industries" non-systematic, cluster specific changes in business default risk are of comparable size. Somewhat bigger cluster specific shocks to default risk are experienced in the retail industry and the "public and subsidized" sector. The industry with by far the largest cluster specific error variance is the "real estate and finance" cluster, with a variance of 0.411. Since the variance of the unsystematic part is fixed at $\sigma_i^2 = 1.645$ for our choice of specification in eq. (5), the implied correlation is 0.2. Column 1 shows that the estimated variance if equality across clusters is assumed, is 0.254. However, a likelihood-ratio test easily rejects the restriction of a common correlation across all clusters is strongly rejected: $LR = 122.7 > \alpha_{0.05}\left(\chi^2_{(6)}\right)$. The clusters should be thus be examined separately, and not as one group. Column three shows that including industry-specific parameters for the firm-specific variables has hardly any effect on the size of the industry correlations. Moreover, column six illustrates that the way in which industry clusters are defined has little or no impact on the quantitative importance of industry specific correlations. Although the relative size of the correlations for the optimal clusters is somewhat different, their overall impact is broadly consistent with the estimates in column two.

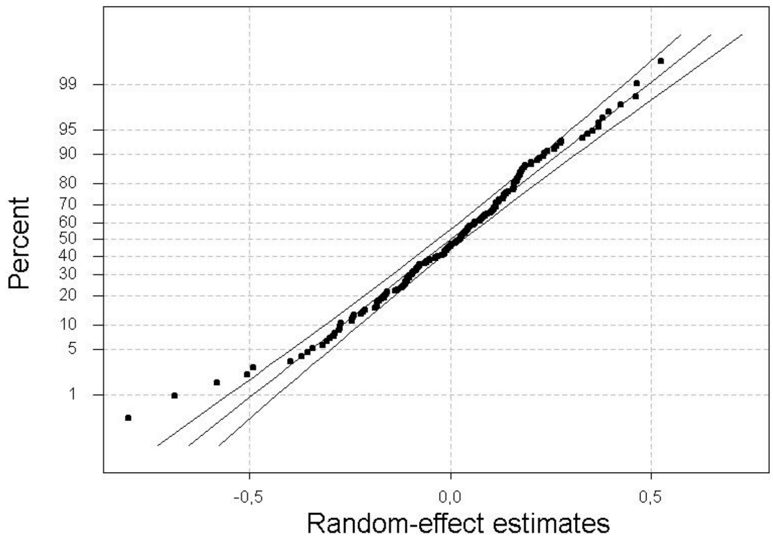
Finally, columns four and five help us understand better what kind of variation in the default rate industry correlations capture. The exclusion of macro explanatory or firm-specific variables from the model has a strongly varying effect on the size of our correlation estimates. Excluding both simultaneously (column five) has virtually no impact on the estimates for the wholesale and construction industries, while other industries, particularly the public/subsidized sector and real estate & finance register substantially larger correlations in the absence of these explanatory variables. This supports our view that industry correlations capture industry-specific effects, and not economy-wide fluctuations in the default frequency. Re-including the macro variables into the model (column four) further strengthens this view, since this once more shows that the impact on the size of the estimated industry correlations varies substantially across industries.

Although we have no formal explanation of the differences between clusters, there may be some heuristic explanations. The financial sector and real estate sector, were exposed to fundamental changes during the 1990s, after financial markets had been deregulated during the late 1980's and early 1990's. In the early 1990's Sweden also experiences a banking crisis, the effects of which may have been dissipated only slowly. The deep recession that struck Sweden more or less simultaneously also had a big impact on the government budget and its contributions to various (semi-)governmental organizations.

Despite there being good theoretical reasons for the inclusion of close business relations within certain industries into models of default or credit risk [56][24], economic theory does not help us much in forming a prior for the size of the industry specific error variances. We presume that an (unconditional) correlation of 0.5 - the equivalent to a cluster variance of 1.645 - or even higher are plausible and that most of such correlation will be related to variation in macroeconomic variables.

Figure 3: A QQ plot for the estimated random effect components.

The figure compares the empirical distribution of estimated random-effect components of the model in Table 3, column 2, to the normal distrbution. The scale of the vertical axis is determined by the dispersion of observations for the normal distribution The actual empirical distribution is depicted by dots. If the empirical distribution were to follow the normal curve exactly, the dots would overlap with the straight line in the plot. The concave/convex lines provide 95 % confidence limits.




When we compare the three models, the most substantial increase in cluster variances comes about when the systematic part is excluded completely (column 4) - although the exact magnitude of the effect varies somewhat between industries. The industry errors now capture part of the variation in the default rate that macro - and firm specific covariates can no longer explain. For obvious reasons, we would like to know if and to what extent this result is solely a consequence of a common impact that the macro-variables have on companies. To cast more light on this issue, we re-estimated the model and included the macro variables while keeping out the other explanatory variables. In the third column of Table 3, the estimates of the cluster variances with this specification are displayed. Once again, the effect varies between sectors, with the macro variables being able to explain more of the cluster variances for the "government" sector and "real estate and finance". Overall, including macro variables and excluding firm specific covariates reduces the estimated variances somewhat, but does not bring them back to the level of the model with a full systematic component. Hence, we find that a substantial part of the difference in the conditional and the unconditional correlations should be attributed to the firm specific variables in the systematic part.

In the estimation procedure, it was assumed that the random-effect component, i.e. $\varepsilon_k^\tau$, follows the normal distribution with expectation equal to zero. To check the validity of this assumption, we re-scaled the estimated random-effects to let them have a unit standard deviation by dividing the estimates of the 175 random-effects components by the relevant cluster variance. Under the above mentioned assumption the standardized estimates should in fact follow the standard Normal distribution. Figure 3 shows, by means of a Q-Q-plot, to what extent the distribution of the standardized estimates match the percentiles of the standard Normal distribution. If the correspondence between these two is perfect, the points will be placed along the straight line that lies in between the accompanying lines that reflect the upper and lower 95% confidence limits. The Q-Q-plot suggests that normality assumption is quite acceptable.

From the discussion in Section 2 we recall that the cluster specific shocks may well be persistent and that one thus may observe patterns of autocorrelation, for example an AR(1) structure as outlined in eq. (3). Therefore, we also used the estimated random-effects to check for the presence of any autoregressive structure in the correlations, by examining the autocorrelation function for the series of 25 observations we have available for each cluster. The strongest (and in fact only) sign of such a pattern was found for the "retail" industry, that has a first-order autocorrelation of about 0.25. However, even this estimate was insignificant and we thus conclude that no autoregressive structure is present in this data.[19]

### 4.2 Estimating portfolio credit risk distributions and VaR

Having estimated a duration model of default risk with cluster specific effects, we now turn to applying this model to the estimation of a portfolio credit loss distribution. In order to get a better understanding of the extent to which our modeling approach leads to improved estimates of credit loss distributions, we will compare the Value-at-Risk (VaR) estimates, the statistics we use to summarize the loss distribution, with those derived from two simple benchmark models.

