Quantifying geographical and macroeconomic effects on bank branch deposits using linear mixed models*

Eike Christian Brechmann†, Claudia Czado‡ and Peggy Ng§

Abstract

The assessment of performance and potential is central to decisions pertaining to the location of bank branches. A common method for evaluating branch performance is data envelope analysis in which in-branch variables are typically considered. This paper adopts an alternate methodology that quantifies the influence of local socio-economic variables on bank deposits (a common measure of performance) using linear mixed models (LMM). It also illustrates the potential of using LMM to build a predictive model to support branch location decisions.

1 Introduction Commercial banks can operate as a single unit bank or develop a network of bank branches, which act as the key contact point between customers and the central bank. As such, branches occupy key positions in banking organizations and their locations reflect important strategic decisions and operating policies. The rationale for developing a branch network, beyond the significant effect it has on banks' market shares [15], is threefold: diversification of risk, customer convenience and market knowledge.

Key words and phrases: Bank performance, branching, linear mixed models.

AMS 2000 subject classifications. Primary 90B50; secondary 62J05, 62-07.

†*Mailing Address:* Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, D-85747 Garching, Germany. *E–mail:* eike.brechmann@mytum.de.

‡*Mailing Address:* Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, D-85747 Garching, Germany. *E–mail:* cczado@ma.tum.de.

§*Mailing Address:* School of Administrative Studies, York University, 4700 Keele Street, Toronto, Canada. *E–mail:* peggyng@yorku.ca.

Branching facilitates geographic diversification, which allows banks to diversify their assets by improving access to different industries that may respond to shocks differently [13]. However, the extent to which geographic diversification reduces the risk depends on how economically diverse the different geographic areas are (see for example [20]).

Through their local branches banks are able to better obtain and process market-specific knowledge. Jayaratne and Strahan in [14] have shown that real growth in bank income can emanate from the improved loan screening and monitoring that is facilitated by branch network proliferation. These banks may have local information about certain borrowers, local economic conditions or market trends that they may not have had without their local branches and this can insulate certain branches from competitive forces (see [9]).

The optimum number of branches and their optimum locations are interrelated issues that have to be addressed by bank managers. Chelst, Schultz and Sanghvi in [7] provide a general procedure to facilitate this task. One central topic is the performance evaluation of the current branch network. Yet, such assessments are complex multidimensional processes. In fact, Doyle, Fenwick and Savage in [10] found 38 independent variables needed to fully describe branch performance. Boufounou in [4] analyzed a similar number of variables for a commercial bank in Greece while Avkiran in [1] tested 91 potential variables and six performance variables for evaluating branch performance.

A common measure of branch performance is budgeting, which is however criticized for its emphasis on expenses rather than profitability. Measuring the performance of a branch by its profit, which includes earnings from a wide range of services such as loans and mortgages, suffers from problems of suitably allocating revenues and expenses [8]. In this paper we consider the performance measure of total deposits as in [4]. Drawbacks of this simple measure are that it does not distinguish the different kinds of deposits which bring various profit margins and it ignores revenues which are generated from loans (see [1]). However it is certainly one of the main business drivers of banks and easily collected and amenable for a statistical analysis.

Traditional methods for evaluating branches are the performance index (see [19]), econometric methods (see for example [11]) and the commonly used data envelope analysis (DEA). DEA is a non-parametric linear programming technique used to compute a comparative ratio of inputs to outputs for each unit. See [2] for a demonstration of DEA and [16] for a summary of research conducted using DEA. The analysis can incorporate in-branch (discretionary) as well as out-of-branch (non-discretionary) variables (see e.g. [18]).

However often the variables under consideration are all in-branch variables (see [16] and [3]), such as employees space, branch expenses (rent, marketing, operating costs, etc.) and acquired equipment.

The focus of our study is to quantify the influence of out-of-branch variables such as geographical and macroeconomic variables on the performance measure total deposits of a branch. We are especially interested in investigating the effect of local wealth (as measured by county unemployment rates and county income per capita) and local bank competition. For local bank competition a variable depending on the sum of distances of a branch to other branches of other banks is constructed and shown to influence the branch total deposit. For this we build an adequate statistical model, which allows the adjustment of longitudinal and cluster effects. The statistical model chosen is from the class of linear mixed models. We will show that the inclusion of interaction effects significantly improves the model fit. A non-hierarchical model specification with interaction effects is shown to be preferred over a standard linear model as well as a hierarchical specification. The presence of interaction effects points to complex multidimensional influences on the total deposits of a branch. Finally the predictive capabilities of the models are investigated.

The paper is organized as follows: in Section 2 an introduction to the theory of linear mixed models is given as they are used extensively in Section 3 in order to model the determinants for total branch deposits. Section 3 develops and analyzes our main model and evaluates its goodness compared to other models. Finally Section 4 summarizes the main findings and discusses our approach with respect to other methods.

2 Linear mixed models Introductions to the theory and the use of linear mixed models can be found e.g. in [22] and in [17]. The latter also describes the *R*-library *nlme* which is designed for statistical analyses with mixed models. An illustrative approach of fitting linear mixed models using the *nlme* library is given in [12].

The well-known standard linear model can be written as

$$Y = X\beta + \varepsilon,$$

where $Y \in \mathbb{R}^n$ denotes the response vector, $X \in \mathbb{R}^{n\times p}$ is the design matrix, $\beta \in \mathbb{R}^p$ are the regression coefficients, and $\varepsilon \in \mathbb{R}^n$ is the vector of random errors. Usually one assumes $\varepsilon \sim N_n(0, \sigma^2 I_n)$, where $N_n(\mu, \Sigma)$ denotes the n-dimensional normal distribution with mean vector μ and covariance matrix Σ.

However such linear models are not always appropriate to deal with data sets. In linear models independent response variables are assumed, but often this is not the case. Data could be clustered, i.e. the response is measured once for each subject and each subject belongs to a group of subjects (cluster), or longitudinal, i.e. the response is measured at several time points and the number of time points is not too large (the repeated measurements then form a group of dependent observations). For such dependent data structures the linear model has to be extended by allowing group-specific random effects in so-called linear mixed models which can easily be formulated for each group i as an extension of the standard linear model:

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i, \tag{2.1}$$

where $Y_i \in \mathbb{R}^{n_i}$ denotes the n_i observations in group i, $X_i \in \mathbb{R}^{n_i \times p}$ is the design matrix for the fixed effects, $\beta \in \mathbb{R}^p$ are the fixed-effect coefficients, and $\varepsilon_i \in \mathbb{R}^{n_i}$ indicates the errors. Moreover, $Z_i \in \mathbb{R}^{n_i \times q}$ is the design matrix for the random effects with $q \leq p$ and $b_i \in \mathbb{R}^q$ being the random-effect coefficients. Since random effects and errors are random, distributions for them have to be specified. A common choice is:

$$\begin{aligned} \varepsilon_i &\sim N_{n_i}(0, R_i) \\ b_i &\sim N_q(0, G), \end{aligned} \tag{2.2}$$

where ε_i and b_i are independent. Here $R_i \in \mathbb{R}^{n_i \times n_i}$ and $G \in \mathbb{R}^{q \times q}$ are the covariance matrices for the errors and the random effects, respectively. Usually $R_i = \sigma^2 I_{n_i}$ with $\sigma^2 > 0$ and I_{n_i} the n_i-dimensional identity matrix is assumed. Also note that G is assumed to be the same for all groups i.

