of dummy variables that capture the distance between two headquarter locations. For this purpose we define distance classes and set the respective dummy to one if the distance between two firms belongs to a certain distance class, i.e. for some $d < d'$ we put

$$D_{i,j}^{d,d'} = \begin{cases} 1 \text{ if } d \text{ miles} \leq \text{distance between firms } i \text{ and } j < d' \text{ miles}, \\ 0 \text{ else.} \end{cases} \tag{4}$$

The empirical correlation of residual returns for firms i and j is given by

$$CORR_{i,j} = \frac{\sum_{t=1}^{N_{i,j}} (\hat{\varepsilon}_{i,t} - \bar{\hat{\varepsilon}}_i)(\hat{\varepsilon}_{j,t} - \bar{\hat{\varepsilon}}_j)}{\sqrt{\sum_{t=1}^{N_{i,j}} (\hat{\varepsilon}_{i,t} - \bar{\hat{\varepsilon}}_i)^2 \sum_{t=1}^{N_{i,j}} (\hat{\varepsilon}_{j,t} - \bar{\hat{\varepsilon}}_j)^2}}, \tag{5}$$

[5] Formula (3) is based on the assumption that the estimated values $\widehat{\kappa}^{(1)}(r), \ldots, \widehat{\kappa}^{(T)}(r)$ are independently sampled, which we test with Fisher's g-test (see Brockwell and Davis, 1991). The null hypothesis of Gaussian white noise is not rejected (at a significance level of 5%) for all but one distance.

where $\hat{\varepsilon}_{i,t}$ denotes the residual return of firm i at time t and $\bar{\hat{\varepsilon}}_i$ is the mean residual return of firm i. Further, $N_{i,j}$ is the number of available pairwise observations for firms i and j. To obtain a valid estimate for the empirical correlation, we demand a minimum of 60 pairwise observations. The general regression model has the form

$$CORR_{i,j} = \alpha + \beta^{\top} D_{i,j} + u_{i,j}, \tag{6}$$

where u denotes the error term, D is a vector of dummy distance variables, α denotes the regression constant, and $\beta^{\top} = (\beta_1, \ldots, \beta_l)$ is the transposed vector of respective coefficients, with l being the number of distance classes considered. One might be concerned about the appropriateness of linear regression because the dependent variable, being a correlation coefficient, is bounded between -1 and 1. However, since we examine the correlation of residual returns, which are typically small, the majority of observations in our sample do not come close to the bounds. The minimum (maximum) correlation computed for residual returns from 01/2000 - 12/2008 is -0.69 (0.92). The rather large maximum value is computed for the pair Fannie Mae and Freddie Mac and is an outlier, since the 1% and 99% percentiles amount to -0.29 and 0.31, respectively.

By construction the observations in our sample are not independent. Each firm contributes to multiple observations because we consider pairwise correlations. Consequently, the OLS standard errors are biased downward, which leads to inflated t-statistics for $\widehat{\alpha}$ and $\widehat{\beta}$. Following Barker and Loughran (2007) we address this problem by estimating the standard errors through a bootstrap simulation. More precisely, we generate 1000 bootstrap samples by drawing with replacement from the original data and then run an OLS regression for each bootstrap sample. This results in 1000 bootstrap coefficient estimates $\widehat{\alpha}^{(i)}$ and $\widehat{\beta}^{(i)}$ $(i = 1, \ldots, 1000)$. The standard error of the k-th component of the OLS sample estimate $\widehat{\beta}$ (or $\widehat{\alpha}$) is set to the bootstrap standard error

$$\mathrm{SE}^*(\widehat{\beta}_k) = \sqrt{\frac{1}{999} \sum_{n=1}^{1000} \Big(\widehat{\beta}_k^{(n)} - \frac{1}{1000} \sum_{m=1}^{1000} \widehat{\beta}_k^{(m)}\Big)^2}\,, \tag{7}$$

and an approximate 95% confidence interval is given by

$$\big(\widehat{\beta}_k - z_{0.975}\mathrm{SE}^*(\widehat{\beta}_k),\ \widehat{\beta}_k + z_{0.975}\mathrm{SE}^*(\widehat{\beta}_k)\big)$$

(e.g. Sprent and Smeeton, 1994).