We calculate VaR based on the portfolio at $\tau$ for a one quarter ahead horizon by mean of the following algorithm:

1. Let $\widehat{\beta}$ and $\widehat{\sigma}_{\varepsilon_k}$ be consistent estimates of $\beta$ and $\sigma_{\varepsilon_k}$ in the model in equation (2).[20]

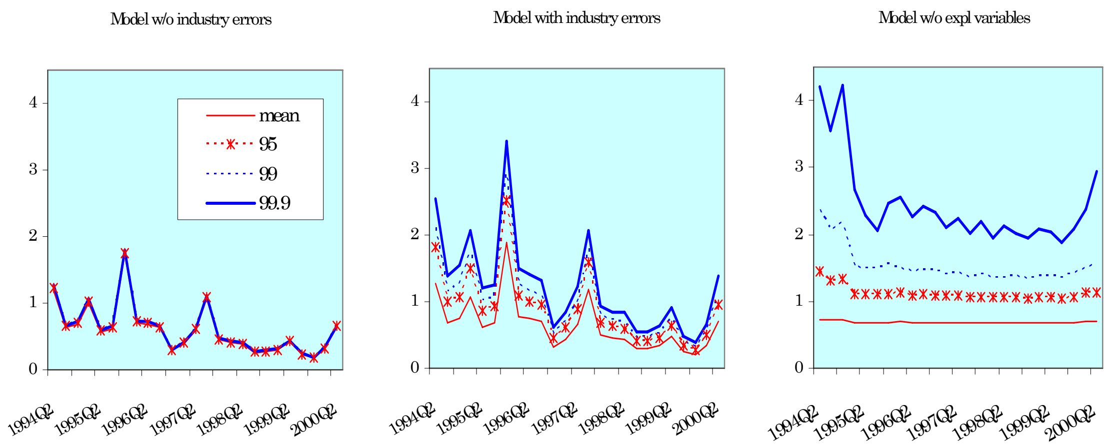
2. Draw $\varepsilon_k^\tau$ from the normal distribution with expectation equal to zero and variance equal

[19] This approach does not estimate the autoregressive part jointly with other parameters in the model. The GLIMMIX macro in SAS provides this option but the implementation failed due to lack of memory. The procedure of estimating the $AR(1)$-part, conditional on the estimates for $\beta$, $\sigma_i^2$ and $\sigma_k^2$, gave no indication of an auto-regressive structure in the data. Most likely, a joint estimation procedure would have produced similar results, and thus we did not pursue the effort to solve this technical problem.

[20] Here we assume that the model has been estimated without an autoregressive part. In the case of unequal variance it is necessary to adjust (i) such that the draws from the normal-distribution is taken with the appropriate standard deviation for the cluster. If there is an autoregressive part in the model, then the draws in (i) should be taken from the corresponding autoregressive process.

Figure 4: Percentiles of the default rate distributions implied by three models.

The boxes display the mean and three percentiles from the credit loss distributions that have been generated using three different models of default risk. In Box 1, the default rate distribution has been generated with a conventional duration model. A cluster specific error structure has been added to arrive at the model used in Box 2 . In Box 3 all explanatory variables have been deleted from the cluster error model.




to $\widehat{\sigma}_k^2$

3. Draw $\varepsilon_i^\tau$ from either the standard normal distribution or the extreme-value distribution (depending on the estimated model)

4. Repeat steps (2) and (3) for $k = 1, ..., K$

5. Calculate $\widehat{\pi}_i^\tau = \Phi\left[x_i(\tau)\widehat{\beta} + \varepsilon_i^\tau + \varepsilon_k^\tau[i \in k] + \gamma_t\right]$ or $\widehat{\pi}_i^\tau = 1 - \exp\left[-\exp\left(x_i(\tau)\widehat{\beta} + \varepsilon_i^\tau + \varepsilon_k^\tau[i \in k] + \gamma_t\right)\right]$ (depending on the estimated model)

6. Let $Loss^\tau = \sum_{i=1}^{n^\tau} \widehat{\pi}_i^\tau s_i^\tau / \sum_{i=1}^{n^\tau} s_i^\tau$, where $s_i^\tau$ is the size of loan $i$ at time $\tau$ and $n^\tau$ is the number of loans in the loan portfolio at time $\tau$.

7. Repeat (2)-(5) $R$ times and order the $R$ observations of $Loss^\tau$ by increasing size, where $R$ should be large enough to guarantee that the loss distribution has converged.

8. Let $VaR_z^\tau$ equal the z-percentile in the distribution of $Loss^\tau$

Because our statistical model does not incorporate a description of (any possible relation between default risk and) the loan size or recovery rates, a comparison of *loss* distributions will not strictly give us an evaluation of the suitability of our statistical model for generating portfolio *credit loss* distributions. Sample specific, non-stochastic variation in the loan size may add noise to the portfolio *default rate* distribution generated by the model. For this reason, we start out by studying how the distribution of quarterly *default* rates is affected by using model (2) instead of a conventional specification. We can do this by following the procedure described in steps (1)-(7) above, but replacing the $Loss^{\tau}$ rate by a $defaultrate^{\tau} = \sum_{i=1}^{n^{\tau}} \widehat{\pi}_i^{\tau}$ in step (5). This will give us a purer picture of the merits of the model specification. Having done this, we

**Table 4: The effect of including industry specific errors on the estimation of default risk distributions.**
Table shows the mean and various percentiles of the portfolio default risk distribution that is generated by (i) the model with industry specific common errors, and (ii) the model without common industry specific errors. Normally styled figures refer to the model with industry specific errors, italics to the model witout industry errors.

