THE RIMINI CENTRE FOR ECONOMIC ANALYSIS

WP 13_11

**Dimitris Korobilis**
Université Catholique de Louvain, Belgium
The Rimini Centre for Economic Analysis (RCEA), Italy

**Michelle Gilmartin**
University of Strathclyde, UK

# ON REGIONAL UNEMPLOYMENT: AN EMPIRICAL EXAMINATION OF THE DETERMINANTS OF GEOGRAPHICAL DIFFERENTIALS IN THE UK



The *Rimini Centre for Economic Analysis* (RCEA) was established in March 2007. RCEA is a private, nonprofit organization dedicated to independent research in Applied and Theoretical Economics and related fields. RCEA organizes seminars and workshops, sponsors a general interest journal *The Review of Economic Analysis*, and organizes a biennial conference: *The Rimini Conference in Economics and Finance* (RCEF) . The RCEA has a Canadian branch: *The Rimini Centre for Economic Analysis in Canada* (RCEA-Canada). Scientific work contributed by the RCEA Scholars is published in the RCEA Working Papers and Professional Report series.



The Rimini Centre for Economic Analysis
Legal address: Via Angherà, 22 – Head office: Via Patara, 3 - 47900 Rimini (RN) – Italy
www.rcfea.org - secretary@rcfea.org

# On regional unemployment: an empirical examination of the determinants of geographical differentials in the UK*

Dimitris Korobilis†
Université Catholique de Louvain

Michelle Gilmartin
University of Strathclyde

**Abstract**

This paper considers the determinants of regional disparities in unemployment rates for the UK regions at NUTS-II level. We use a mixture panel data model to describe unemployment differentials between heterogeneous groups of regions. The results indicate the existence of two clusters of regions in the UK economy, characterised by high and low unemployment rates respectively. A major source of heterogeneity seems to be caused by the varying effect (between the two clusters) of the share of employment in the services sector, and we trace its origin to the fact that the "high unemployment" cluster is characterised by a higher degree of urbanization.

*Keywords:* distribution dynamics, regional labour markets, unemployment differentials.

*JEL Classification:* C23, J08, J64, R12, R23.

*Michelle Gilmartin is a Fellow of the Fraser of Allander Institute. Dimitris Korobilis is a Fellow of the Rimini Center for Economic Analysis.

†Corresponding author. Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, 34 Voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium. Tel: +32 (0)10 474351, e-mail: dimitrios.korompilis@uclouvain.be

# 1 Introduction

Despite the UK economy having experienced a downward trend in the national unemployment rate since the mid 1980s, significant and sustained unemployment differentials have been observed amongst the regions throughout the same period. Unemployment differentials between regions are often greater than between countries (Taylor and Bradley, 1997), suggesting that regional unemployment data offer an additional source of information for investigating the causes of aggregate unemployment fluctuations. The conjecture that spatial disparities may give rise to inefficiencies implies that appropriately-designed interventionist policies could lead to significant improvements in national economic performance. Furthermore, eliminating such differentials could alleviate the adverse socioeconomic effects associated with spatial concentrations of high unemployment (Elhorst, 2003).

To better understand the dynamics of these differentials in the case of the UK, Figure 1 plots the unemployment rates for 5 NUTS-II level regions. The data are annual for the period 1999 to 2008. Although all five regions belong to the same NUTS-I level[1] region (Northwest England), there are some interesting patterns observed which provide an example of the influence of the geographical disparity of unemployment rates at the regional level. Merseyside, Greater Manchester and (to a lesser degree) Cheshire, are characterized by $U$-shaped curves. This shape is consistent with the behavior of national unemployment rates during the same period: a decline from 1999 to mid 2000's, and an increase in unemployment around one year prior, and during, the 2007-2008 global financial crisis. However the region of Cumbria has an obvious downward trend (even during the last years of the sample), while Lancashire's unemployment rate has stayed relatively constant until 2005 and then increased only in 2006 and 2007. Similarly, in terms of absolute unemployment rates, Merseyside stands out from all other four regions. Therefore, if we were to group regions into "clusters of similarity" in order to study them more effectively, it would be misleading to classify Lancashire with Merseyside just because they are spatially proximate. At the same time, in order to design more efficient regional economic policies, it is not only potential heterogeneity in the absolute level of unemployment which matters, but also potential heterogeneity in factors affecting unemployment.

Economic theory provides guidance only as to where we should seek the reasons of these regional disparities. However, knowing how regions are distributed into high or low unemployment groups/clusters is an empirical issue which allows us to study and summarize the origins of these developments. In a pioneering study, Overman and Puga (2002) cluster 150 NUTS-II European regions according to different similarity characteristics, and Lopez-Bazo, del Barro and Artis (2005) extent this analysis by additionally considering the effects of factors that affect unemployment rates (including equilibrium/disequilibrium variables, demographic characteristics etc.). Cracolici, Cuffaro and Nijkamp (2007) use a panel data model with spatial dependence effects in order to examine the geographical differentials in Italian regions.