Sometimes it is more illustrative to express linear mixed models in a hierarchical (nested) form which is easier to interpret (see e.g. [12]). This will be done and explained in Section 3.3. However, note that such a hierarchical model specification will not always be possible for linear mixed models, since effects can be 'crossed' (e.g. variables might have a time and a geographical level but these levels are not nested, i.e. the geographical variables might not be measured over time but only once).

In linear mixed models the fixed-effect coefficients and the covariance parameters of R_i and G have to be estimated. This is usually done using restricted maximum likelihood (REML) estimation which is preferred to standard maximum likelihood (ML) estimation because it produces unbiased estimates [22]. Random effects can be predicted (rather than 'estimated' as they are random variables) using conditional expectations and the estimated

covariances. The predicted values are referred to as empirical best linear unbiased predictors (EBLUPs).

As in linear models, one often wants to test certain hypotheses in order to determine the goodness of fit of a model. Since a linear mixed model incorporates random effects, the choice of an appropriate covariance model for them is crucial. Likelihood ratio tests (LRTs) can be used to test hypotheses with regard to covariance parameters (e.g. to test the assumption of heterogeneous error variances). As usual, models have to be nested to conduct an LRT, i.e. there is a full and a reduced model: the reduced model has less parameters than the full model which incorporates all parameters of the reduced model. Then the LRT statistic is given as $2(\ell_{full} - \ell_{reduced})$ where ℓ denotes the estimated log-likelihood in the respective models. However, the usual null distribution of the LRT statistic is no longer valid in the context of linear mixed models, since null hypotheses are often on the boundary of the parameter space. In particular, it is of interest to compare two nested models with a different number of random effects (and the same fixed effects). In the simple case of one model with q random effects and the other with $q+1$ random effects, the difference in the number of covariance parameters is $q + 1$. The corresponding null hypothesis is on the boundary of the parameter space and the null distribution can be determined as a 50:50 mixture of chi-squared distributions with q and $q + 1$ degrees of freedom. In general, when comparing models with q and $q + k$ ($k > 1$) random effects (or other specifications of covariance parameters), the determination of the null distribution is more complicated (see [21]). We like to note that the corresponding method which is implemented in the *R*-library *nlme* does not use the correct null distributions, but is more conservative, i.e. a null hypothesis is not as easily rejected as under the correct null distribution [17].

Model selection regarding fixed effects is often done using $t - tests$ ($H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$) with test statistic $t = \frac{\hat{\beta}_i}{\hat{se}(\hat{\beta}_i)}$, where $\hat{se}(\hat{\beta}_i)$ denotes the estimated asymptotic standard error of $\hat{\beta}_i$. Unlike in standard linear model theory the null distribution is in general no longer an exact t distribution [22]. LRTs are not appropriate for testing hypotheses regarding fixed effects when using REML estimation, since they are based on ML estimation.

3 Geographical and macroeconomic determinants for total branch deposits

3.1 Data The data considered in our study consists of 2,988 branch-year records of a major US bank in the state of New York with multiple branches. 506 branches are included with observations over the period from 1994 to 2002. The data is clustered (the branches within a county form a cluster) and also longitudinal, since it is observed over a period of nine years. Therefore, a mixed model approach seems to be appropriate to model the dependencies in the data that arise from the clusters (counties) and from the measurements taken on the same subjects (branches, counties, state). Figure 4.1 illustrates the hierarchical structure of the data.

Because of the clustered data, there are three types of variables: variables on the state, county and branch level. All variables are measured over time, but some measurements on the branch level are not taken in all years, since branches closed or opened. To achieve a more homogeneous error variance the logarithm of the total deposits, *log.dep*, is used as dependent variable. The precise definition of variables is given in Table 4.1.

Note that the set of explanatory variables is intentionally chosen rather small to demonstrate the usability of the hierarchical modeling technique for the research purpose. In addition, the chosen variables capture the gist of the population demography. Other variables such as schooling or occupation could be considered, but these effects are correlated with per capita income and unemployment rates (see e.g. [6]). An inclusion of these variables might therefore lead to multicollinearities among the variables and thus to computational problems. For similar reasons we decided to focus on five variables on the macroeconomic level that are closely related the bank's overall performance and therefore closely linked to the performance of single branches. As we aim at building a rather simple illustrative model, we refrained as well from considering local business variables such as number and size of companies.

An impression of the relationship between *log.dep* and all three levels of covariates can be obtained by examining the corresponding scatter plots (see Figure 4.2). The first plot shows that there is a weak positive overall influence of *comp* on *log.dep*, but variation is high for a high level of competition. The panels in Figure 4.2 corresponding to *pop* and *inc.pc* indicate weak positive influences on *log.dep*, whereas *unemp* shows no clear influence on *log.dep*. Individual examinations of the influences per branch and per county show that the effects on *log.dep* vary a lot across branches and counties. However, individual linear regression fits per county do not show a need for random slopes on the county level (see [5]). Finally, the remaining panels of Figure 4.2 do not show a clear influence of any of

the state variables on *log.dep*. There are possibly weak positive influences of *mshare* and *av.dep* on the response *log.dep*.

Interactions between the variables may also be present as the effects on different levels are likely to be interrelated with each other and therefore important information would be omitted if interactions were not taken into account. Possible interactions are thus examined in [5]. To reduce the complexity of the model, only second order interactions are considered.

The analyses underlying this case study are performed with *R* using the libraries *nlme* and *lattice* for the data examination. *R*-commands and -outputs can be found in [5].