| | | Simulated portfolio default rates at distribution percentiles | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Quarter | mean | *mean* | 95 | *95* | 97.5 | *97.5* | 99.9 | *99.9* |
| 1994Q2 | 1,28 | *1,19* | 1,81 | *1,23* | 1,95 | *1,23* | 2,55 | *1,27* |
| 1994Q3 | 0,69 | *0,63* | 0,99 | *0,65* | 1,06 | *0,66* | 1,39 | *0,67* |
| 1994Q4 | 0,74 | *0,68* | 1,07 | *0,70* | 1,16 | *0,71* | 1,54 | *0,73* |
| 1995Q1 | 1,07 | *0,99* | 1,50 | *1,03* | 1,61 | *1,03* | 2,07 | *1,06* |
| 1995Q2 | 0,62 | *0,56* | 0,87 | *0,58* | 0,94 | *0,59* | 1,21 | *0,60* |
| 1995Q3 | 0,68 | *0,62* | 0,93 | *0,64* | 0,99 | *0,65* | 1,24 | *0,66* |
| 1995Q4 | 1,88 | *1,72* | 2,53 | *1,76* | 2,71 | *1,77* | 3,41 | *1,79* |
| 1996Q1 | 0,78 | *0,70* | 1,08 | *0,72* | 1,16 | *0,72* | 1,49 | *0,73* |
| 1996Q2 | 0,74 | *0,68* | 1,01 | *0,70* | 1,08 | *0,70* | 1,41 | *0,72* |
| 1996Q3 | 0,69 | *0,63* | 0,95 | *0,64* | 1,02 | *0,64* | 1,31 | *0,65* |
| 1996Q4 | 0,32 | *0,29* | 0,45 | *0,30* | 0,48 | *0,30* | 0,62 | *0,30* |
| 1997Q1 | 0,44 | *0,39* | 0,61 | *0,41* | 0,65 | *0,41* | 0,85 | *0,42* |
| 1997Q2 | 0,65 | *0,59* | 0,88 | *0,61* | 0,94 | *0,61* | 1,22 | *0,62* |
| 1997Q3 | 1,17 | *1,07* | 1,58 | *1,09* | 1,68 | *1,10* | 2,06 | *1,11* |
| 1997Q4 | 0,50 | *0,45* | 0,69 | *0,46* | 0,74 | *0,46* | 0,94 | *0,47* |
| 1998Q1 | 0,46 | *0,41* | 0,63 | *0,42* | 0,67 | *0,43* | 0,85 | *0,43* |
| 1998Q2 | 0,43 | *0,38* | 0,59 | *0,39* | 0,64 | *0,40* | 0,84 | *0,40* |
| 1998Q3 | 0,29 | *0,26* | 0,40 | *0,27* | 0,43 | *0,27* | 0,55 | *0,27* |
| 1998Q4 | 0,30 | *0,27* | 0,41 | *0,28* | 0,43 | *0,28* | 0,55 | *0,29* |
| 1999Q1 | 0,33 | *0,30* | 0,45 | *0,30* | 0,48 | *0,31* | 0,64 | *0,31* |
| 1999Q2 | 0,47 | *0,42* | 0,64 | *0,43* | 0,68 | *0,44* | 0,91 | *0,44* |
| 1999Q3 | 0,26 | *0,23* | 0,35 | *0,23* | 0,37 | *0,24* | 0,47 | *0,24* |
| 1999Q4 | 0,19 | *0,17* | 0,27 | *0,18* | 0,29 | *0,18* | 0,38 | *0,18* |
| 2000Q1 | 0,35 | *0,31* | 0,49 | *0,32* | 0,52 | *0,32* | 0,67 | *0,33* |
| 2000Q2 | 0,70 | *0,63* | 0,96 | *0,65* | 1,04 | *0,65* | 1,38 | *0,66* |

will calculate the credit loss distributions, to complete the illustration of how the model could be used by banks for in risk-management applications and by regulatory authorities for monitoring purposes.

Figure 4 displays, for each of the 25 quarters in the sample period, the mean and three commonly used percentiles from the default rate distribution generated by three different models of firm default risk. Table 4 provides some further numerical details on the default rate distributions.[21]

[21] It should be noted that the calculation of VaR presumes that $\sigma_{\varepsilon_k} = \widehat{\sigma}_{\varepsilon_k}$, which of course is susceptible to

The first box contains the distribution that results when model (4) is used, the middle box the distribution for our model with cluster errors, and the percentiles in the third box are derived using a cluster error model without any explanatory variables. A first glance at the three boxes reveals both similarities and important differences. Both the first and second model appear to capture the broad movements in default rates over the sample period. The peaks in 1995 and 1997 are clearly captured and even the general downward trend up to the end of 1999 is identified and the mean default rate is more or less identical for models 1 and 2. The model without industry specific errors has default rate distributions, for each point in time, that are highly concentrated though. The lower and upper percentiles almost coincide, while for the model with industry specific errors, default rate percentiles are at least .5 percentage points and up to 1.6 percentage points higher. Note that the mean and higher percentiles in the first box of Figure 4 are closer to each other than is common in typical estimates of VaR. This is proximity explained by the fact that we calculate loss distributions conditional on the outcome of $x_i(\tau)$ and the point estimate of $\widehat{\beta}$, whereas many applications of VaR compute loss distributions unconditionally.and thus incorporate uncertainty about $x_i(\tau)$. Uncertainty about $\widehat{\beta}$ is typically not taken into account. The distributions that underly boxes 2 and 3 are produced in the same way. Although it would be easy to adjust our estimated loss distribution for this factor uncertainty, we prefer to isolate the correlation effect from factor uncertainty. Allowing for factor uncertainty would only shift the curves in each box upward by the same percentage, since $x_i(\tau)$, $\varepsilon_i^\tau$, and $\varepsilon_k^\tau$ are independent of each other.

When contrasted with the work by other authors, there are both differences and similarities. Duffie et al. [21] find that the impact of their unobserved heterogeneity factor on the *average* default rates of U.S. corporations is roughly a factor of 2. In this paper, industry dependencies create big effects in the *tails* of the default rate distribution but no effects on the *mean* default rate. This difference between the approaches has a theoretical and an empirical two source. The theoretical difference lies in the way correlation is approached. The unobserved heterogeneity approach reflects both variation that could potentially be picked up by including additional risk factors *and* clustering of defaults that cannot be captured by a Poisson model. The random effects approach captures simultaneous clustering of defaults, without affecting the firm-specific default risk predictions.

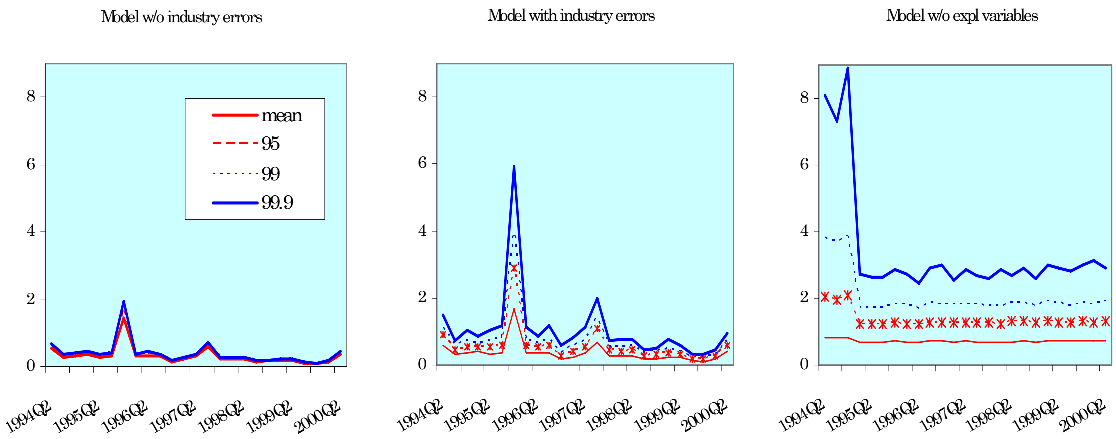
The empirical difference lies in the fact that Duffie et al. use a long panel with fewer firm specific variables. Including fewer observable risk factors thereby creates a potentially larger role for the unobserved risk factor. This paper use a somewhat shorter but broader/deeper panel, which is likely to give a conservative estimate of the importance of correlations for credit risk.

Another distinction can be found in the different interpretation of default clustering, namely the explanation of correlation by unobserved factors or industry shocks..

uncertainty. If $\sigma_{\varepsilon_k} > \widehat{\sigma}_{\varepsilon_k}$, then VaR would be greater than reported.

Figure 5: Percentiles of the loss rate distribution implied by three models.