In this paper we follow a similar idea using a novel econometric methodology, and apply it in the case of UK regions. We use a mixture panel data model with equilibrium and disequilibrium variables, in order to cluster regions according to similar (descriptive) characteristics. Our model extends a "traditional" panel data model in many ways. Coefficients are pooled across clusters, implying that regions belonging to the same cluster are similar. The clustering is done in a data-

[1] In total UK has 11 NUTS-I regions, and within them there are 37 NUTS-II regions. In this paper we focus on unemployment differentials at the NUTS-II level.

based fashion, rather than being defined by an arbitrary measure of proximity. Additionally, the model offers a good balance between bias and estimation error. By defining cluster specific coefficients, our model is not as restrictive as a regression with coefficients pooled across all regions, nor subject to the large estimation errors that can occur if we would allow a different regression coefficient for each individual region. The latent panel data approach we use is part of a series of recent developments in classification of regions or countries using mixtures (see for example Tsionas, 2000; Canova, 2004; Paap, Franses and vanDijk, 2005; Fruewirth-Scnatter and Kaufmann, 2008, to name but a few).

The remainder of the paper proceeds as follows: in Section 2 we describe our model and the econometric methodology. Section 3 details our empirical dataset and in Section 4 we discuss the empirical results of applying our novel methodology to the issue of regional differentials. Finally, in Section 6 we provide concluding comments and insights from a policy perspective.

# 2 An empirical model of regional unemployment differentials

## 2.1 Model formulation

Let $y_{it}$ be the annual unemployment rate in region $i$ at time $t$, for $i = 1, ..., N$ and $t = 1, ..., T$. Dynamic panel data analysis of $y_{it}$ involves autoregressive/distributed lag models of the form (Baltagi, 2005)

$$y_{it} = \widetilde{\varphi} y_{it-1} + \widetilde{\beta}_i x_{it} + \varepsilon_{it} \tag{1}$$

where $x_{it}$ may include strictly exogenous variables, and $y_{it-1}$ are the lagged values of $y_{it}$, and $\varepsilon_{it} \sim N\left(0, \sigma_i^2\right)$ is the error term. With regional variables, like annual unemployment rates, it is usually the case that the number of regions is large, while the time-series dimension is short. In that case, estimators of the regression coefficients $\widetilde{\beta}_i$ from the data $y_i$ will exhibit large estimation errors. In order to deal with this problem the researcher can pool across regions and instead estimate a joint parameter $\widetilde{\beta}_i \equiv \beta$. That way we end up studying only one "representative" region which is not very useful. Spillovers and interactions across regions are important and no region should be studied in isolation (Quah, 1996)

The regional empirical model we use is an extension of the panel data model used in Paap et al. (2005). Unemployment rates are modeled as a mixture of $C$ unobserved clusters, where coefficients are pooled only across regions having similar characteristics (i.e. belonging in the same cluster). Let $r_i = j$ *index that region* $i$ *belongs to cluster* $j$, for $j = 1, ..., C$ clusters in total, and with a respective probabilities $p_{i1} = P[r_i = 1]$, $p_{i2} = P[r_i = 2]$,..., $p_{iC} = P[r_i = C]$, where $0 \leq p_{i1}, p_{i2}, ..., p_{iC} \leq 1$ and $\sum_{j=1}^{C} p_{ij} = 1$. The model we use can be written as

$$\begin{aligned} y_{it} - \alpha_j - \beta_j x_{it} &= \varphi_j \left(y_{it-1} - \alpha_j - \beta_j x_{it}\right) + \eta_{it}, & (2) \\ \text{if } r_i &= j,\ j = 1, ..., C & (3) \end{aligned}$$

where $y_{it-1}$ are lagged values of unemployment in each region with autoregressive coefficients $\{\varphi_1, \varphi_2, ..., \varphi_C\}$. The model is written in steady-state form, $y - \mu_j = \varphi\left(y_{-1} - \mu_j\right) + error$, meaning that the quantity $\mu_j = \alpha_j + \beta_j x_{it}$ is the *unconditional* mean of unemployment. In order to characterize only $\alpha_j$ as the long-run average equilibrium level (steady-state) of unemployment in cluster $j$, we transform the regressors $x_{it}$ to have mean zero. The errors $\eta_{it}$ are

distributed as $\eta_{it} \sim N(0, \Omega)$, where $\Omega$ is a full $N \times N$ covariance matrix. To preserve parsimony in the covariance structure of the error vector $\eta_{it}$ we decompose it into a common component, $\delta_i f_t$, plus a Normal error term $\varepsilon_{it}$ which has a parsimonious diagonal covariance matrix, $\Sigma = diag\left\{\sigma_1^2, ..., \sigma_N^2\right\}$. We write this as

$$\eta_{it} = \delta_i f_t + \varepsilon_{it} \tag{4}$$

so that in the decomposition above we need to estimate the $N$ elements $\delta_i$ and the $N$ elements $\sigma_i^2$ , instead of the $N(N+1)/2$ unknown elements of a full covariance matrix $\Omega$. This decomposition is followed in practice because we have too few time series observations for each of the regions ($T \ll N$) and subsequently we want to save degrees of freedom; see also Paap et al. (2005) for a similar specification.

There are numerous statistical methods available for the empirical classification of spatial or panel data, while it is notable that most of those have been applied in the regional economic growth literature. For instance, the spatial regimes regressions of Fischer and Stirbock (2004) and Baumont, Ertur and Le Gallo (2003) have the additional (compared to our model) advantage of explicitly modelling the spatial correlations, while grouping similar regions at the same time. Nevertheless, as it is the case in Overman and Puga (2002), the grouping is done using geographical proximity as a measure of similarity for within-group members.