An open question in the bootstrap simulation is the resampling technique. Barker and Loughran (2007) suggest to randomly draw firms with replacement from the set of all firms and then construct the bootstrap sample by selecting pairwise combinations of the randomly drawn firms from the original sample (pairwise combinations for the same firm are omitted). In our application the number of firms to be drawn is 484. We will refer to this resampling approach as *firm-by-firm resampling.*

Alternatively and in the style of the block bootstrap discussed in Efron and Tibshirani (1993), one could draw observations from blocks with correlated observations. We suggest to define block i to be the set of all observations in which firm i is one of the two firms whose correlation is the dependent variable. The number of blocks is then equal to the number of firms. The bootstrap sample is constructed by drawing such blocks with replacement. We will refer to this approach as *blockwise resampling.*

A bootstrap sample constructed by blockwise resampling always includes observations for all firms from the original sample. With firm-by-firm resampling, firms that were not drawn do not appear in a particular bootstrap sample. This implies a smaller variation for blockwise bootstrap samples and we therefore expect smaller standard errors for the coefficient estimates using this technique. However, since the dependence structure in our sample is rather complicated it is not obvious from the literature which approach is superior. We therefore consider both approaches for the bootstrap simulation.

4. Empirical Results

We first present results of the spatial correlation analysis for the period 01/2000 - 12/2008. Section 4.2 reports results for the analysis on subsamples, splitting the data with regard to time and population size. Section 4.3 analyzes the link between local correlations and investor sentiment.

4.1. Correlation of residual returns – January 2000 to December 2008

The results from the analysis using the mark correlation function are presented graphically as the mark correlation function provides quasi-continuous values. The estimated spatial correlations for a distance up to 500 miles along with the pointwise 95% confidence interval are shown in Figure 2. The mean

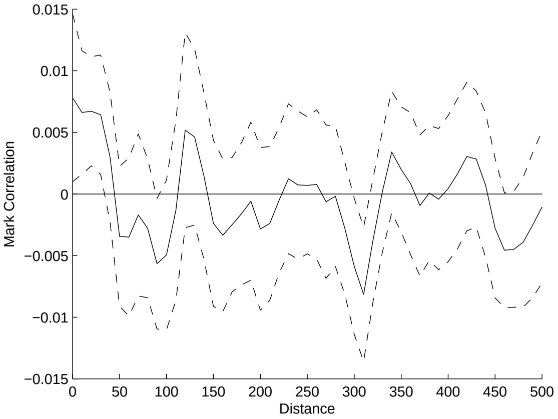


Figure 2: **Residual correlation.** Correlation of residual returns (01/2000 - 12/2008) for a distance up to 500 miles from the mark correlation function. Dashed lines show the pointwise 95% confidence interval.

mark correlation function $\bar{\kappa}(r)$ fluctuates around zero, with the zero line being included in the confidence interval almost everywhere for distances larger than 50 miles. It is only for distances up to 50 miles that the results suggest a statistically significant positive residual correlation.

We investigate this finding further by the following two specifications of the dummy regression model (equation 6):

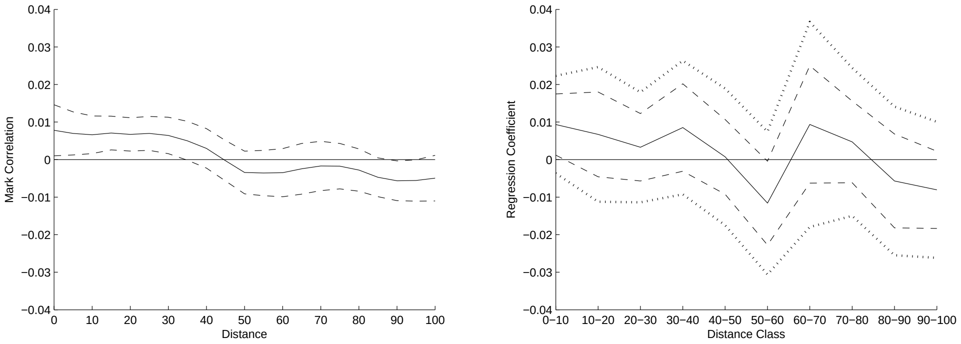
$$CORR_{i,j} = \alpha + \sum_{k=1}^{10} \beta_k D_{i,j}^{50(k-1),50k} + u_{i,j}. \tag{8}$$

$$CORR_{i,j} = \alpha + \sum_{k=1}^{10} \beta_k D_{i,j}^{10(k-1),10k} + u_{i,j}. \tag{9}$$

The first specification considers 10 distance classes, each capturing a distance of 50 miles, and a reference class for all distances exceeding 500 miles. The second specification takes a closer look on short distances, considering 10 smaller distance classes, each for a distance of 10 miles, and a reference class for all distances exceeding 100 miles.