The boxes display the mean and three percentiles from the credit loss distributions that have been generated using three different models of default risk. In Box 1, the loss distribution has been generated with a conventional duration model. A cluster specific error structure has been added to this model in Box 2. In Box 3 all explanatory variables have been deleted from the cluster error model.




The similarity lies

To be able to single out the contribution of the cluster error structure to the shape of the loss and default rate distributions, it is useful to consider how other factors will affect these. Firstly, there's the size of the portfolio: larger portfolios, with more borrowers, will produce loss distributions that tend to converge to the mean and have smaller tail losses. Secondly, the *skewness* of the distribution of loan sizes will influence the location of the tails: portfolios in which the biggest loans are exceptionally large in relation to the remainder, will tend to generate bigger tail losses. Finally, there's the cluster errors, that capture the interdependency between firms within industries. A closer look at the boxes in Figures 4 and 5 makes what their impact is and that they, in this particular case, are the single driving force behind the larger default rates in the tail percentiles of the cluster model (Box 2). To see this, consider that in Figure 4 the single factor causing differences between Boxes 1 and 2 is the addition of the cluster errors. Hence, it is the existence of an interdependence between companies within industries that causes the higher percentiles of the default rate distribution to exceed the mean default rate by up to one and a half percent points. When we instead look at the loss distributions, in Figure 5,

**Table 5: The effect of including industry specific errors on the estimation of Value-at-Risk.**
Table shows the mean and various percentiles of the loss distribution that is generated by (i) the model with industry specific common errors, and (ii) the model without common industry specific errors. Normally styled figures refer to the model with industry specific errors, italics to the model witout industry errors.

| | | Simulated portfolio loss rates at loss distribution percentiles | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Quarter | mean | *mean* | 95 | *95* | 97.5 | *97.5* | 99.9 | *99.9* |
| 1994Q2 | 0,59 | *0,54* | 0,90 | *0,60* | 1,00 | *0,62* | 1,51 | *0,67* |
| 1994Q3 | 0,30 | *0,27* | 0,47 | *0,32* | 0,52 | *0,32* | 0,73 | *0,36* |
| 1994Q4 | 0,34 | *0,30* | 0,55 | *0,35* | 0,62 | *0,36* | 1,04 | *0,40* |
| 1995Q1 | 0,39 | *0,35* | 0,56 | *0,38* | 0,60 | *0,39* | 0,85 | *0,43* |
| 1995Q2 | 0,33 | *0,29* | 0,54 | *0,33* | 0,61 | *0,33* | 1,05 | *0,36* |
| 1995Q3 | 0,38 | *0,33* | 0,61 | *0,37* | 0,70 | *0,38* | 1,18 | *0,41* |
| 1995Q4 | 1,66 | *1,44* | 2,88 | *1,65* | 3,32 | *1,70* | 5,91 | *1,93* |
| 1996Q1 | 0,36 | *0,31* | 0,57 | *0,34* | 0,65 | *0,35* | 1,11 | *0,38* |
| 1996Q2 | 0,37 | *0,32* | 0,56 | *0,37* | 0,62 | *0,39* | 0,87 | *0,44* |
| 1996Q3 | 0,36 | *0,31* | 0,57 | *0,34* | 0,65 | *0,35* | 1,16 | *0,38* |
| 1996Q4 | 0,17 | *0,15* | 0,27 | *0,16* | 0,31 | *0,17* | 0,57 | *0,18* |
| 1997Q1 | 0,25 | *0,21* | 0,42 | *0,23* | 0,48 | *0,24* | 0,81 | *0,25* |
| 1997Q2 | 0,35 | *0,30* | 0,55 | *0,32* | 0,63 | *0,33* | 1,11 | *0,35* |
| 1997Q3 | 0,68 | *0,59* | 1,10 | *0,64* | 1,24 | *0,66* | 2,00 | *0,71* |
| 1997Q4 | 0,28 | *0,24* | 0,43 | *0,26* | 0,48 | *0,27* | 0,73 | *0,29* |
| 1998Q1 | 0,25 | *0,22* | 0,40 | *0,24* | 0,45 | *0,24* | 0,79 | *0,26* |
| 1998Q2 | 0,26 | *0,22* | 0,44 | *0,24* | 0,50 | *0,25* | 0,78 | *0,27* |
| 1998Q3 | 0,16 | *0,14* | 0,26 | *0,15* | 0,30 | *0,16* | 0,47 | *0,17* |
| 1998Q4 | 0,19 | *0,16* | 0,30 | *0,18* | 0,33 | *0,18* | 0,52 | *0,20* |
| 1999Q1 | 0,21 | *0,18* | 0,37 | *0,19* | 0,42 | *0,20* | 0,77 | *0,21* |
| 1999Q2 | 0,22 | *0,19* | 0,32 | *0,20* | 0,36 | *0,20* | 0,57 | *0,22* |
| 1999Q3 | 0,12 | *0,10* | 0,17 | *0,11* | 0,19 | *0,11* | 0,30 | *0,12* |
| 1999Q4 | 0,10 | *0,09* | 0,16 | *0,09* | 0,18 | *0,09* | 0,31 | *0,10* |
| 2000Q1 | 0,17 | *0,14* | 0,25 | *0,16* | 0,28 | *0,16* | 0,47 | *0,18* |
| 2000Q2 | 0,39 | *0,34* | 0,58 | *0,39* | 0,65 | *0,40* | 0,97 | *0,45* |

taking account of the cluster error structure appears to lead to increases in the VaR percentiles during most quarters that are comparable to those in the default rate distribution in Figure 4 and Table 4. Depending on the percentile and the quarter one considers, the model without cluster errors underestimates tail losses by up to four percentage points. Clearly, the exceptional peak in the loss distribution in the beginning of 1996, that exceeds the jump in the default rate distribution, has to be attributed to fact that, coincidentally, large exposures were being lost at a time when the average rate of default was also high. Table 5 provides numerical details for the boxes in Figure 5.

When we compare Box 2 in Figures 4 and 5 to the model with cluster errors but without explanatory variables in Box 3, one observes clearly that, through time, the difference between expected risk and tail risk in the outer tails increases when the mean default rate rises. Aggregate or "macro" risk, beyond affecting expected default rates, thus appears to be important for portfolio credit risk in the sense that periods with high rates of default among businesses are associated with more than proportional increases in unexpected defaults. The third model is, however, in both its mean and the lower percentiles, much less successful at capturing the trend in the average default rate. Only the 99.9th and, to a lesser extent, the 99th percentile, manage to produce a peak at the start and a slight downward movement over the remainder of the sample period. But the peak at the start is disproportionately high, while other increases in the default rate in the middle of the sample period, especially the one in 1997 are more or less fail to be captured. So although a model without a systemic part can generate peaks and troughs in the relevant range of the VaR percentiles, it is quite uninformative about when a bank should hold a large economic capital stock. In other words: a model with systematic factors but without a cluster error structure is well able to fit the broad trends in mean portfolio default rates and credit losses, but fails to appropriately capture the credit losses in bad times. Typically these are the times when banks are hit by large "unexpected" credit losses.[22] This model thus performs best when it is least needed. A model that accounts for the dependencies within each industry but lacks a systematic part produces quite some variation in the tails, but the shifts in the estimated loss distribution do not match those in the bank's actual loss experience. It would thus force the bank to hold an unnecessarily large economic capital stock for most of the time. Only the model that includes both a systematic part and cluster specific variances manages to follow both the trend in credit losses and produce industry driven fluctuation in losses around that trend. Consequently, the economic capital requirements that are derived from it are larger for periods when times of large "aggregate" disturbances occur and smaller when the economy is doing well. To what extent actual capital buffers will follow this economic requirement will depend on individual banks' "buffer smoothing" policy. Most likely, many banks will avoid too large fluctuations in their capital stock over time.