3.2 Statistical analysis In contrast to [4] and [1] we fit a regression model with mixed effects as described in Section 2. Thus our initial mixed model includes not only fixed effects for all branch, county and state variables, and their interactions, but also random intercepts and slopes on the branch level as well as random intercepts on the county level. It is stated as in the definition of a linear mixed model (2.1) for branch i in county j in year t. The effects in bold face in equation (3.1) indicate those which will prove to be significant at the 5% level in the final model.

$$
\begin{aligned}
log.dep_{ijt} = {} & \boldsymbol{\beta_0} + \beta_1 comp_{ijt} + \boldsymbol{\beta_2 pop_{jt}} + \boldsymbol{\beta_3 inc.pc_{jt}} + \boldsymbol{\beta_4 unemp_{jt}} \\
& + \beta_5 comp_{ijt} pop_{jt} + \beta_6 comp_{ijt} inc.pc_{jt} + \beta_7 comp_{ijt} unemp_{jt} \\
& + \boldsymbol{\beta_8 no.fail_t} + \boldsymbol{\beta_9 mshare_t} + \boldsymbol{\beta_{10} branch.total_t} + \boldsymbol{\beta_{11} dep.total_t} \\
& + \boldsymbol{\beta_{12} av.dep_t} + \beta_{13} no.fail_t comp_{ijt} + \beta_{14} no.fail_t pop_{jt} \\
& + \beta_{15} no.fail_t inc.pc_{jt} + \boldsymbol{\beta_{16} no.fail_t unemp_{jt}} + \beta_{17} mshare_t comp_{ijt} \\
& + \beta_{18} mshare_t pop_{jt} + \beta_{19} mshare_t inc.pc_{jt} + \boldsymbol{\beta_{20} mshare_t unemp_{jt}} \\
& + \beta_{21} branch.total_t comp_{ijt} + \beta_{22} branch.total_t pop_{jt} \\
& + \beta_{23} branch.total_t inc.pc_{jt} + \boldsymbol{\beta_{24} branch.total_t unemp_{jt}} \\
& + \beta_{25} dep.total_t comp_{ijt} + \beta_{26} dep.total_t pop_{jt} + \beta_{27} dep.total_t inc.pc_{jt} \\
& + \boldsymbol{\beta_{28} dep.total_t unemp_{jt}} + \beta_{29} av.dep_t comp_{ijt} + \beta_{30} av.dep_t pop_{jt} \\
& + \boldsymbol{\beta_{31} av.dep_t inc.pc_{jt}} + \beta_{32} av.dep_t unemp_{jt} \\
& + \mathbf{b_{ij0}} + \mathbf{b_{ij1} comp_{ijt}} + b_j + \varepsilon_{ijt}
\end{aligned}
\tag{3.1}
$$

As in definition (2.2), the following independent distributions for the errors and the random effects are assumed (where n_i denotes the number of observations of branch i):

$$\varepsilon_{ij} = (\varepsilon_{ij1}, ..., \varepsilon_{ijn_i})^T \sim N_{n_i}(0, \sigma^2 I_{n_i}) \tag{3.2}$$

$$b_j \sim N(0, g_{00}^2) \tag{3.3}$$

$$b_{ij} = (b_{ij0}, b_{ij1})^T \sim N_2(0, G) \text{ with } G = \begin{pmatrix} g_0^2 & g_{01} \\ g_{01} & g_1^2 \end{pmatrix} \tag{3.4}$$

This model is not hierarchical (nested) because the random effects b_{ij0} and b_{ij1} are crossed with the fixed effects of the county variables (e.g. pop_{jt}). A hierarchical model is considered in Section 3.3.

In order to improve the model fit, the structure of the random effects is examined at first by testing whether the random effects specified in model (3.1)-(3.4) should be included as described in the introduction. Whereas the random effects for the branch level intercept (b_{ij0}) and slope of *comp* (b_{ij1}) are significant ($p-value < .0001$), the random effects for the county level intercept (b_j) were found to be not significant. They are subsequently removed from the model. Indeed according tests show that both branch level random effects stay significant.

Since the number of observations and the values of *log.dep* in each year are varying, the within-group errors might be varying for each year, too. Therefore, heterogeneous residual variances σ_t^2 for each year $t, t = 1994, \ldots, 2002$ are considered; i.e. we assume

$$\varepsilon_{ijt} \sim N(0, \sigma_t^2). \tag{3.5}$$

In order to achieve identifiability of the parameters σ_t^2, it is assumed that $\sigma_t^2 = \delta_t^2 \sigma^2$ for each year $t, t = 1994, \ldots, 2002$ and $\delta_{1994} = 1$ (see [17]). The test of this variance structure ($H_0 : \sigma_t^2 = \sigma^2$, i.e. $\delta_t = 1$ for each year t, $t = 1994, \ldots, 2002$ at the 5% level) shows that there is a significant improvement in the model fit ($p - value < .0001$).

The above variance structure can be extended further: since the observations are taken longitudinally on the same subjects, the within-group (i.e. within-branch) errors are probably autocorrelated. Considering the few time points available (9 time points for the years 1994-2002) only the first three or four lags should be considered. Because the empirical autocorrelation function from the residuals of the previous model shows that the autocorrelation of the first lag is significantly not equal to zero, an $AR(1)$ model is chosen as correlation structure. An additional moving average term is also included, i.e. we allow for

a $ARMA(1,1)$ model as correlation structure:

$$\varepsilon_{ijt} = \phi_1 \varepsilon_{ijt-1} + \theta_1 a_{t-1} + a_t, \tag{3.6}$$

where $\{a_t, t \geq 1\}$ is a zero mean white noise process with constant variance σ_a^2. Testing $H_0 : \phi_1 = \theta_1 = 0$ at the 5% level confirms the significance of this extended variance structure ($p-value < .0001$). As a result this heterogeneous autoregressive variance structure of the errors is included in the model.

Finally the model is reduced by a stepwise approach based on $t-tests$ of the fixed effects at 5% level ($H_0 : \beta_i = 0$). In the resulting model all fixed effects are significant at the 5% level or left in the model in order to maintain the hierarchical structure of the fixed effects. All significant effects are marked in the initial model formulation (3.1) using bold face and displayed with their estimated regression coefficients in Table 4.2.

Having assumed specific error distributions in this final model it has to be checked whether these assumptions are appropriate. At first, the assumptions on the within-group (i.e within-branch) errors are checked (compare (3.2), (3.5) and (3.6)), and subsequently the random effects are examined.

The within-group (i.e. within-branch) errors are assumed to have a heterogeneous autoregressive variance structure: $\varepsilon_{ijt} \sim N(0, \sigma_t^2)$ and $\varepsilon_{ijt} = \phi_1 \varepsilon_{ijt-1} + \theta_1 a_{t-1} + a_t$, where $\{a_t, t \geq 1\}$ is a zero mean white noise process with constant variance σ_a^2. Therefore the errors depend on those of the years before and on t. Since $\hat{\phi}_1 = 0.73$, one expects approximately similarly distributed standardized within-group residuals per year. In fact 93.1% of the residuals lie in the $[-2, 2]$-band, i.e. the approximate 95% confidence band. Checking the assumption of normality, QQ-plots for each year (Figure 4.3) show that the model fit is quite good for some years (e.g. in 1997), but there are also some deviations from normality (e.g. in 2001).

The final model includes random effects for the intercept and for the slope of *comp* on the branch level with distribution given by (3.4). A look at the EBLUPs of the random effects for each branch confirms the zero-mean assumption (left panel in Figure 4.4), while the marginal normality of the random effects can be investigated by QQ-plots (right panel in Figure 4.4). These show that the assumption is approximately appropriate.