Regression results are presented in Table 1. We report coefficient estimates from an OLS regression on the original sample[6] and t-statistics that

[6] The means of the bootstrap coefficient estimates are almost identical for the two



(a) Mark correlation function

(b) Regression analysis

Figure 3: **Residual correlation.** Correlation of residual returns (01/2000 - 12/2008) with pointwise 95% confidence intervals for a distance up to 100 miles using the mark correlation function (Subfigure (a)) and regression analysis (Subfigure (b)). Dotted and dashed lines in Subfigure (b) are pointwise 95% confidence intervals resulting from firm-by-firm and blockwise resampling, respectively.

are based on bootstrap standard errors. We use the firm-by-firm as well as the blockwise resampling technique. The latter results in smaller standard errors, which can be ascribed to the smaller variation in the blockwise resampling samples.

Figure 3 provides a graphical comparison of results for distances up to 100 miles. Subfigures (a) and (b) show the estimated mark correlation and results for the second regression specification (equation 9), respectively. The illustration of results from the regression analysis derives from a linear interpolation of the estimated coefficients. Both graphs show a similar pattern for spatial correlations: an approximately constant level of correlation up to a distance of 30 to 40 miles and a sharp decline thereafter. Moreover, both approaches suggest a rather small residual correlation for nearby firms. The mean mark correlation for firms located in the same zip code is approximately 0.008; the coefficient of $D^{0,10}$ implies an average increase of correlation by 0.009 as compared to the reference class.

The results differ somewhat in terms of statistical significance for very nearby firms. Confidence intervals are narrower in the mark correlation approach. In the regression approach, the dummy variable for distances be-

techniques and very close to the values obtained from the OLS regression.

Specification 1		Specification 2	
$D^{0,50}$	0.0073	$D^{0,10}$	0.0093
	(1.709)[2.633]		(1.411)[2.261]
$D^{50,100}$	-0.0035	$D^{10,20}$	0.0067
	(-0.754)[-1.247]		(0.734)[1.176]
$D^{100,150}$	0.0086	$D^{20,30}$	0.0033
	(1.422)[2.459]		(0.439)[0.721]
$D^{150,200}$	0.0011	$D^{30,40}$	0.0085
	(0.250)[0.416]		(0.940)[1.444]
$D^{200,250}$	0.003	$D^{40,50}$	0.0007
	(0.638)[1.073]		(0.078)[0.141]
$D^{250,300}$	0.0035	$D^{50,60}$	-0.0116
	(0.861)[1.566]		(-1.194)[-2.043]
$D^{300,350}$	0.0009	$D^{60,70}$	0.0094
	(0.236)[0.389]		(0.671)[1.182]
$D^{350,400}$	0.0052	$D^{70,80}$	0.0047
	(1.136)[1.819]		(0.469)[0.844]
$D^{400,450}$	0.0043	$D^{80,90}$	-0.0057
	(0.925)[1.518]		(-0.563)[-0.885]
$D^{450,500}$	-0.0004	$D^{90,100}$	-0.0081
	(-0.096)[-0.163]		(-0.871)[-1.541]
Constant	0.0036	Constant	0.0042
	(3.164)[5.064]		(3.881)[6.209]
Firm Pairs	(116301)[117066]	Firm Pairs	(116301)[117066]
Adj. R^2	(0.0005)[0.0003]	Adj. R^2	(0.0005)[0.0003]

Table 1: **Regression results.** Dependent variable is the pairwise correlation of residual stock returns. t-statistics are computed based on the firm-by-firm (in parentheses) and blockwise (in square brackets) resampling techniques. Specification 1 (Specification 2) includes dummy variables for 10 equidistant distance classes with a reference class for distances larger that 500 miles (100 miles). Adjusted R^2 and firm pairs are averages from 1000 bootstrap simulations.