## 5 Discussion

Currently available portfolio credit risk models fail to allow for default risk dependency across loans other than through common risk factors. As a consequence they ignore the close ties that usually exist between companies, due to legal, financial and business relations, and tend to lead to clustering of bankruptcies. As a result, many models commonly used by regulators and financial institutions are likely to underestimate credit risk peaks in periods when industry specific shocks, as opposed to firm specific (idiosyncratic) and country wide factors, manifest themselves.

To address this shortcoming, we have attempted to integrate the insights from theoretical models of default correlation, like Giesecke [24] and Zhou [57], into a conventional duration model of default risk and the implied portfolio credit risk estimates by explicitly introducing

[22] When studying credit losses, authors generally refers to the difference between the credit loss at a specific percentile of the loss distribution minus the mean losses as "unexpected" credit losses.

a relation between firms' default risk through industry specific errors. In this paper, we also illustrate how this model of default risk can be operationalized by practitioners by applying it to a pooled data set from two Swedish banks' business loan portfolios over the period 1994-2000.

The results from the application are encouraging. Our estimates reveal substantial differences in the intra-industry inter-firm correlation of defaults. The retail industry and the "public and subsidized" sector display substantially higher cluster variation than most other industries, while the "real estate and finance" business demonstrates by far the largest correlation.While accounting for dependencies within industries does not affect estimates of individual default risk to any large extent, intra-industry dependencies are shown to be quantitatively important for credit risk. Estimates of VaR can increase by up to four percentage points, depending on the point in time and the aggregate economic conditions. Most commonly used credit risk models, that ignore these intra-industry effects, will thus substantially underestimate actual portfolio risk - by a degree that varies over the business cycle. "Conventional" models that include only systematic factors, although able to fit the broad trends in credit losses, fail to capture the fluctuations in unexpected credit losses, particularly in bad times, when banks typically are confronted with larger "unexpected" credit losses. The model we propose does manages to follow both the trend in credit losses and produces industry driven, time-varying fluctuations in losses around that trend. Our model will thus be a better aid for banks in determining the appropriate size of economic capital buffers.

What do these findings imply for financial institutions, financial supervisors and financial regulators? As we noted, the economic capital requirements that result from our model will be larger compared with the those implied by "conventional" models, especially in periods with large "aggregate" disturbances. All other equal, this should imply costlier lending for banks as they will be in need of larger capital buffers than one has thought until now. Another way to look at these results is that the trade-off banks face when determining how much to diversify is different. Manove, Padilla and Pagano ([40]) and Acharya, Hasan and Saunders ([3]) have, for example, discussed the costs that diversification bring about for banks and the trade-off they may face when balancing these against the gain from reducing tail credit losses. In this contaxt, our results seem to imply that the optimal degree of diversification is higher for banks than they inferred until now from less accurate models.

From a regulatory perspective, our findings will not have any immediate consequences, since regulatory capital requirements in the Basel II framework only depend on *average* probabilities of default. This paper is one in a larger literature that shows that the Basel risk weight mappings may not capture the actual risk properties of credit portfolios and that the success of implementing the Basel II framework will hinge on the specific characteristics (parameters) of banks' portfolios. For supervisors, it does mean, however, that the fragility of banks, and thus most likely the banking system, has been underestimated. In the long run, our findings ought to be a reason to make capital requirements contingent on portfolios' industrial composition, i.e. degree of diversifiaction.

As any model, ours assumes a specific structure. Obviously, there may be other ways to model intra-industry comovements between firms. We imposed that shocks to default risk are i.i.d. and firm default risk within an industry was positively correlated. One could think of some firms (within one or several industries) with a horizontal relation to display negative comovements, for example because a one firm's default creates new business opportunities for another. Another avenue may be to impose more structure on the cluster errors by using "input-output" data or other information documenting the intensity of contacts between firms and/or clusters. We believe that future research could explore these channels of inter-firm interaction.

## References


[1] Allen, Linda and Anthony Saunders, (2004), Incorporating systemic influences into risk measurements: A survey of the literature, *Journal of Financial Services Research.*26 (2), pp. 161-191.

[2] Acharya, Viral, Sreedhar Bharat and Anand Srinivasan, (2004), Understanding the recovery rates on defaulted securities, Forthcoming in :*Journal of Finance.*

[3] Acharya, Viral, Iftekhar Hasan and Anthony Saunders, (2006), Should Banks Be Diversified? Evidence from Individual Bank Loan Portfolios, *Journal of Business.*79 (3), pp. 1355-1412

[4] Altman, Edward I., (1968), Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* 23 (4), pp. 589-611.

[5] Altman, Edward I., (1971), Railroad bankruptcy propensity, *Journal of Finance* 26 (2), pp. 333-345.

[6] Altman, Edward I., (1973), Predicting railroad bankruptcies in America, *Bell Journal of Economics* 4 (1), pp. 184-211.

[7] Altman, Edward I., (1984), The success of business failure prediction models, *Journal of Banking and Finance* 4 (X), pp. 171-198.

[8] Arminger, Gerhard, Daniel Enache and Thorsten Bonne, (1997), Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis and feedforward networks, *Computational Statistics* 12, 293-310.

[9] Ashby, M., Neuhaus, J. M., Hauck, W. W., Bacchetti, P., Heilbron, D. C., Jewell, N. P., Segall, M. R. and R. E. Fusaro, (1992), An annotated bibliography of methods for analysing correlated categorical data, *Statistics in Medicine* 11:67-99.

[10] Bangia, Anil, Francis X. Diebold, André Kronimus, Christian Schagen and Til Schuermann, (2002), Ratings migration and the business cycle, with application to credit portfolio stress testing, *Journal of Banking and Finance* 26, pp. 445–474.

[11] Basel Committee, (2003), *The New Basel Capital Accord, Third Consultative paper*, www.bis.org/publ/bcbsca.htm.

[12] Breslow, N. E. and D. G. Clayton (1993), Approximate inference in generalized linear mixed models, *Journal of American Statistical Association*, 88:421, pp. 9-25.

[13] Butler, J. S. and Moffitt, R., (1982), A computational efficient quadrature procedure for the one-factor multinomial probit model, *Econometrica* 50(3):761-764.