The final model also contains five interactions of county variables with state variables. Among those, four are interactions with *unemp* and one with *inc.pc*. The interactions complicate the interpretation of the different effects, but include important additional information. We therefore have to examine these interactions, since we cannot interpret the

main effect solely when interaction terms are present. In order to facilitate this, Table 4.3 gives the expected deposits at different levels of the covariates: in the left part of the table each county variable is considered at its observed 25%- and its 75%-quantile (denoted by 'low' and 'high'), the state variables are taken at their respective empirical medians, while in the right part of the table it is the other way around and, for reasons of clarity, only the best four and the worst four combinations of the variables are displayed. This shows that the influence of *pop* is the strongest. The effect of *inc.pc* is also clearly positive. The interaction with *av.dep* is important to understand this overall effect: if we consider a 3D interaction plot (Figure 4.5), which displays the effect of *inc.pc* and *av.dep* on the deposits while the remaining covariates are set to their respective mean values, we see a weak positive overall influence of *inc.pc* on the deposits. The influence of *unemp* is positive, but not as strong as those of the other two county variables. 3D interaction plots of the four interactions between state variables and *unemp* can be found in [5]. In total, the effect of all three covariates together at their 75%-quantiles is a 25% increase in the deposits compared to the value with all covariates at their 25%-quantiles. These results reflect the findings of the explorative data analysis. Among the state variables, the examination of all 32 possible combinations of covariate levels showed that the influence of the *av.dep* is naturally very strong but slightly weaker than the effect of *branch.total*. The effect of *dep.total* is fairly negative though. Furthermore, the effect of $no.fail$ is weakly negative and the effect of $mshare$ is moderately positive.

The model fitted above is useful, but may not be correct. As a next step we investigate if the linear mixed model is an improvement in the model fit or if a standard linear model is sufficient to model the influences on *log.dep*. The comparison to a linear model with the same fixed effects shows that the AIC of the mixed model is much smaller: 531 vs. 7937 of the linear model. Now one could argue that this is possibly because the variance structure of the linear model is not as sophisticated as in the mixed model. However a so-called generalized least squares (GLS) model which allows to fit heteroscedastic and correlated within-group errors (but no random effects) also has a larger AIC than the mixed model: 531 vs. 3675 of the GLS model. Thus the additional fitted random effects in a mixed model, modeling the variability and dependency, lead to a considerable improvement in the model fit.

3.3 Hierarchical model Since the model developed in Section 3.2 is not hierarchical, it is of interest to consider a hierarchical model as well. To do this the branch and county variables have to be averaged over time in order to eliminate the time effect, i.e. we define $av.comp_{ij} = \frac{1}{\text{No. of obs. of branch } i} \sum_t comp_{ijt}$ and $av.pop_j = \frac{1}{9} \sum_t pop_{jt}$ (analogously for *inc.pc* and *unemp*). With these adjusted variables the following three-stage model can be fit: first, the time effects are represented by the state variables.

$$\begin{aligned} log.dep_{ijt} = \alpha_{ij0} + \alpha_{ij1} no.fail_t + \alpha_{ij2} mshare_t + \alpha_{ij3} branch.total_t \\ + \alpha_{ij4} dep.total_t + \alpha_{ij5} av.dep_t + \varepsilon_{ijt} \end{aligned} \tag{3.7}$$

Second, the intercepts and slopes depend on branch-specific effects. A random effect is included for the intercept.

$$\begin{aligned} \alpha_{ij0} &= \gamma_{0j0} + \gamma_{1j0} av.comp_{ij} + b_{ij} \\ \alpha_{ijk} &= \gamma_{0jk} + \gamma_{1jk} av.comp_{ij} \qquad\qquad k = 1, \ldots, 5 \end{aligned}$$

Third, the county-specific effects are modeled:

$$\begin{aligned} \gamma_{0j0} &= \delta_{000} + \delta_{010} av.pop_j + \delta_{020} av.inc.pc_{jt} + \delta_{030} av.unemp_{jt} + b_{0j} \\ \gamma_{1j0} &= \delta_{100} + \delta_{110} av.pop_j + \delta_{120} av.inc.pc_{jt} + \delta_{130} av.unemp_{jt} + b_{1j} \end{aligned}$$

$$\gamma_{0jk} = \delta_{00k} + \delta_{01k} av.pop_j + \delta_{02k} av.inc.pc_{jt} + \delta_{03k} av.unemp_{jt}, \quad \gamma_{1jk} = \delta_{10k}, \quad k = 1, \ldots, 5$$

We assume that errors are distributed as in (3.2). Similar to (3.3) and (3.4) the following independent distributions of the random effects are assumed (where n_i denotes the number of observations of branch i):

$$\begin{aligned} b_{ij} &\sim N(0, g_{00}^2) \\ b_j = (b_{0j}, b_{1j})^T &\sim N_2(0, G) \text{ with } G = \begin{pmatrix} g_0^2 & g_{01} \\ g_{01} & g_1^2 \end{pmatrix} \end{aligned}$$

Note that random effects are included only for γ_{0j0} and γ_{1j0} to reduce the model complexity. The coefficients $\gamma_{1j1}, \ldots, \gamma_{1j5}$ are modeled only by an intercept because third-order interactions are not considered.

All these equations can be substituted into (3.7) and then the initial hierarchical model can also be written in the standard notation of linear mixed models (2.1). The resulting model is similar to the non-hierarchical one: it incorporates the same main effects and

interactions (substitute *av.pop* for *pop*, etc.) but two random effects on the county level and just one on the branch level.

A similar model analysis as in Section 3.2 shows that random effects are only needed for the branch level intercept and that a heterogeneous autoregressive variance structure for the errors is also appropriate (for more details see [5]). Finally using a stepwise approach based on $t - tests$ at the 5% level, the following main effects and interactions are kept in the model: an intercept and all branch, county and state variables except for $no.fail$ as well as the interactions $av.comp \times mshare$, $av.pop \times mshare$, $av.unemp \times mshare$, $av.unemp \times branch.total$ and $av.unemp \times dep.total$.

Model diagnostics for this final hierarchical model show that the distributional assumptions on the errors are probably not accurate (QQ-plots show clear deviations from normality), but nevertheless the residuals show some good characteristics. The distributional assumptions on the random effects are approximately appropriate.

3.4 Prediction In order to compare and evaluate the two models, the predictive capability is checked by taking the following approach: the final hierarchical and non-hierarchical models are estimated with the data of 1994 to 2001, i.e. the same fixed and random effects as well as the same variance structures for the errors as in the respective final models are chosen, but the time period is restricted. Then the values of 2002 are predicted using these restricted models. Certainly, the estimated parameters of the models change, but no new model reduction and diagnostics are performed, since the restricted models are supposed to represent the unrestricted ones (note that this is no complete cross-validation but a computationally less demanding approach).

A comparison of predicted and observed values of 2002 (Figure 4.6) shows that the predictive capability of the non-hierarchical model is quite good and better than that of the hierarchical model which underestimates the observed values. The sum of squared residuals of the hierarchical model is much larger: 126 vs. 18 of the non-hierarchical model. This reflects the loss of time information when variables are averaged over time.