Similarly there are shortcoming of other approaches, like the multiple regime regression model of Durlauf and Johnson (1995), which is used to classify observations into clusters by using two specific control variables. All these models are subject to the sensitivity of choosing subjectively a control variable or measure of proximity. Crone (2005) uses an old and established data-based clustering technique called K-means clustering, combined with extracting leading indicators of regional variables based on a dynamic factor model. The K-means clustering technique, however, is only based on similarities of the measured variable $y_{it}$ (regional unemployment rates in our case), and cannot provide cluster specific effects $\beta_j$ of factors affecting unemployment ($x_{it}$ in our model). In contrast, our model gives cluster specific coefficients for each variable affecting unemployment, allowing for better understanding of the economic structure of each cluster of regions. The following subsection explains the mechanics of how this clustering happens in practice.

## 2.2 Estimation

The parameters of this model consist of $\theta_j = (\alpha_j, \beta_j, \varphi_j)$, $j = 1, ..., C$, which are pooled across clusters, and $\delta_i$ which are different across regions $i$. As explained in the Introduction, the cluster-specific coefficients imply that regions belonging to the same cluster are defined by common effects, while regions belonging in different clusters have structural differences in their unemployment determinants. A different interpretation of our model occurs if we solve for only the current level of unemployment in the left-hand side of equation (2) and replace $\eta_{i,t}$ as in (4). For instance, in the case of two clusters we have

$$y_{it} = \begin{cases} \widetilde{\alpha}_1 + \varphi_1 y_{it-1} + \widetilde{\beta}_1 x_{it} & \\ \widetilde{\alpha}_2 + \varphi_2 y_{it-1} + \widetilde{\beta}_2 x_{it} & \end{cases} + \delta_i f_t + \varepsilon_{it}, \begin{array}{l} if\ r_i = 1 \\ if\ r_i = 2 \end{array}. \tag{5}$$

where $\widetilde{\alpha}_j = \alpha_j (1 - \varphi_j)$ and $\widetilde{\beta}_j = \beta_j (1 - \varphi_j)$. The formulation above shows that the unemployment in each of the $N$ regions $i$ is specified as following the model with parameters

$\theta_1 = (\alpha_1, \beta_1, \varphi_1)$ with probability $p_{i1}$, and the model with parameters $\theta_2 = (\alpha_2, \beta_2, \varphi_2)$ with probability $p_{i2}$.

We briefly describe the estimation steps. More results can be found in Paap, Franses and vanDijk (2005) and Basturk, Paap and vanDijk (2008). The data likelihood function implied by the 2-cluster model in equation (5) is

$$L\left(y_t|, \theta_j, \delta_i\right) \propto \prod_{i=1}^{N} \left( p_1 \prod_{t=1}^{T} \frac{1}{\sigma_i} \exp\left(-\frac{\varepsilon_{it}}{\sigma_i}\right)^2 \right)^{\mathbf{1}(r_i=1)} \left( p_2 \prod_{t=1}^{T} \frac{1}{\sigma_i} \exp\left(-\frac{\varepsilon_{it}}{\sigma_i}\right)^2 \right)^{\mathbf{1}(r_i=2)}$$

where $\mathbf{1}\left(r_i = c\right)$ is an indicator variable which takes the value $1$ if the condition $r_i = c$ is satisfied, and is $0$ otherwise.

Since the variables that index which region belongs to each cluster, $r_i$, are unobserved (latent), maximization of the likelihood function is not straightforward. However it is convenient to use the expectation-maximization (EM) algorithm in order to find the maximum likelihood (ML) estimate; see McLachlan and Krishnan (1997). The main mechanism behind the EM algorithm is this:

1. Given starting (proposed) values for the parameters, we compute the probability that a region $i$ belongs to cluster $j$ (which we denoted $p_j$ above) as the ratio of the value of the log-likelihood for cluster $j$ to the sum of the log-likelihood value for all $C$ clusters (subsequently we estimate $C$ different probabilities for each of the $N$ different regions).

2. Now that $p_j$ is known, the log-likelihood can be evaluated conditional on the proposed parameters (expectation or E-step of the algorithm).

3. Additionally, given $p_j$ we also find the values of the parameters which maximize the log-likelihood (maximization or M-step).

The algorithm will cycle through steps 2 and 3, where the parameter values obtained from the M-step are used as proposals in the E-step above. Converge is achieved when the change in the likelihood from one cycle to the next is minimal (less than $1 \times 10^{-9}$).

# 3 Regression results

In this section we apply the model in equations (2) - (4) to study convergence and determinants of regional unemployment rates in the UK. We first describe the full dataset and choice of variables. Then we provide model estimates for the full sample of our data, but also for subsamples.

We use annual measurements on several regional variables for the UK for the period 1999 - 2008. All data are from Eurostat (http://ec.europa.eu/eurostat) and are at the NUTS-II level. North Eastern Scotland, and Highlands & Islands are excluded from the analysis due to missing observations. The empirical results are based on the remaining 35 NUTS-II UK regions, a list of which is provided in the first column of Table 3.

The dependent variable in our study, $y_{i,t}$, is the log of regional unemployment rate for people 15 years and over. For the factors in $x_{i,t}$, i.e. the R.H.S variables in our regression, we consider several variables that potentially affect regional unemployment rates. We consider employment growth (EMPLG) as a disequilibrium factor. However in light of potential endogeneity we use the first lag of this variable (see Lopez-Bazo, del Barro and Artis, 2002). We also include the shares

of employment in agriculture (EAGR), manufacturing (EMAN), and services (ESERV) as market equilibrium variables (see Taylor and Bradley, 1997; Bean, 1994). Age effects are captured using the ratio of students that have at least started high school over working age population (HCAP), as well as the shares of people 15-29 years old (YOUNG) and people 65+ old (OLD) over total population (see for example Cracolici, Cuffaro and Nijkamp, 2007). As additional demographic effects, we use the participation rate of women (FEMP) and men (MALP), which are defined as the ratio of female (male) labour force over total females (males) at working age. Demographic equilibrium effects are captured through measurements on net migration balance (MIG) and the population density (DENS) of each region. Data on wages or real labor costs were not available at the NUTS-II (not even at the NUTS-I) level, and subsequently are absent from the empirical model.