[14] Calem, Paul and Michael LaCour-Little, (2004), Risk-based capital requirements for mortgage loans, *Journal of Banking and Finance* 28 (3), pp. 647-672

[15] Campbell, John Y., Jens Hilscher, and Jan Szilagyi, (2006), *In search of distress risk*, NBER Working Paper 12362

[16] Carey, Mark, (1998), Credit risk in private debt portfolios, *Journal of Finance*, 52 (4), pp. 1363-1387.

[17] Carling, Kenneth, Tor Jacobson, Jesper Lindé and Kasper Roszbach (2007), Corporate credit risk modeling and the macroeconomy, *Journal of Banking and Finance.*

[18] Carey, Mark and Mark Hrycay, (2001), Parameterizing credit risk models with rating data, *Journal of Banking and Finance* 25, pp. 197-270.

[19] Das, Sanjiv R., Darrell Duffie, Nikunj Kapadia, and Leandro Saita, (2007), Common Failings: How Corporate Defaults are Correlated, Forthcoming in: *Journal of Finance.*

[20] DeServigny, A. and O. Renault, (2002), *Default correlation: empirical evidence*, Working Paper, Standard and Poor's.

[21] Duffie, Darrell, Andreas Eckner, Guillaume Horel and Leandro Saita, (2006), Frailty Correlated Default, manuscript Stanford University.

[22] Frydman, Halina, Edward I. Altman, and Duen-Li Kao, (1985), Introducing recursive partitioning for financial classification: the case of financial distress, *Journal of Finance*, Vol. XL No. 1, pp. 269-291.

[23] Egloff, Daniel, Markus Leippold, and Paolo Vanini, (2004), *A Simple Model of Credit Contagion*, mimeo University of Zurich.

[24] Giesecke, Kay, (2004), Correlated default with incomplete information, Forthcoming in: *Journal of Banking and Finance.*

[25] Giesecke, Kay, and Stefan Weber, (2004), Cyclical correlations, credit contagion, and portfolio losses, Forthcoming in: *Journal of Banking and Finance.*

[26] Gordy Michael B., (2000), A comparative anatomy of credit risk models, *Journal of Banking and Finance*, 24 (1-2) , pp. 119-149.

[27] Gordy, Michael B., (2003), A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation* 12, pp. 199-232.

[28] Greene, W. H., (1991), Limdep version 6, Econometric Software, New York.

[29] Guilkey, D. K. and Murphy, J. L., (1993), Estimation and testing in the random effects probit model, *Journal of Econometrics* 59, pp.301-317.

[30] Hackbarth, D., J. Miao, and E. Morellec, (2006), Capital structure credit risk and macroeconomic conditions, Forthcoming in: *Journal of Financial Economics.*

[31] Heckman, J. J. and Willis, R., (1976), Estimation of a stochastic model of reproduction: An econometric approach, in: N. Terleckyj, ed. Houshold production and consumption, National Bureau of Economic Research, New York.

[32] Honjo, Yuji, (2000), Business failure of new firms: an empirical analysis using a multiplicative hazards model, *International Journal of Industrial Organization 18*, pp. 557-574.

[33] Jacobson, Tor, Jesper Lindé and Kasper Roszbach, (2006), Internal ratings systems, implied credit risk and the consistency of banks' risk classification policies, *Journal of Banking and Finance* 30 (7), pp.1899-1926

[34] Jacobson, Tor, Jesper Lindé and Kasper Roszbach (2005), Exploring interactions between real activity and the financial stance, *Journal of Financial Stability* 1 (3), pp.308-342.

[35] Jorion, Philippe, and Gaiyan Zhang, (2007), *Intra-industry Credit Contagion: Evidence from the Credit Default Swap and Stock Markets*, Forthcoming in: *Journal of Financial Economics.*

[36] Koopman, Siem Jan, Roman Kräussl, André Lucas, and André Monteiro, (2005), *Credit Cycles and Macro Fundamentals*, working paper Tinbergen Institute.

[37] Koyluoglu, H. Ugur, and Andrew Hickman, (1998), *A generalized framework for credit risk portfolio models*, Working paper, Oliver, Wyman & Company and Credit Suisse Financial Products. Published in abridged version as: "Reconcilable Differences", *Risk* (October 1998), pp. 56-62.

[38] Li, Kai, (1999), Bayesian analysis of duration models: an application to Chapter 11 bankruptcy, *Economics Letters* 63, pp. 305-312.

[39] Littell, R. C., Milliken, G. A., Stroup, W. W. and R. D. Wolfinger, (1996), *The SAS system for mixed models*, SAS Inst. Inc., Cary, North Carolina.

[40] Manove, Michael, Jorge Padilla and Marco Pagano, (2001), Collateral vs. Project Screening: A Model of Lazy Banks, *RAND Journal of Economics*, 32 (4).

[41] McCullagh, P. and Nelder, J. A., (1989), *Generalized linear models*, London: Chapman and Hall.

[42] Merton, Robert C., (1974), On the Pricing of Corporate Debt: The Risk Structure of Interest Rates", *Journal of Finance* 29, pp. 449-470

[43] Mingo, John J., (2000), Policy implications of the Federal Reserve study of credit risk models at major U.S. banking institutions, *Journal of Banking and Finance* 24 (1-2) , pp. 15-33.

[44] Narendranathan, W. and Stewart, M., (1993), Modelling the probability of leaving unemployment: Competing risks models with flexible baseline hazard, *Applied Statistics* 42, pp. 63-83.

[45] Olsson, U, (2002), "Generalized Linear Models: an applied approach", Studentlitteratur, Lund.

[46] Nickell, Pamela, William Perraudin and Simone Varotto, (2000), Stability of rating transitions, *Journal of Banking & Finance* 24, pp.203-227.

[47] Pesaran, Hashem, Til Schuermann, Björn-Jakob Treutler, and Scott Weiner, (2006), Macroeconomic Dynamics and Credit Risk: A Global Perspective, *Journal of Money Credit and Banking* 38 (5), pp. 1211-1262.

[48] Platt, Harlan D., and Marjorie B. Platt, (1991), A note on the use of industry-relative ratios in bankruptcy prediction , *Journal of Banking and Finance* 15, pp. 1183-1194.

[49] Statistics Sweden, (2002), Industry code - SNI2002, Stockholm, Sweden. www.scb.se/templates/Listning2____35025.asp

[50] Shumway, Tyler, (2001), Forecasting bankruptcy more accurately: a simple hazard model, *Journal of Business* 74 (1), pp. 101-124.

[51] Song, J. X. and C. W. Ahn, (2003), An evaluation of methods for the stratified analysis of clustered binary data in community intervention trials, *Statistics in Medicine* 22:2205-226.

[52] Szegö, Giorgio, (2002), Measures of risk, *Journal of Banking and Finance* 26, pp. 1253-1272.

[53] Treacy, William and Mark Carey, (2000), Credit risk rating systems at large US banks, *Journal of Banking and Finance* 24 (1-2), pp. 167-201.

[54] Wilson, Thomas, (1997), Portfolio credit risk (I), *Risk* 10 (9), pp. 111-116

[55] Wolfinger, R. and M. O'Connell, (1993), Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation Simulation* 48, pp. 233-243.