tween 0 and 50 miles is significant on the 10% (firm-by-firm resampling) and 1% (blockwise resampling) levels. For the mark correlation approach the 95% confidence interval excludes the zero line for distances up to approximately 30 miles (see Figure 3); additional computations show that the average correlation between 0 and 50 miles is significant at the 1% level. The marginal significance in firm-by-firm resampling should give rise to some caution. Nevertheless, it seems an apt summary to say that there is evidence for stock market correlation being higher for firms that are less than 50 miles apart. The economic significance, however, appears to be very limited, if not negligible. What makes typical retail portfolios risky is unlikely to be the distance effect documented here. It is the low degree of diversification (many investors hold less than three stocks[7]) and, if portfolios are local, the local clustering of firm characteristics. When we remove the industry indices from our factor model, the residual return correlation in the distance class 0 to 50 miles increases from 0.007 to 0.03. While industry clustering has a strong effect on local correlations, the effect of distance per se is much smaller.

Moving on to larger distances, there is no consistent evidence for differences in stock market correlations once the distance is larger than 50 miles. Within the regression approach, firm-by-firm resampling does not lead to coefficients that are significant at a level of 5% or better; with blockwise resampling, there is one significant dummy (100 to 150 miles). The mark correlation function, finally, exhibits only a few isolated cases where the distance is significant at the 5% level.

Our results therefore do not support the findings by Barker and Loughran (2007), who conclude that geographical proximity is an important factor explaining monthly return correlations. In particular, these authors report an increase in correlation of returns by 12 basis points for each reduction in distance by 100 miles.

Barker and Loughran (2007) regress pairwise correlations of stock returns on distance as well as a set of control variables. To capture industry effects, for example, these authors include the industry correlation as an explanatory variable; to capture differences in systematic risk, they include the differences of the Fama-French 3-factor betas.

To ensure that our different finding does not arise from differences in

[7] In the sample of Barber and Odean (2000), for example, the mean household holds 4.3 stocks.

Industry Correlation	0.3300	(14.776)
In Same Industry	-0.0215	(-1.630)
Information Technology	0.0820	(4.046)
2000 miles apart	-0.1681	(-2.613)
Distance $<$ 2000	-0.0024	(-4.850)
Distance $>$ 2000	0.0054	(2.025)
Const.	0.0694	(5.543)
Firm Pairs		116294
Adj. R^2		0.2084

Table 2: **Replication of the regression approach of Barker and Loughran (2007).** Dependent variable is the pairwise correlation of stock returns. t-statistics are in parentheses. Adjusted R^2 and firm pairs are averages from 1000 bootstrap simulations. The definition of variables and the bootstrap (firm-by-firm resampling) follows Barker and Loughran (2007, Model 2 in Table 3).

samples, but only from differences in methodologies, we estimate Barker and Loughran's Model 2 using our data set. In this model, correlation coefficients of stock returns are regressed on correlation of industry returns and two dummy variables indicating whether two firms operate in the same industry or belong to the information technology sector. The distance effect is captured by a dummy variable set to one if two firms are 2000 miles or further apart along with two variables measuring the distance between firms (in 100 miles). The first one measures distances up to a distance of 2000 miles and the second one those beyond 2000 miles.[8]

Results reported in Table 2 imply a conclusion similar to that of Barker and Loughran (2007). We find the variable of interest (Distance $<$ 2000) having the right sign and being highly statistically significant. This suggests that for each decrease in distance by 100 miles the correlation of returns (even after controlling for industry correlation) increases significantly. The difference between the results of Barker and Loughran and the ones presented above are therefore due to methodology.

We favor our approach because it appears that the Barker and Loughran approach does not sufficiently control for correlation due to factors other

[8] Barker and Loughran (2007) also present results for a richer model including the differences of the Fama-French 3-factor betas. These additional variables, however, have no effect on the magnitude or significance of the distance coefficients.