Possible remaining endogeneity can only be assumed but not tested, due to the lack of NUTS-II level variables that could be used as instruments. This is true for other previous studies which have used similar a dataset combined with panel estimation; see Lopez-Bazo, del Barro and Artis (2005) and Cracolici, Cuffaro and Nijkamp (2007) for a discussion of these issues. The results were quite robust to specifications in which variables with non-significant coefficients were removed from the list of regressors, so in the next we present empirical results using the full set of variables described above.

Finally we need to decide on the measure of common determinant $f_t$ of regional unemployment rates. Our choice is to use the national unemployment rate, which is a reasonable choice following theory and previous empirical evidence (Elhorst, 2003). Alternatively, the common component could be extracted using a latent factor model (principal component) on the regional unemployment rates.

The next step in our analysis is to determine the number of clusters for the UK regions. In order to achieve that we succesively run models with number of clusters $C$ equal to 1 (i.e. the fixed effects model), 2, 3, 4, 5, and 35 (i.e. the random effects model). Then we decide upon the number of clusters using information criteria, which are summarized in Table 1. Along with traditional criteria for model selection, like the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), we provide values of the AIC3 and the AIC3 adjusted. The latter two measures, which replace the factor 2 in the usual AIC by a factor 3, are justified from the simulations of Bozdogan (1994) who shows their superior performance in selecting the correct number of clusters in mixture models. Finally, we present values of all criteria both based on the number of regions only ($N$), and based on the total number of observations ($N \times T$).

The criteria values in Table 1 show that the model with 2 clusters gets the most support from the AIC and AIC-3 criteria. The BIC criterion suggests the model with one cluster, although its value is not very different for the models with 2 and 3 clusters. It is well known that the BIC criterion - as an approximation to the Bayesian marginal likelihood - will always select the more parsimonious among two " equally good" models. Here note that conditional on the number of clusters, we estimate the model with all the variables, and then we eliminate variables based on the t-statistics of their respective coefficients. Given the evidence in the information criteria, we will proceed our analysis comparing the models with one and two clusters.

Estimation results from these two models are presented in Table 2. The two-cluster model indicates a separation of the two clusters: 41% (or 14 regions) belong to a high unemployment group (cluster 1) and 59% (or 21 regions) belong to a low unemployment group (cluster 2). The separation into high/low unemployment groups is indicated from the fact that the intercept $\alpha_j$,

which is equal to unconditional mean $\mu$ of regional unemployment rates[2], is higher in the first cluster. Additionally, the coefficients on the exogenous variables $\beta_j$ on this same cluster are of equal magnitude or higher (in absolute values) compared to cluster 2. Of the 11 determinants of unemployment it is only YOUNG which seems to be insignificant in the one-cluster ($C = 1$) and two-cluster ($C = 2$) models. While MIG, ESERV and DENS are not significant at the 5% significance level in the model with $C = 1$, they all become siginificant in cluster 2 of the model with $C = 2$. This shows that treating all regions homogeneously, and assigning an average effect on them, is a very dangerous practice for regional models. In contrast, (especially since regional data are not abundant) using a parsimonious model with just two levels of heterogeneity can give more informative estimates of the true determinants of unemployment.

In that respect, we can also observe that the signs and magnitudes of the coefficients $\beta_j$ are relatively different between the two clusters when we assume $C = 2$. For the ESERV variable in particular we observe significant coefficients with opposite signs[3], meaning that regions belonging in cluster 1, and specialize in the services sector, present higher unemployment rates compared with regions in cluster 2. This is a very interesting result, since as we will see below, the regions in the high unemployment group (cluster 1) are the ones which host some of the largest cities in the UK (whose structure obviously depends more on services employment).

For the rest of the variables the magnitudes vary to some degree between the two clusters, but the signs are the ones expected. Regions which generate more opportunities for employment should enjoy lower unemployment rates, as it is revealed from the negative sign of employment growth (EMPLG). The employment shares in manufacturing (EMAN) and agriculture (EAGR) are all negative and significant in both models. A high number of students that have at least started high school as a proportion of working age population (HCAP) implies that their participation in the labour force will be delayed and unemployment is expected to be higher. Nevertheless, note that the expected negative sign shows up in the two-cluster model but not in the model with one cluster. Male (MALP) and female (FEMP) participation rates usually show up in the literature with a negative sign (see the review in Elhorst, 2003), which is the case in Table 2 as well. Unemployment (especially in European regions, including UK regions) tends to be very high among young persons (YOUNG) as opposed to older population (OLD), while expectations about the sign of the two last demographic variables (MIG and DENS) are mixed in the literature (Elhorst, 2003). For instance, a higher density of a region might imply a better matching between jobs and workers, but at the same time there might be a higher cost of congestion for firms and workers (Patridge and Rickman, 1997). In our models MIG is positive, while DENS has a positive sign on the one-cluster model and a negative sign on the low unemployment group of the two-cluster model (in the high unemployment group the coefficient is positive but insignificant).