3.5 Interpretation of results from the statistical analysis Since the previous analyses showed that the non-hierarchical model's predictive capability is superior to that of the hierarchical model, the following interpretation regarding the influences on the de-

posits of a bank branch is mainly based on the results of the non-hierarchical mixed model as developed in Section 3.2. Nevertheless, the hierarchical model allows for an easy representation of the effects in the hierarchical three-stage structure (see [5]).

The examination of the data showed that the deposits of a bank branch significantly depend on geographic effects such as local wealth and local competition as well as on bank-performance effects on the macroeconomic level. As expected an increase of market share, the average deposit per bank and the share of the number of branches in NY compared to the USA have a positive influence on the bank branch deposits. Only the effect of the share of branches is somewhat surprising. Perhaps an increase in the number of branches in NY compared to the USA means enforced marketing activities in NY, i.e. the bank concentrates on its branches in NY. At the same time, there are negative influences of the number of branches that closed during a year and the share of the total deposits in NY compared to the USA. While the first effect is easy to explain (closures of branches are probably a result of a bad market environment), the second effect is not. Note that these effects interact with geographic effects and that a useful statistical model aimed at determining a good locality for bank branches should include these variables for statistical control.

Geographically the deposits depend positively on the county's population and on the per capita income, since it is obvious that there are more deposits if there are more people and if people earn more. However the overall effect of the unemployment rate is unclear, since it interacts with other effects. Obviously unemployed people have less cash flow and therefore one might expect less deposits in an area with high unemployment, but people, who recently lost their job, possibly save more money in the short term because of the financial insecurity of the next months and who live in an area with an increasing unemployment rate, might also save more money because they feel threatened by unemployment as well and thus they want to be financially prepared.

Likewise there is no uniform influence of the local competition on bank deposits, since there are probably opposing trends if the competition increases: on the one hand, competition stimulates business and if for example the population in an area increases, the number of branches in that area increases in order to get new customers and more deposits. On the other hand, if there is more competition, each branch cannot have as much deposits as if there were less branches. This shows that it would be too easy just to give competition effects a negative sign as most might expect it.

To investigate these influences the following approach is helpful: all branches are classified as being in a 'rural'/'urban', 'poor'/'rich' area with a low/high unemployment rate (the classification is done using the empirical medians of the respective variables). For all branches in a specific area the branch-specific effects of the competition are averaged and then compared.

The comparison of branches in rural and urban areas shows that the effect of the competition is slightly stronger in rural areas than in urban areas. A possible explanation for this observation is that the market environment in rural areas is less developed. Thus a higher competition might have a stronger impact than in urban areas, since e.g. marketing activities can influence people to a greater extent, while urban people are more used to such activities. They have a broad choice of banks and have chosen their bank deliberately.

Comparing the effect of the competition in rich and poor areas it can be said that rich people, who have more deposits than poor people anyway, have more deposits if there is more competition, i.e. they possibly think more about where to put their money and like to chose their bank deliberately. More competition in a rich area therefore increases the deposits in a branch. This effect is much smaller in poor areas because people living there do not have much more money to increase their deposits. They are happy if they have some deposits at the bank and do not care much about which bank it is.

At last, the comparison of the competition effects in areas with high and low unemployment rates shows that people in areas with a high unemployment rate more strongly increase their deposits if the competition increases than people in areas with a low unemployment rate do. This can possibly be explained by the fact that unemployed people or people that are threatened by unemployment are much more worried about their money than employed people are. Therefore such people are easier to be influenced by marketing activities and new offers which may be a result of an increasing competition, while employed people worry less about small differences in offers in order to earn a few cents or dollars more of interest.

Besides these dependencies there is an additional branch-specific effect: some branches have more deposits than others if all other influences are disregarded. This can be explained by specific characteristics of a branch such as a long-term customer loyalty or a particular good location in an area.

In the Figure 4.7 one can see the influence of the branch-specific effects for four randomly chosen branches from Rockland (533, rural and poor area with a low unemployment rate),

Suffolk (657, urban and poor area with a low unemployment rate), Nassau (5052, rural and rich area with a low unemployment rate) and New York (435, urban and rich area with a high unemployment rate).

4 Summary and discussion Our approach illustrates the potential of linear mixed models in the context of measuring branch performance and deciding about the location of new branches. Compared to DEA, regression analysis indicates directly causes of low performance, gives performance information about *all* branches in the sample and can be used to forecast deposits of new branches [4]. It therefore allows an easy evaluation of a single existing branch and of the potential of a new location. In contrast to [1] and [4] we include interactions and random effects in our models in order to take into account different local market environments and thus make the model more reliable. On the one hand, geographic effects such as local wealth (as measured by county unemployment rates and county income per capita) and local competition are found to significantly influence the branch performance, while, on the other hand, bank-performance effects on the macroeconomic level such as the number of branches that close and the bank's market share also have to be considered in order to asses the performance accurately.

The specific performance measure of deposits is easily available for a statistical analysis. Since the study is longitudinal, it also considers new business of a branch to some extent. However, the flexibility of the mixed-effects regression model easily allows for the use of other performance variables such as fee income or the number of new deposit and/or lending accounts. Multiple performance measures at once could be included e.g. by a weighted sum of these measures with weights possibly determined by banking executives. More independent variables such as in-branch variables or other competitive situation features could also be included in the modeling but would further increase the computational complexity. A detailed explorative data analysis is therefore crucial to identify potential random effects and interactions before fitting an initial model. Now that we know the aptness of the mixed-effects regression method, a more computationally intense effort could be utilized to model the bank wealth on a holistic set of variables reflecting the banking environment.

Acknowledgement C. Czado is supported by the DFG (German Research Foundation) grant CZ 86/1-3. We like to thank Jon Kerr for sharing data and knowledge in the banking

industry.

REFERENCES

[1] Avkiran, N. K. (1997). Models of retail performance for bank branches: predicting the level of key business drivers. *International Journal of Bank Marketing*, **15(6)** 224–237.

[2] Avkiran, N. K. (1999). An application reference for data envelope analysis in branch banking: helping the novice researcher. *International Journal of Bank Marketing*, **17(5)** 206–220.

[3] Berger, A. N. and Humphrey, D. B. (1997). Efficiency of financial institutions: International survey and directions for future research. *European Journal of Operational Research*, **98(2)** 175–212.

[4] Boufounou, P. V. (1995). Evaluating bank branch location and performance: A case study. *European Journal of Operational Research*, **87** 389–402.

[5] Brechmann, E. C. (2009). *Linear mixed models applied to bank branch deposit data*. Project, http://www-m4.ma.tum.de/diplarb/projekt_brechmann.pdf, Zentrum Mathematik, Technische Universität München, Garching bei München.

[6] Card, D. (1999). *The Causal Effect of Education on Earnings*, in Ashenfelter, O. and Card, D. (eds.), Handbook of Labor Economics, Vol 3, Elsevier, Amsterdam.