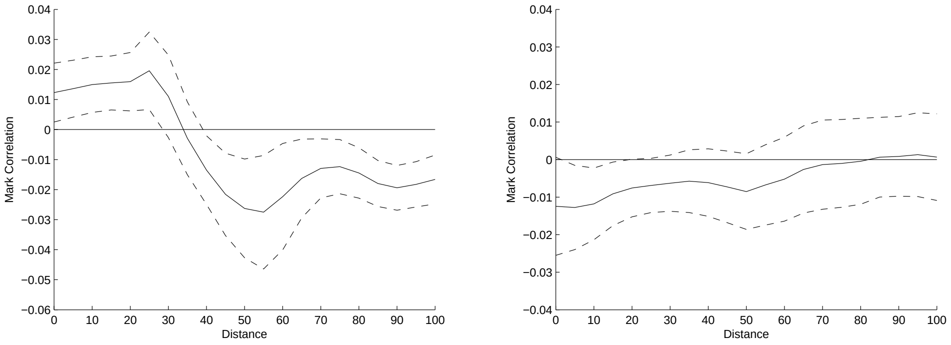
than distance. Consider, for example, three firms from the same industry, two of them exhibiting average exposure to industry risk, and one a lower exposure (e.g. because some operations belong to another industry). In the Barker and Loughran model one would implicitly assume that the correlation of these three firms is the same. In our model, differences will be captured through the industry coefficient in the factor regression. Similarly, consider two pairs of firms: the betas of the pair {firm 1, firm 2} are $\beta_1 = 0.5$ and $\beta_2 = 1$, the betas of {firm 3, firm 4} are $\beta_3 = 1$ and $\beta_4 = 1.5$. The beta difference used as an explanatory variable by Barker and Loughran (2007) is the same for each pair ($\beta_2 - \beta_1 = \beta_4 - \beta_3 = 0.5$). However, ceteris paribus, firms 3 and 4 should have a higher correlation than firms 1 and 2 because the covariance that is due to common variation with the market return is $0.5Var(R_M)$ and $1.5Var(R_M)$ for pairs {firm 1, firm 2} and {firm 3, firm 4}, respectively.

4.2. Correlation of residual returns – Subsamples

Previous studies consider the size of headquarter cities (measured by the population number) as a potential factor to influence the relation of distance and stock return correlation. The findings are controversial. Pirinsky and Wang (2006) report that correlation of stock returns is higher in larger cities. Barker and Loughran (2007) draw the opposite conclusion from the observation that the difference in the populations of headquarter cities does not explain correlations.[9]

We control for the size effect by splitting the sample with respect to population and analyzing each subsample separately. More precisely, we classify cities as large if the population number exceeds 100000 inhabitants and as small otherwise. Results from the analysis using the mark correlation function are presented in Figure 4 for each subsample. The results support the findings of Pirinsky and Wang (2006). The correlation of residual stock returns is higher for nearby firms located in large cities (Subfigure (a)) than for nearby firms located in small cities (Subfigure (b)). The regression analysis yields a coefficient of 0.012 with standard errors 0.007 (firm-by-firm resampling) and 0.005 (blockwise resampling) for the dummy variable $D^{0,50}$ in the

[9] The approach chosen by Barker and Loughran (2007) does not distinguish between pairs of small cities and large cities. Consider firms located in two equally sized large and two equally sized small cities. The difference in populations for the large and for the small pair is zero and does not control for the difference between the pairs.



(a) Large cities

(b) Small cities

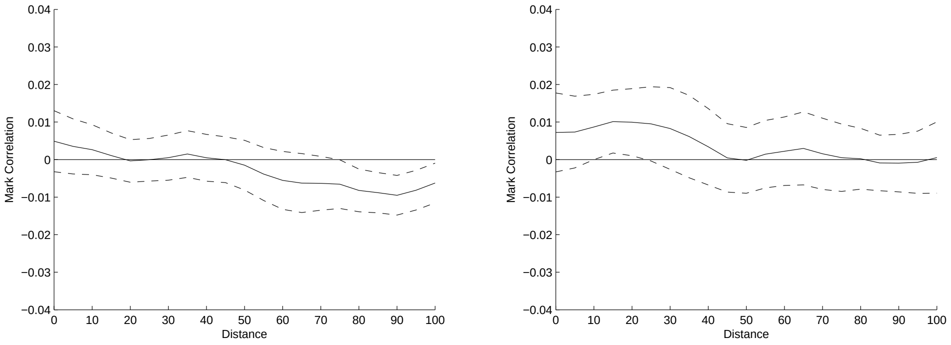
Figure 4: **Residual correlation (subsamples).** Correlation of residual returns measured by the mark correlation function for firm pairs located in large cities (Subfigure (a)) and small cities (Subfigure (b)). A city is classified as large, if it has more than 100000 inhabitants. Pointwise 95% confidence intervals are shown as dashed lines.

large cities subsample. Whereas in the small cities subsample, the respective coeffiecient is -0.003 with standard errors 0.007 (firm-by-firm resampling) and 0.004 (blockwise resampling). Furthermore, using the mark correlation function, we unexpectedly detect slight, but significant negative correlations for firms located in large cities at a distance larger than 40 miles.