Table 3 and Figure 2 show quantitatively and graphically (respectively) the distribution of

[2]We remind that we have achieved this by substracting the sample mean from the variables $x_t$ so that they have expectation zero. Subsequently it holds that $E\left(\mu_j\right) = E\left(\alpha_j + \beta_j x_t\right) = \alpha_j + \beta_j E\left(x_t\right) = \alpha_j$, for $j = 1, 2$.

[3]The coefficient of ESERV for the standard panel data model is not statistically different from zero. Given the estimates of the coefficients $\left(\beta_1^{ESERV}, \beta_2^{ESERV}\right) = (-0.078, 0.058)$ of the model with two clusters (and the estimates of the weights/probabilities $(p_1, p_2) = (0.41, 0.59)$) we can see that the average effect of this variable is: $E\left(\widetilde{\beta}^{ESERV}\right) = \left(\beta_1^{ESERV} \times p_1\right) + \left(\beta_2^{ESERV} \times p_2\right) \approx 0$, which gives a rough idea (since we would need to take into account the standard errors in order to properly test whether $E\left(\widetilde{\beta}^{ESERV}\right) = 0$ holds in a statistical sense) why failing to account for heterogeneity can be misleading.

the 35 regions into clusters. The last column of Table 3 shows the sample mean of the unemployment rates (in levels, not in logs) over the sample period 1999-2008. We can observe a few interesting things. First, the model-based separation into high and low unemployment rate clusters can be confirmed from the sample means. All regions which belong in cluster 1 show remarkably high unemployment rates (5%+), while the regions in the second cluster show very good performance, with average rates lower than the national average of 5.2% for that period. Second, the only exemption to this rule is the Eastern Scotland region (Edinburgh) which, while it has an average rate of 5.4% (comparable to Northern Ireland, Derbyshire, West Yorkshire and Greater Manchester), it belongs in the low unemployment region. This result is actually robust to different specifications we tried by adding/removing variables. In fact in 100% of the specifications we experimented with, it is always the case that the two regions of Scotland UKM2 (Edinburgh) and UKM3 (Glasgow) are well separated into different clusters. Third, if we exclude the area of London, the high-unemployment areas are mainly in the Middlands and Highlands, while the whole of the Lowlands is characterized by low unemployment areas. However, this kind of separation is not so strong to suggest a geographical distinction in the economic performance (as we measure it by unemployment rates) between "the south" and "the north" of the UK.

Finally Table 4 presents the other model parameters, $\sigma_i^2$ and $\delta_i$ for $i = 1, ..., 35$ (one for each region), along with their standard errors. Remember that we have used the factor model (4) to decompose the covariance matrix $\Omega$ into the "common component" $\chi = \delta_i f_t$, and the innovation error $\varepsilon_{it}$ which has variance $\sigma_i^2$ (see for instance Latin et al. 2003). Our factor $f_t$ here is observed (national unemployment rates), but a latent factor (principal component) could have been estimated. The $\delta_i$'s can be interpreted as "loadings" i.e. a parameter vector that shows which regions load on the national factor. Subsequently, $\delta_i^2$ is the variance explained by the common component, and $\sigma^2$ the variance left unexplained by the whole model. Hence, the last column of Table 4 estimates the quantity $100 \times \left(\delta_i^2/(\delta_i^2 + \sigma_i^2)\right)$, which shows the total proportion of the variance in regional unemployment rates explained by the common component. These values are quite high for most regions, which signifies the high correlation between national and regional unemployment rates as documented in Elhorst (2003). Notable exemptions from this rule are the regions of Shropshire and Staffordshire (in West Midlands), and Cornwall and Isles of Scilly (in South West).

# 4 Conclusions and policy implications

We study the structural differences in regional unemployment for 35 UK regions at NUTS-2 level for the period 1999 – 2008, using an endogenous classification procedure based on finite mixtures. Crucially, our approach has the implication that the grouping of regions within homogeneous clusters is purely data-driven.

We find that there is evidence of two clusters of regions in the UK, characterized by high and low unemployment rates respectively. As with other studies, we confirm the empirical finding that regional unemployment determinants ultimately are related to factors that affect labour market equilibrium (demand/supply), disequilibrium and other demographic factors. However, between clusters, the results show that the specifics of the determinants differ significantly. Firstly, we find that some potential determinants are significant for one region but not

the other. Secondly, even where the same determinant is significant for both regions, the size and sign of the coefficient differ. This suggests that the unemployment dynamics vary across the two clusters, and these differences are not captured by traditional regressions analyses that group regions with different fundamental characteristics together. Furthermore, we find that geographical proximity does not necessarily provide a good indicator for grouping purposes: structural differences are found to exist between sub-regions which belong to the same broad geographical region.

The major source of heterogeneity between the two clusters seems to come from the share of employment in services. This effect is closely related to the nature of the regions classified in the two clusters. More specifically, the high unemployment cluster comprises regions which host the most populous cities in the UK. Hence this cluster includes large urban centers like: London (UKI1, UKI2), Birmingham (UKG3), Leeds (UKE4), Glasgow (UKM3), Sheffield (UKE3), Bradford (UKE4), Manchester (UKD3) and Liverpool (UKD5). This is an exciting pattern which shows that, even after controlling for the 10 factors affecting regional unemployment rates (EMPLG, EAGR, ESERV, etc.), our model predicts that the degree of urbanization is a very important factor of unemployment heterogeneity in the UK.

## References


Baltagi B. H. (2005). *Econometric analysis of panel data*. 3rd edition, John Wiley and Sons, Chichester.