[7] Chelst, K. R., Schultz, J. P. and Sanghvi, N. (1988). Issues and decision aids for designing branch networks. *Journal of Retail Banking*, **10(2)** 5–17.

[8] Davenport, T. O. and Sherman, H. D. (1987). Measuring branch profitability. *The Bankers Magazine*, **170(5)** 34–38.

[9] DeYoung, R., Hunter, W. C. and Udell, G. F. (2004). The past, present, and probable future for community banks. *Journal of Financial Services Research*, **25** 85–133.

[10] Doyle, P., Fenwick, L. and Savage, G. P. (1979). Management planning and control in multi- branch banking. *Journal of Operational Research Society*, **30(2)** 105–111.

[11] Ferrier, G. D. and Lovell, C. A. K. (1990). Measuring cost efficiency in banking: Econometric and linear programming evidence. *Journal of Econometrics* **46** 229–245.

[12] Fox, J. (2002). *Linear Mixed Models: Appendix to An R and S-SPLUS Companion to Applied Regression*. http://cran.r-project.org/doc/contrib/fox-companion/appendix-mixed-models.pdf.

[13] Gart, A. (1994). *Regulation, Deregulation, Reregulation*. Wiley, New York.

[14] Jayaratne, J. and Strahan, P. E. (1996). The finance-growth nexus: evidence from bank branch deregulation. *Quarterly Journal of Economics*, **111** 639–670.

[15] Kim, M. and Vale, B. (2001). Non-price strategic behavior: the case of bank branches. *International Journal of Industrial Organization*, **19** 1583–1602.

[16] Mostafa, M. (2007). Modeling the efficiency of GCC banks: a data envelopment analysis approach. *International Journal of Productivity and Performance Management*, **56(7)** 623–643.

[17] Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer, New York.

[18] Ramanathan, R. (2003). *An introduction to data envelopment analysis: a tool for performance measurement*. Sage Publications, New Delhi.

[19] Sherman, H. D. and Gold, F. (1985). Bank branch operating efficiency: Evaluation with data envelopment analysis. *Journal of Banking and Finance*, **9** 297–315.

[20] Shiers, A. F. (2002). Bank branching, economic diversity and bank risk. *The Quarterly Review of Economics and Finance*, **42** 587–598.

[21] Stram, D. and Lee, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50** 1171–1177.

[22] West, B. T., Welch, K. B. and Galecki, A. T. (2006). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall/CRC, Boca Raton.

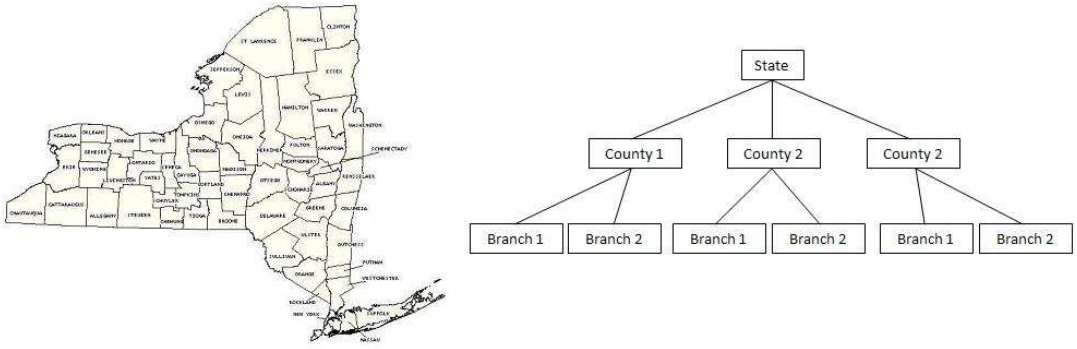

Figure 4.1: Map of New York State and the hierarchical structure of the data.

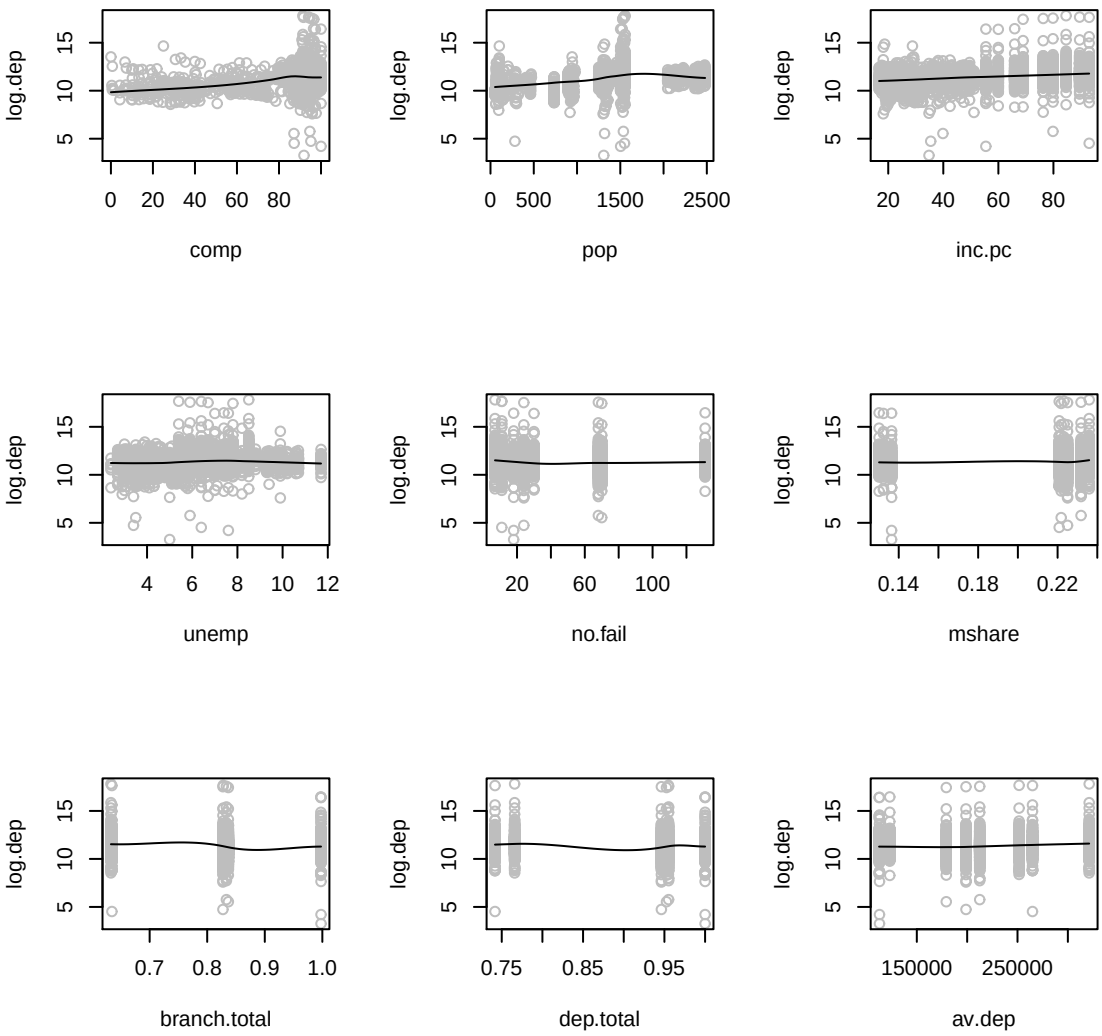

Figure 4.2: Scatter plots with lowess smoothed curves of *log.dep* against all nine covariates.