[56] Zhou, X.-H., Perkins, A. J. and Hui, S. L., (1999), Comparisons of software packages for generalized linear multilevel models, *The American Statistician* 53 (3), pp. 282-290.

[57] Zhou, Chunsheng, (2001), An analysis of default correlations and multiple defaults, *Review of Financial Studies* 14 (2), pp. 555-576.

# A Appendix 1

**Table 6. Definition of clusters and their distribution in the data.**

| Variable | Included SNI-codes | | | | | % share |
|---|---|---|---|---|---|---|
| Public or subsidized sector | 0-9999 | 22100-22330 | 40000-41000 | 75000-91330 | 99000-99999 | 13.3 |
| Wholesale | 51000-51700 | | | | | 12.7 |
| Retail | 50000-50500 | 52000-55529 | 71000-71100 | 71400-71402 | 92000-98999 | 19.4 |
| Transport and communication | 60000-64203 | 71200-71340 | 72000-74849 | | | 23.5 |
| Manufacturing | 27000-35999 | | | | | 7.8 |
| Construction, forest and other industries | 10000-22000 | 23000-26829 | 36000-39999 | 45000-45500 | | 14.9 |
| Estate and finance | 65000-70329 | | | | | 8.4 |

# B Appendix 2

Table A2 shows the frequency by which companies occur in broader sectors of the input-output matrix in our databasedatabase.

The first column (clust) shows the 57 input-output clusters which are the (first) two-digits-level aggregation on SNI codes of the companies. The second column (frequency) shows the

**Table A1. Estimates of the systematic part of default risk duration model with industry interaction terms.**

The table replicates results from Table 2 for a model with macro and financial ratio factors plus additional interaction terms between industry dummies and financial ratios. The model is estimated with cluster specific errors, where correlations are allowed to differ among industries. No-autocorrelation is assumed. The link function is the complementary log-log. To facilitate the reading of the table, the interaction terms have been placed in a separate column for each industry. The "intercept" for the financial ratios should be interpreted as the effect for industry number 7. Industry definitions are given in Table 3 and Table 6a. Significance levels of parameters are indicated as: *** = 1% level, **=5% level, *=10% level.

| | | | Industry 1 | | Industry 2 | | Industry 3 | | Industry 4 | | Industry 5 | | Industry 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Estimate | StdErr | Estimate | StdErr | Estimate | StdErr | Estimate | StdErr | Estimate | StdErr | Estimate | StdErr | Estimate | StdErr |
| **Bank data** | | | | | | | | | | | | | | |
| Short-term credit | 0.516 *** | 0.037 | | | | | | | | | | | | |
| Mix of credits | -0.298 *** | 0.049 | | | | | | | | | | | | |
| Long-term credit | 0.000 | - | | | | | | | | | | | | |
| Bank A | -0.121 *** | 0.038 | | | | | | | | | | | | |
| **Credit bureau data** | | | | | | | | | | | | | | |
| Complete | -2.625 *** | 0.084 | | | | | | | | | | | | |
| Reported previously | 0.583 *** | 0.080 | | | | | | | | | | | | |
| Not reported | -3.561 *** | 0.235 | | | | | | | | | | | | |
| Partially missing | 0.000 | - | | | | | | | | | | | | |
| Payment remark | 0.958 *** | 0.050 | | | | | | | | | | | | |
| Mismanaged loan | 0.832 *** | 0.064 | | | | | | | | | | | | |
| **Accounting data** | | | | | | | | | | | | | | |
| Industry dummy | | | -0.612 * | 0.332 | -0.797 ** | 0.399 | -0.954 ** | 0.384 | -0.919 *** | 0.346 | -0.780 | 0.483 | -1.180 *** | 0.431 |
| Log(TS) | 0.868 *** | 0.152 | | | | | | | | | | | | |
| TS data missing | -0.069 *** | 0.014 | 0.038 ** | 0.016 | 0.025 | 0.020 | 0.046 ** | 0.020 | 0.047 ** | 0.018 | 0.037 | 0.026 | 0.057 ** | 0.023 |
| EBITA/TA | -0.266 ** | 0.131 | 0.221 | 0.142 | -0.077 | 0.160 | 0.078 | 0.152 | 0.096 | 0.146 | -0.140 | 0.212 | -0.072 | 0.164 |
| Inventories/TS | -0.960 * | 0.540 | 0.886 | 0.609 | 2.142 *** | 0.603 | 1.486 *** | 0.573 | 1.837 *** | 0.576 | 1.767 *** | 0.671 | 1.426 | 0.606 |
| TL/TA | 0.804 *** | 0.147 | -0.138 | 0.165 | 0.263 | 0.182 | 0.166 | 0.168 | 0.086 | 0.165 | 0.158 | 0.223 | 0.249 | 0.182 |
| **Macro data** | | | | | | | | | | | | | | |
| Output gap | -0.168 *** | 0.022 | | | | | | | | | | | | |
| Yield curve | -0.217 *** | 0.055 | | | | | | | | | | | | |
| **Duration** | | | | | | | | | | | | | | |
| 1:st year | -0.170 | 0.129 | | - | | | | | | | | | | |
| 2:nd year | 0.062 | 0.129 | | | | | | | | | | | | |
| 3:rd year | 0.151 | 0.130 | | | | | | | | | | | | |
| 4:th year | 0.280 *** | 0.128 | | | | | | | | | | | | |
| 5:th year | 0.294 *** | 0.139 | | | | | | | | | | | | |
| > 5 years | 0.000 | | | | | | | | | | | | | |

Table A1 continued. Estimates of the systematic part of default risk duration model with industry interaction terms.

| | | |
|---|---|---|
| Intercept | -3.544 *** | 0.334 |
| | | |
| Nobs | | 950,693 |
| Loglikelihood | | -4.235E+06 |
| | | |
| Gamma | | 0.990 |
| Kendall's Tau-b | | 0.174 |
| Somers' D def\|P(def) | | 0.506 |
| Pearson correlation | | 0.174 |

**Table A2: Statistics Sweden industry classification**

The table shows activities in industries at 2-digit level. Missing digits reflect unused industries cod