Basturk N., Paap R. and van Dijk D. J. C. (2008). Structural differences in economic growth, Tinbergen Institute Discussion Paper, TI 2008-085/4, Erasmus University Rotterdam.

Baumont C., Ertur C. and Le Gallo J. (2003). Spatial Convergence Clubs and the European Regional Growth Process, 1980-1995. In: Fingleton B. (Ed.), *European Regional Growth*, Springer, Berlin, Heidelberg, New York, pp. 131-158

Bean C. R. (1994). European unemployment: A survey. *Journal of Economic Literature* 32, 573-619.

Bozdogan H. (1994). Mixture-model cluster analysis using model selection criteria and a new information measure of complexity. In: Bozdogan, H. (Ed.), *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, vol. 2. Kluwer, Boston, pp. 69-113.

Canova F. (2004). Testing for convergence clubs in income per capita: A predictive density approach. *International Economic Review* 45, 49-77.

Cracolicci M. F., Cuffaro M. and Nijkamp P. (2007). Geographical Distribution of Unemployment: An analysis of Provincial Differences in Italy. *Growth and Change* 38, 649-670.

Crone T. M. (2005). An alternative definition of economic regions in the United States based on similarities in State business cycles. *The Review of Economics and Statistics* 87, 617-626.

Elhorst J. P. (2003). The mystery of regional unemployment differentials: A survey of theoretical and empirical explanations. *Journal of Economic Surveys* 17, 709-748.

Fischer M. and Stirböck C. (2004). Regional Income Convergence in the Enlarged Europe, 1995-2000: A Spatial Econometric Perspective. Centre for European Economic Research, Discussion Paper No. 04-42.

Früwirth-Schnatter S. and Kaufmann S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics* 26, 78-89.

Lattin, J., Carroll, J.D. and Green, P.E. (2003). *Analyzing Multivariate Data*. Thomson Learning, Pacific Grove CA.

Lopez-Bazo E., del Barro E. T. and Artis M. (2005). Geographical distribution of unemployment in Spain. *Regional Studies* 39, 305-318.

McLachlan G. J. and Krishnan T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons, New York.

Overman H. G. and Puga D. (2002). Unemployment Clusters Across Europe's Regions and Countries. *Economic Policy* 34, 115-147.

Paap R., Franses P. H. and van Dijk D. (2005). Does Africa grow slower than Asia, Latin America and the Middle East? Evidence from a new data-based classification method. *Journal of Development Economics* 77, 553-570.

Partridge M. D. and Rickman D. S. (1997). The Dispersion of US State Unemployment Rates: The Role of Market and Non-Market Equilibrium Factors. *Regional Studies* 31, 503-606.

Quah D. T. (1996). Regional Convergence Clusters across Europe. *European Economic Review* 40, 951-958.

Taylor J. and Bradley S. (1997). Unemployment in Europe: A Comparative Analysis of Regional Disparities in Germany, Italy and the UK. *Kyklos* 50, 221-245.

Tsionas E.G. (2000). Regional Growth and Convergence: Evidence from the United States. *Regional Studies* 34, 231–38

# Appendix: Tables and Figures

Figure 1: Unemployment rates for the five NUTS-II regions belonging to the Nortwest area, 1999-2008.