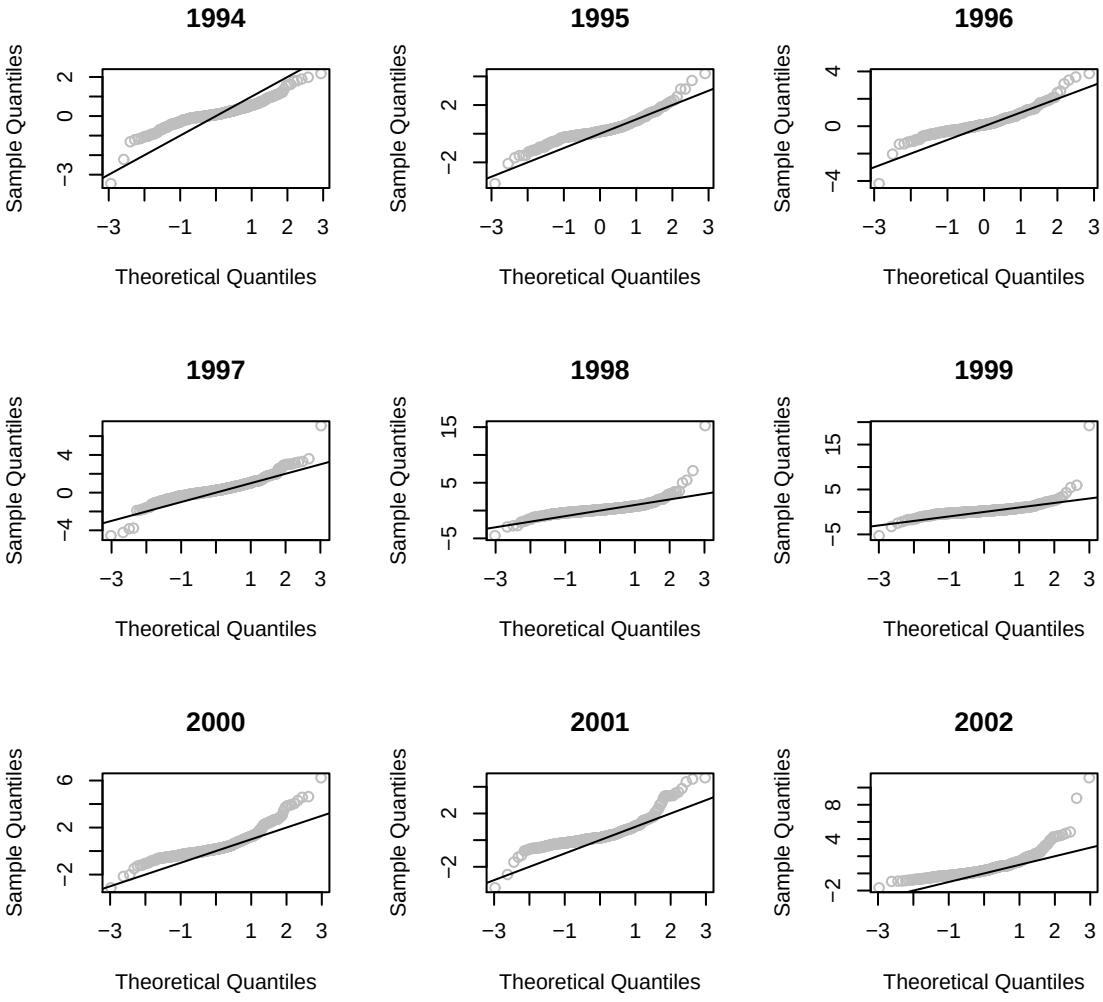

Figure 4.3: QQ-plots of the standardized residuals of the final model.

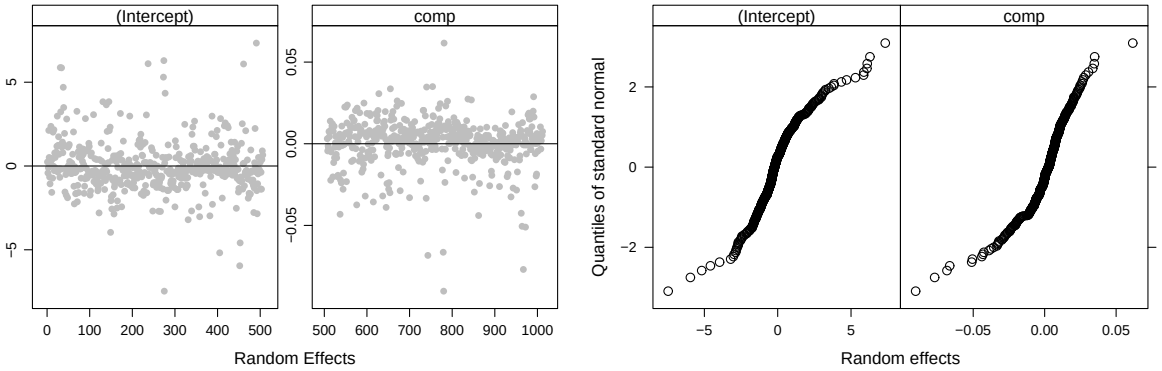

Figure 4.4: EBLUP's $(\hat{b}_{ij0}, \hat{b}_{ij1})$ and their QQ-plots.

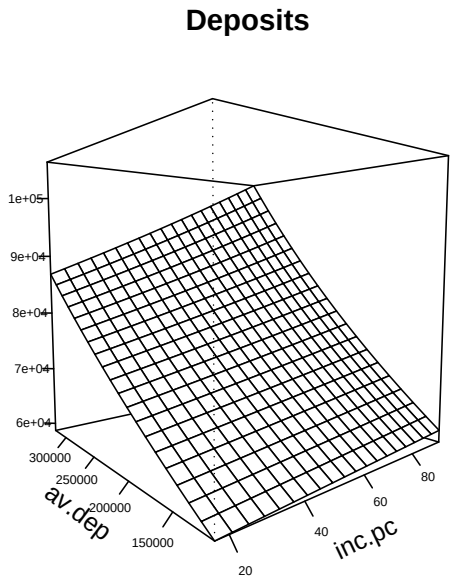

Figure 4.5: 3D interaction plot to assess joint influence of *inc.pc* and *av.dep* on *log.dep*, when all other variables are taken at their average value for the final model.