In Figure 5 we present estimates for the mark correlation for subsamples after splitting the sample with respect to time. We consider a sample including returns for the period 01/2000 - 12/2004 (this is the time span used by Barker and Loughran (2007)) and a second sample for the period 01/2005 - 12/2008. For both subsamples the results are close to those of the entire sample, with the significance of the proximity of headquarter locations being more pronounced for the later period.

4.3. Residual correlations and sentiment over time

Extant papers favor the view that local information or local exchange generates locally correlated trading patterns because investors have a tendency to hold local stocks (see Pirinsky and Wang, 2006; Barker and Loughran, 2007). This line of argument has not been subject to a rigorous test because data on local events and their impact on trading behavior are difficult to collect. Here, we suggest to examine a time-series implication of the argument. If trading activity of locally focused investors is high, the impact of any local



(a) 01/2000 - 12/2004

(b) 01/2005 - 12/2008

Figure 5: **Residual correlation (subsamples).** Correlation of residual returns measured by the mark correlation function for subsamples including returns from 01/2000 - 12/2004 (Subfigure (a)) and from 01/2005 - 12/2008 (Subfigure (b)). Pointwise 95% confidence intervals are shown as dashed lines.

event on stock returns can be expected to be high, too. Since locally focused investors are typically associated with people subject to behavioral biases, we use a measure of investor sentiment to proxy their trading activity.

Our hypothesis is therefore that local correlations vary positively with investor sentiment. To quantify the latter, we use the monthly investor sentiment index constructed by Baker and Wurgler (2007). It is available (on Jeffrey Wurgler's website (see Wurgler, 2009)) until December 2005, which means that we can cover 60% of our 2000-2008 sample. To measure monthly intensity of local correlations, we use the mark correlation approach. As described in Section 3.1 it produces a time series of monthly geographical correlations and is therefore well suited to study the time-series behavior of correlations. What is left to determine is how to measure the magnitude of local correlation effects. In Section 4.1 we have seen that correlations are only significant if the distance between firm headquarters is smaller than 50 miles. We therefore suggest to examine the average correlation for all distances smaller than 50 miles, quantified by $\bar{\kappa}_t^{0,45} = \frac{1}{10}\sum_{i=0}^{9}\kappa_t(5i)$. This variable is regressed on the contemporaneous sentiment index from Baker and Wurgler (2007). To control for the possibility that sentiment is correlated with general shifts in correlation, we also include the average correlation for distances between 50 and 100 miles ($\bar{\kappa}_t^{50,100} = \frac{1}{11}\sum_{i=10}^{20}\kappa_t(5i)$). Estimating the regression with OLS leads to (t-statistics in parentheses):

$$\begin{aligned}&\bar{\kappa}_t^{0,45} = \underset{(1.56)}{0.003} + \underset{(2.72)}{0.007}\ \text{Sentiment} + \underset{(0.01)}{0.001}\ \bar{\kappa}_t^{50,100}\\ &\text{N} = 72,\ \text{Adj. R}^2 = 0.073,\ \text{DW-statistic} = 2.205\end{aligned}$$

Local correlation does not vary with more distant correlations. It does, however, vary significantly (p-value better than 1%) with sentiment. The standard deviation of the sentiment index is 0.79. A two-standard deviation move in sentiment therefore changes average correlation in the 0 to 50 miles range by 0.55 percentage points. While this may appear small, it is just as large as the average correlation in that range (0.56%). Thus, if investor sentiment is very high, local correlation can be expected to be larger than 1%. If sentiment is very low, local correlation approaches zero. This result provides direct support to the behavioral-based explanation of local correlations suggested in the literature.