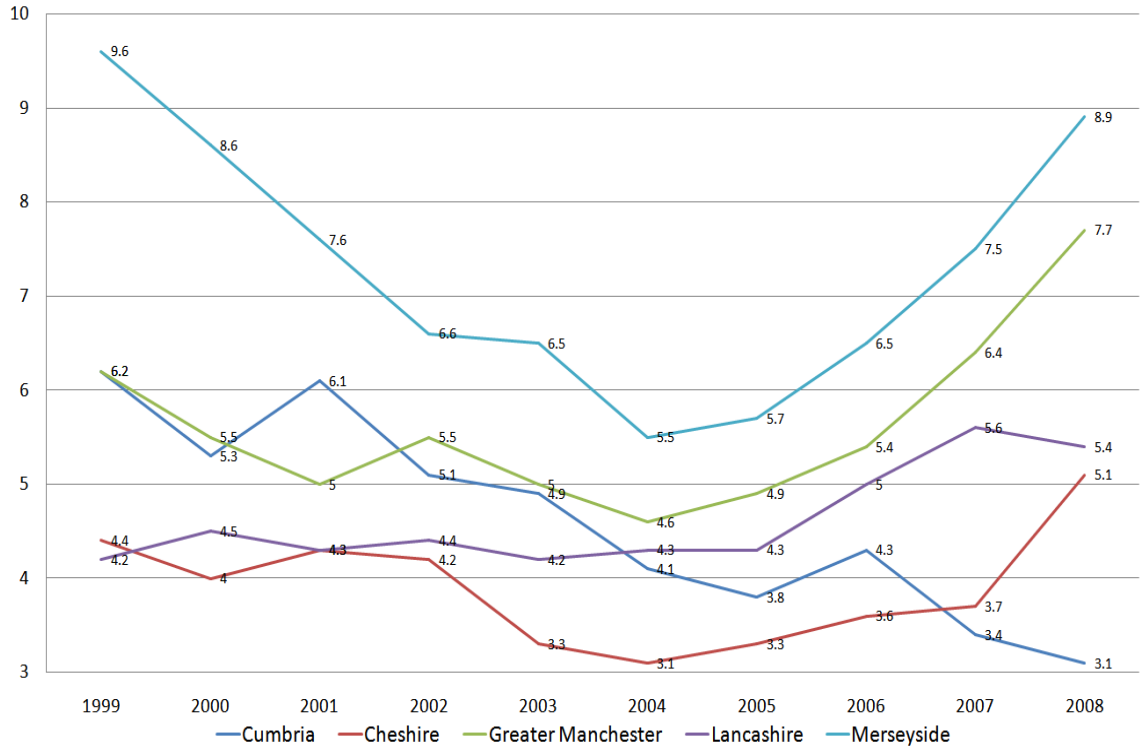


Table 1: Cluster selection based on information criteria

| | C=1 | C=2 | C=3 | C=4 | C=5 | C=35 |
|---|---|---|---|---|---|---|
| Information criteria with number of series as number of observations | | | | | | |
| AIC | -17.850 | **-18.343** | -18.072 | -18.251 | -18.108 | -9.711 |
| BIC | **-12.695** | -12.655 | -12.651 | -12.296 | -11.886 | 4.243 |
| AIC-3 | -14.535 | **-14.686** | -14.586 | -14.423 | -14.108 | -0.739 |
| AIC-3 adj | -17.764 | **-18.143** | -17.929 | -17.994 | -17.794 | -7.739 |
| Information criteria with total number of observations | | | | | | |
| AIC | -1.983 | **-2.038** | -2.008 | -2.028 | -2.012 | -1.079 |
| BIC | **-0.601** | -0.513 | -0.555 | -0.432 | -0.344 | 2.662 |
| AIC-3 | -1.615 | **-1.632** | -1.621 | -1.603 | -1.568 | -0.082 |
| AIC-3 adj | -1.974 | **-2.016** | -1.992 | -1.999 | -1.977 | -0.860 |

Table 2. Cluster-specific parameters

| | Model 1 ($C = 2$) | | Model 2 ($C = 1$) |
|---|---|---|---|
| Coef. on Variables | Cluster 1 | Cluster 2 | |
| | $\alpha_1$ | $\alpha_2$ | $\alpha$ |
| Intercept | 1.633 | 1.572 | 1.559 |
| | (124.9) | (230.1) | (157.8) |
| | $\beta_1$ | $\beta_2$ | $\beta$ |
| EMPLG | -0.135 | -0.237 | -0.196 |
| | (4.66) | (5.84) | (5.05) |
| EAGR | -0.218 | -0.202 | -0.336 |
| | (3.70) | (18.43) | (3.26) |
| EMAN | -0.157 | -0.081 | -0.433 |
| | (9.70) | (5.51) | (3.78) |
| ESERV | 0.078 | -0.058 | -0.108 |
| | (6.12) | (4.29) | (1.46) |
| HCAP | -0.048 | -0.018 | 0.036 |
| | (2.52) | (1.38) | (5.29) |
| FEMP | -0.071 | -0.101 | -0.088 |
| | (9.93) | (14.75) | (21.67) |
| MALP | -0.085 | -0.115 | -0.103 |
| | (22.85) | (80.66) | (35.36) |
| YOUNG | 1.359 | 0.322 | 0.704 |
| | (1.04) | (0.91) | (0.71) |
| OLD | -2.993 | -2.078 | -2.522 |
| | (1.87) | (8.42) | (9.97) |
| MIG | 1.673 | 1.214 | 0.491 |
| | (1.17) | (3.83) | (1.86) |
| DENS | 0.012 | -0.205 | 0.022 |
| | (0.73) | (2.68) | (1.17) |
| | $\varphi_1$ | $\varphi_2$ | $\varphi$ |
| autoregr. coeffcient | 0.821 | 0.828 | 0.872 |
| | (15.23) | (19.43) | (12.85) |
| | $p_1$ | $p_2$ | $p$ |
| mixing proportions | 0.41 | 0.59 | 1.00 |
| Log-likelihood | 796.51 | | 750.42 |

Table 3. Description of regions

| NUTS region | Code | Cluster Membership | Average Unemployment |
|---|---|---|---|
| **North East** | UKC | | |
| Tees Valley and Durham | UKC1 | 1 | 7.21 |
| Northumberland, Tyne and Wear | UKC2 | 1 | 7.04 |
| **North West** | UKD | | |
| Cumbria | UKD1 | 2 | 4.63 |
| Cheshire | UKD2 | 2 | 3.90 |
| Greater Manchester | UKD3 | 1 | 5.62 |
| Lancashire | UKD4 | 2 | 4.62 |
| Merseyside | UKD5 | 1 | 7.30 |
| **Yorkshire and the Humber** | UKE | | |
| East Yorkshire & Northern Lincolnshire | UKE1 | 1 | 6.21 |
| North Yorkshire | UKE2 | 2 | 3.30 |
| South Yorkshire | UKE3 | 1 | 6.30 |
| West Yorkshire | UKE4 | 1 | 5.49 |
| **East Midlands** | UKF | | |
| Derbyshire & Nottinghamshire | UKF1 | 1 | 5.14 |
| Leicestershire, Rutland & Northants | UKF2 | 2 | 4.50 |
| Lincolnshire | UKF3 | 2 | 4.74 |
| **West Midlands** | UKG | | |
| Herefordshire, Worcestershire & Warks | UKG1 | 2 | 3.74 |
| Shropshire and Staffordshire | UKG2 | 2 | 4.49 |
| West Midlands | UKG3 | 1 | 7.77 |
| **Eastern** | UKH | | |
| East Anglia | UKH1 | 2 | 4.08 |
| Bedfordshire, Hertfordshire | UKH2 | 2 | 3.98 |
| Essex | UKH3 | 2 | 4.21 |
| **London** | UKI | | |
| Inner London | UKI1 | 1 | 8.66 |
| Outer London | UKI2 | 1 | 6.01 |
| **South East** | UKJ | | |
| Berkshire, Bucks and Oxfordshire | UKJ1 | 2 | 3.44 |
| Surrey, East and West Sussex | UKJ2 | 2 | 3.66 |
| Hampshire and Isle of Wight | UKJ3 | 2 | 3.82 |
| Kent | UKJ4 | 2 | 4.78 |
| **South West** | UKK | | |
| Gloucestershire, Wiltshire & Bristol/Bath | UKK1 | 2 | 3.38 |
| Dorset and Somerset | UKK2 | 2 | 3.61 |
| Cornwall and Isles of Scilly | UKK3 | 2 | 4.89 |
| Devon | UKK4 | 2 | 4.45 |
| **Wales** | UKL | | |
| West Wales and The Valleys | UKL1 | 1 | 5.91 |
| East Wales | UKL2 | 2 | 4.80 |
| **Scotland** | UKM | | |
| Eastern Scotland | UKM2 | 2 | 5.39 |
| South Western Scotland | UKM3 | 1 | 6.91 |
| **Northern Ireland** | UKN | | |
| Northern Ireland | UKN1 | 1 | 5.30 |