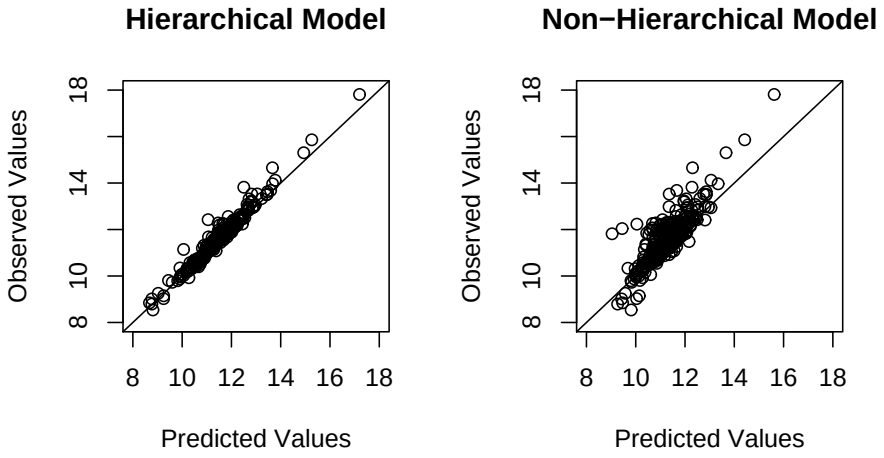

Figure 4.6: Comparison of the predicted values of the final models.

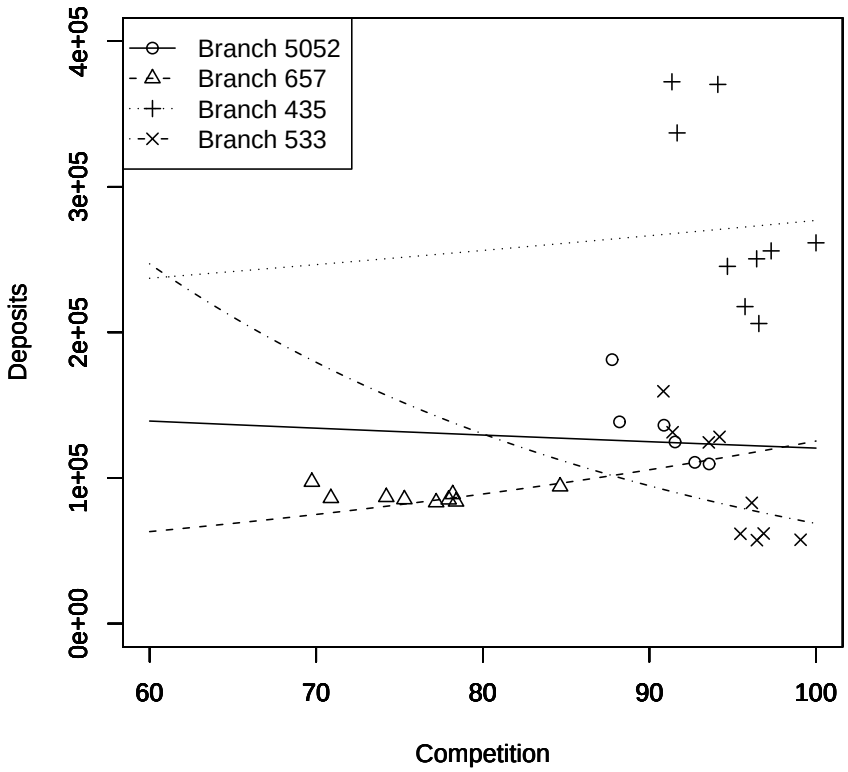

Figure 4.7: Relationship between deposits and competition for four randomly chosen branches (observed and estimated values).

Level	Name	Description
State	$no.fail$	number of branches that closed in NY during the year
	$mshare$	market share in NY
	$branch.total$	share of the number of branches in NY compared to the USA
	$dep.total$	share of the total deposits of the bank in NY compared to the USA
	$av.dep$	average deposit per bank in NY
County	pop	population in the county (in 1000)
	$inc.pc$	per capita income (in 1000)
	$unemp$	unemployment rate in the county
Branch	$branch$	branch identity number (constant over the years)
	$log.dep$	total deposits (in USD) in the branch in log form
	$comp$[1]	measure of geographical competition of the branch (different for each year; values between 0 and 100, where a value of 100 is an indication of a high geographical competition)

Table 4.1: Variable description classified by state, county and branch level.

[1]The sum of all distances between the branch and all branches of other banks which have only one single branch or multiple branches, respectively, are given by the variables $SingleDensity$ and $MMCDensity$. Since these variables are not easy to interpret and highly correlated (94%), they are merged, standardized by their medians, and scaled in order to have values between 0 and 100: $a = \frac{SingleDensity}{\text{median}(SingleDensity)} + \frac{MMCDensity}{\text{median}(MMCDensity)}$, $b = a - \min(a)$, and finally $comp = (1 - \frac{b}{\max(b)}) \cdot 100$.

Variable	Estimate	Std. Error	p-value
Intercept	1.12 E+1	4.41 E−1	0.0000
pop	5.77 E−4	8.83 E−5	0.0000
inc.pc	−7.32 E−4	1.44 E−3	0.6121
unemp	−2.69 E−1	6.91 E−2	0.0001
no.fail	5.50 E−4	2.95 E−4	0.0624
mshare	−6.23 E+0	2.23 E+0	0.0054
branch.t	−3.90 E+0	1.80 E+0	0.0303
dep.total	3.44 E+0	1.68 E+0	0.0410
av.dep	1.70 E−6	2.87 E−7	0.0000

Interact.	Estimate	Std. Error	p-value
unemp× *no.fail*	−1.04 E−4	4.39 E−5	0.0184
unemp× *mshare*	1.37 E+0	3.67 E−1	0.0002
unemp× *branch.t*	1.05 E+0	2.82 E−1	0.0002
unemp× *dep.total*	−9.43 E−1	2.66 E−1	0.0004
inc.pc× *av.dep*	1.05 E−8	3.94 E−9	0.0078

Table 4.2: Significant effects with their estimates, standard errors and p-values in the final model ($branch.total = branch.t$) ($\alpha = 0.05$).

pop	*inc.pc*	*unemp*	Deposits
high	high	high	77,368
high	high	low	75,756
high	low	high	73,061
high	low	low	71,539
low	high	high	66,975
low	high	low	65,580
low	low	high	63,248
low	low	low	61,930

no.f	*msh*	*branch.t*	*dep.t*	*av.dep*	Deposits
low	high	high	low	high	139,487
high	high	high	low	high	139,043
low	low	high	low	high	118,323
high	low	high	low	high	117,947
low	high	low	high	low	58,184
high	high	low	high	low	57,999
low	low	low	high	low	49,356
high	low	low	high	low	49,199

Table 4.3: Expected deposits at different levels of the county and state variables, respectively, using the abbreviations $no.fail = no.f$, $mshare = msh$, $branch.total = branch.t$ and $dep.total = dep.t$.