| SNI code | Nobs | Share | Activities |
|---|---|---|---|
| 01 | 785 | 0.88 | Agriculture, hunting and related service activities |
| 02 | 380 | 0.43 | Forestry, logging and related service activities |
| 05 | 87 | 0.10 | Fishing, operation of fish hatcheries and fish farms; service activities incidental to fishing |
| 10 | 25 | 0.03 | Mining of coal and lignite; extraction of peat |
| 11 | 41 | 0.05 | Extraction of crude petroleum and natural gas; service activities incidental to oil and gas extraction excluding surveying |
| 12 | 10 | 0.01 | Mining of uranium and thorium ores |
| 13 | 21 | 0.02 | Mining of metal ores |
| 14 | 115 | 0.13 | Other mining and quarrying |
| 15 | 784 | 0.88 | Manufacture of food products and beverages |
| 16 | 2 | 0.00 | Manufacture of tobacco products |
| 17 | 293 | 0.33 | Manufacture of textiles |
| 18 | 137 | 0.15 | Manufacture of wearing apparel; dressing and dyeing of fur |
| 19 | 56 | 0.06 | Tanning and dressing of leather; manufacture of luggage, handbags, saddlery, harness and footwear |
| 20 | 964 | 1.09 | Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials |
| 21 | 211 | 0.24 | Manufacture of pulp, paper and paper products |
| 22 | 2016 | 2.27 | Publishing, printing and reproduction of recorded media |
| 23 | 15 | 0.02 | Manufacture of coke, refined petroleum products and nuclear fuel |
| 24 | 373 | 0.42 | Manufacture of chemicals and chemical products |
| 25 | 659 | 0.74 | Manufacture of rubber and plastic products |
| 26 | 294 | 0.33 | Manufacture of other non-metallic mineral products |
| 27 | 173 | 0.19 | Manufacture of basic metals |
| 28 | 2643 | 2.98 | Manufacture of fabricated metal products, except machinery and equipment |
| 29 | 1678 | 1.89 | Manufacture of machinery and equipment n.e.c |
| 30 | 122 | 0.14 | Manufacture of office machinery and computers |
| 31 | 522 | 0.59 | Manufacture of electrical machinery and apparatus n.e.c. |
| 32 | 266 | 0.30 | Manufacture of radio, television and communication equipment and apparatus |
| 33 | 724 | 0.82 | Manufacture of medical, precision and optical instruments, watches and clocks |
| 34 | 407 | 0.46 | Manufacture of motor vehicles, trailers and semi-trailers |
| 35 | 306 | 0.34 | Manufacture of other transport equipment |
| 36 | 735 | 0.83 | Manufacture of furniture; manufacturing n.e.c. |
| 37 | 46 | 0.05 | Recycling |
| 40 | 325 | 0.37 | Electricity, gas, steam and hot water supply |
| 41 | 13 | 0.01 | Collection, purification and distribution of water |
| 45 | 8384 | 9.44 | Construction |
| 50 | 24869 | 28.00 | Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel |
| 51* | | | Wholesale trade and commission trade, except of motor vehicles and motorcycles |
| 52* | | | Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods |
| 55 | 2723 | 3.07 | Hotels and restaurants |
| 60 | 4545 | 5.12 | Land transport; transport via pipelines |
| 61 | 941 | 1.06 | Water transport |
| 62 | 609 | 0.69 | Air transport |
| 63 | 1252 | 1.41 | Supporting and auxiliary transport activities; activities of travel agencies |
| 64 | 111 | 0.12 | Post and telecommunications |
| 65 | 478 | 0.54 | Financial intermediation, except insurance and pension funding |
| 66 | 4 | 0.00 | Insurance and pension funding, except compulsory social security |
| 67 | 815 | 0.92 | Activities auxiliary to financial intermediation |
| 70 | 6859 | 7.72 | Real estate activities |
| 71 | 1317 | 1.48 | Renting of machinery and equipment without operator and of personal and household goods |
| 72 | 2251 | 2.53 | Computer and related activities |
| 73 | 272 | 0.31 | Research and development |
| 74 | 13490 | 15.19 | Other business activities (mainly highly educated services) |
| 75 | 15 | 0.02 | Public administration and defence; compulsory social security |

| | | | |
|---|---|---|---|
| 80 | 607 | 0.68 | Education |
| 85 | 1671 | 1.88 | Health and social work |
| 90 | 214 | 0.24 | Sewage and refuse disposal, sanitation and similar activities |
| 91 | 54 | 0.06 | Activities of membership organizations n.e.c. |
| 92 | 1363 | 1.53 | Recreational, cultural and sporting activities |
| 93 | 668 | 0.75 | Other service activities |
| 95 | 79 | 0.09 | Activities of households as employers of domestic staff |
| 99 | | | Extra-territorial organizations and bodies |
| All | 88819 | 100.00 | |

* In the optimal cluster generation, these industries were recoded as 50 to create compatibility with the input-output classification.

number of companies present in a particular input-output cluster in our database. To make the database workable, we have applied the following modifications of the data:

1. The companies appearing with different SNI codes in different time periods have been considered as though they appeared with the latest valid SNI code over the whole period.

2. Entries with a missing SNI code, a zero value as SNI code, a SNI code with fewer than five digits and other (infrequently occurring) forms of invalid SNI codes have been discarded.

3. To be consistent with the coding of the input-output matrix, companies having the first two digits of the SNI code equal to 50, 51 or 52 have been recoded as 50.

**Table 6b. Mapping from SNI industry codes to optimal clusters**

The table shows how all industries, according to the Statistics Sweden industry classification system, map into the seven optimal clusters. All clusters were optimized by maximizing the intra-cluster purchases at the two digit industry code level. Source of industry classification: http://www.scb.se

| Cluster | Included SNI-codes | | | | | | | | | | | | | | | | Obs | % share |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 13 | 50 | 80 | 85 | all missing SNI codes | | | | | | | | | | | | 345458 | 36.3 |
| 2 | 10 | 20 | 22 | 30 | 37 | 67 | 72 | 74 | 93 | | | | | | | | 184491 | 19.4 |
| 3 | 02 | 24 | 25 | 45 | 91 | | | | | | | | | | | | 101069 | 10.6 |
| 4 | 05 | 11 | 15 | 18 | 23 | 26 | 31 | 33-34 | 36 | 41 | 63-65 | 70 | 75 | 90 | 92 | 95 | 143564 | 15.1 |
| 5 | 12 | 19 | 60 | 66 | | | | | | | | | | | | | 47026 | 5.0 |
| 6 | 14 | 16-17 | 21 | 55 | 61 | 71 | 73 | | | | | | | | | | 56621 | 6.0 |
| 7 | 01 | 27-29 | 32 | 35 | 40 | 62 | | | | | | | | | | | 72464 | 7.6 |
| Total | | | | | | | | | | | | | | | | | 950693 | 100.0 |

**Table 6c: Main activities in optimal clusters**

Columns two and three determine the main activities in an optimal cluster by the number of companies in our sample that belong to this cluster. This does not take into account that some industries have few but very large players. The purpose of identifying the main activities is to facilitate a comparison with the exogenous ("old") clusters in Table 6a. Column three identifies which old cluster the optimal cluster resembles most.
Source of industry classification: http://www.scb.se

| Cluster | Major activities | Minor activities | Has much overlap with following "old" clusters |
|---|---|---|---|
| 1 | Retail and wholesale trade, incl automotive services | Education and health & social services | Retail, Wholesale |
| 2 | Other business activities (mainly professional services) | Manufacturing of wood, publishing, computer activities | Transport, communication and professional services |
| 3 | Construction | Forestry, chemical industry | Construction, forest and other industries |
| 4 | Real estate activities, financial intermediation | Food industry, telecom, travel agencies, some manufacturing | Real estate and finance |
| 5 | Land transport | Insurance and pension funding excl. social security | Transport, communication and professional services<br>Real estate and finance |
| 6 | Hotels and restaurants | Some manufacturing, R&D, renting services | Retail<br>Transport, communication and professional services |
| 7 | Manufacturing (metal products, other) | Agriculture, electricity, air transport | Manufacturing |