5. Conclusion

We have analyzed spatial correlations of stock returns for firms included in the S&P 500. We find that geographical distance does not influence stock return correlations once the distance exceeds 50 miles, which contradicts the results of Pirinsky and Wang (2006) and Barker and Loughran (2007). We show that the difference in results is purely methodological, confirming that the choice of research methodology is crucial for examining the effects of distance on stock correlations. Barker and Loughran (2007) criticize the methodology by Pirinsky and Wang (2006) and obtain different results for large firms. We further modify the regression approach by Barker and Loughran (2007) but also suggest an alternative approach from spatial statistics, the mark correlation function. The analysis using the mark correlation function leads to the same conclusions as our modified regression approach and thus strengthens the confidence in the robustness of results.

An advantage of the mark correlation function is that it only needs one cross-section of returns to generate estimates of geographical correlation. We exploit this advantage to examine the time-series behavior of local correlations. We document that the magnitude of local correlation varies with a proxy of investor sentiment, supporting the notion that local correlation is driven by behavioral biases.

References

Baker, M., Wurgler, J., 2007. Investor sentiment in the stock market. Journal of Economic Perspectives 21, 129-151.

Barber, B.M., Odean, T., 2000. Trading is hazardous to your wealth: the common stock investment performance of individual investors. Journal of Finance 55, 773-806.

Barker, D., Loughran, T., 2007. The geography of S&P 500 stock returns. Journal of Behavioral Finance 8, 177-190.

Brockwell, P.J., Davis, R.A., 1991. Time series: theory and methods, 2nd ed. Springer, New York.

Carhart, M.M., 1997. On persistence in mutual fund performance. Journal of Finance 52, 57-82.

Coval, J.D., Moskowitz, T.J., 2001. The geography of investment: informed trading and asset prices. Journal of Political Economy 109, 811-841.

Cressie, N., 1993. Statistics for spatial data. J.Wiley & Sons, New York.

Eckel, S., Fleischer, F., Grabarnik, P., Schmidt, V., 2008. An investigation of the spatial correlations for relative purchasing power in Baden-Württemberg. Advances in Statistical Analysis 92, 135-152.

Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap. Chapman and Hall, New York.

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3-56.

Fama, E.F., French K.R., 1997. Industry costs of equity. Journal of Financial Economics 43, 153-193.

French, K.R., 2009. http://mba.tuck.dartmouth.edu/pages/faculty/ken.french.

Grinblatt, M., Keloharju, M., 2001. How distance, language, and culture influence stockholdings and trades. Journal of Finance 56, 1053-1073.

Heath, C., Tversky, A., 1991. Preference and belief: ambiguity and competence in choice under uncertainty. Journal of Risk and Uncertainty 4, 5-28.

Horowitz, J.L., 2001. The bootstrap. In: Heckman J.J., Leamer, E.E. (Eds), Handbook of Econometrics Vol. 5, North-Holland, Amsterdam.

Huberman, G., 2001. Familiarity breeds investment. Review of Financial Studies 14, 659-680.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. Statistical analysis and modelling of spatial point patterns. J.Wiley & Sons, Chichester.

Ivkovic, Z., Weisbenner, S., 2005. Local does as local is: information content of the geography of individual investors' common stock investments. Journal of Finance 60, 267-306.

Mattfeldt, T., Eckel, S., Fleischer, F., Schmidt, V., 2009. Statistical analysis of labelling patterns of mammary carcinoma cell nuclei on histological sections. Journal of Microscopy 235, 106-118.

Mayer, J., Schmidt, V., Schweiggert, F., 2004. A unified simulation framework for spatial stochastic models. Simulation Modelling Practice and Theory 12, 307-326.

Pirinsky, C.A., Wang, Q., 2006. Does corporate headquarters location matter for stock returns? Journal of Finance 61, 1991-2015.

Stoyan, D., Stoyan, H., 1994. Fractals, random shapes and point fields. J.Wiley & Sons, Chichester.

Sprent, P., Smeeton, N.C., 1994. Applied nonparametric statistical methods, 3rd ed. Chapman & Hall/CRC, Boca Raton.

US Census Bureau, 2009. http://www.census.gov.

Wurgler, J., 2009. http://pages.stern.nyu.edu/∼jwurgler.

Zwillinger, D. (Ed.), 1995. Spherical geometry and trigonometry. CRC Press, Boca Raton.