Table 4. Other model parameters

| Region | $\sigma_i^2$ | s.e. $\sigma_i^2$ | $\delta_i$ | s.e. $\delta_i$ | $100\left(\frac{\delta_i^2}{(\sigma_i^2+\delta_i^2)}\right)$ |
|---|---|---|---|---|---|
| Tees Valley and Durham | 0.00028 | 0.00001 | -0.518 | 0.095 | 100 |
| Northumberland, Tyne and Wear | 0.00017 | 0.00003 | -0.312 | 0.048 | 100 |
| Cumbria | 0.00055 | 0.00023 | 0.101 | 0.149 | 95 |
| Cheshire | 0.00043 | 0.00018 | 0.263 | 0.118 | 99 |
| Greater Manchester | 0.00017 | 0.00004 | 0.062 | 0.048 | 96 |
| Lancashire | 0.00025 | 0.00003 | 0.059 | 0.103 | 93 |
| Merseyside | 0.00029 | 0.00003 | -0.341 | 0.098 | 100 |
| East Yorkshire and Northern Lincolnshire | 0.00007 | 0.00002 | -0.062 | 0.000 | 98 |
| North Yorkshire | 0.00605 | 0.00112 | 0.513 | 0.425 | 98 |
| South Yorkshire | 0.00124 | 0.00059 | -0.324 | 0.182 | 99 |
| West Yorkshire | 0.00013 | 0.00008 | 0.234 | 0.052 | 100 |
| Derbyshire and Nottinghamshire | 0.00004 | 0.00029 | 0.302 | 0.000 | 100 |
| Leicestershire, Rutland and Northants | 0.00059 | 0.00027 | 0.032 | 0.138 | 64 |
| Lincolnshire | 0.00031 | 0.00014 | -0.049 | 0.141 | 88 |
| Herefordshire, Worcestershire & Warks | 0.00203 | 0.00098 | 0.472 | 0.253 | 99 |
| Shropshire and Staffordshire | 0.00012 | 0.00001 | 0.008 | 0.086 | 33 |
| West Midlands | 0.00018 | 0.00005 | -0.267 | 0.051 | 100 |
| East Anglia | 0.00009 | 0.00000 | -0.059 | 0.058 | 97 |
| Bedfordshire, Hertfordshire | 0.00013 | 0.00002 | 0.208 | 0.074 | 100 |
| Essex | 0.00072 | 0.00033 | 0.270 | 0.155 | 99 |
| Inner London | 0.00053 | 0.00030 | -0.330 | 0.131 | 100 |
| Outer London | 0.00015 | 0.00004 | -0.021 | 0.071 | 75 |
| Berkshire, Bucks and Oxfordshire | 0.00097 | 0.00045 | -0.035 | 0.183 | 56 |
| Surrey, East and West Sussex | 0.00076 | 0.00036 | 0.170 | 0.153 | 97 |
| Hampshire and Isle of Wight | 0.00151 | 0.00072 | 0.159 | 0.218 | 94 |
| Kent | 0.00007 | 0.00000 | 0.038 | 0.069 | 96 |
| Gloucestershire, Wiltshire & Bristol/Bath area | 0.00060 | 0.00028 | -0.031 | 0.136 | 62 |
| Dorset and Somerset | 0.00506 | 0.00241 | 0.186 | 0.413 | 87 |
| Cornwall and Isles of Scilly | 0.00014 | 0.00002 | 0.006 | 0.107 | 21 |
| Devon | 0.00057 | 0.00027 | -0.093 | 0.139 | 94 |
| West Wales and The Valleys | 0.00029 | 0.00012 | 0.048 | 0.088 | 89 |
| East Wales | 0.00026 | 0.00011 | 0.114 | 0.119 | 98 |
| Eastern Scotland | 0.00001 | 0.00000 | -0.225 | 0.005 | 100 |
| South Western Scotland | 0.00022 | 0.00008 | -0.037 | 0.065 | 86 |
| Northern Ireland | 0.00058 | 0.00027 | -0.123 | 0.169 | 96 |

Note: s.e. stands for standard error.

Figure 2: Map showing the geographical distribution of the data-based estimates of the cluster membership for each NUTS-II region

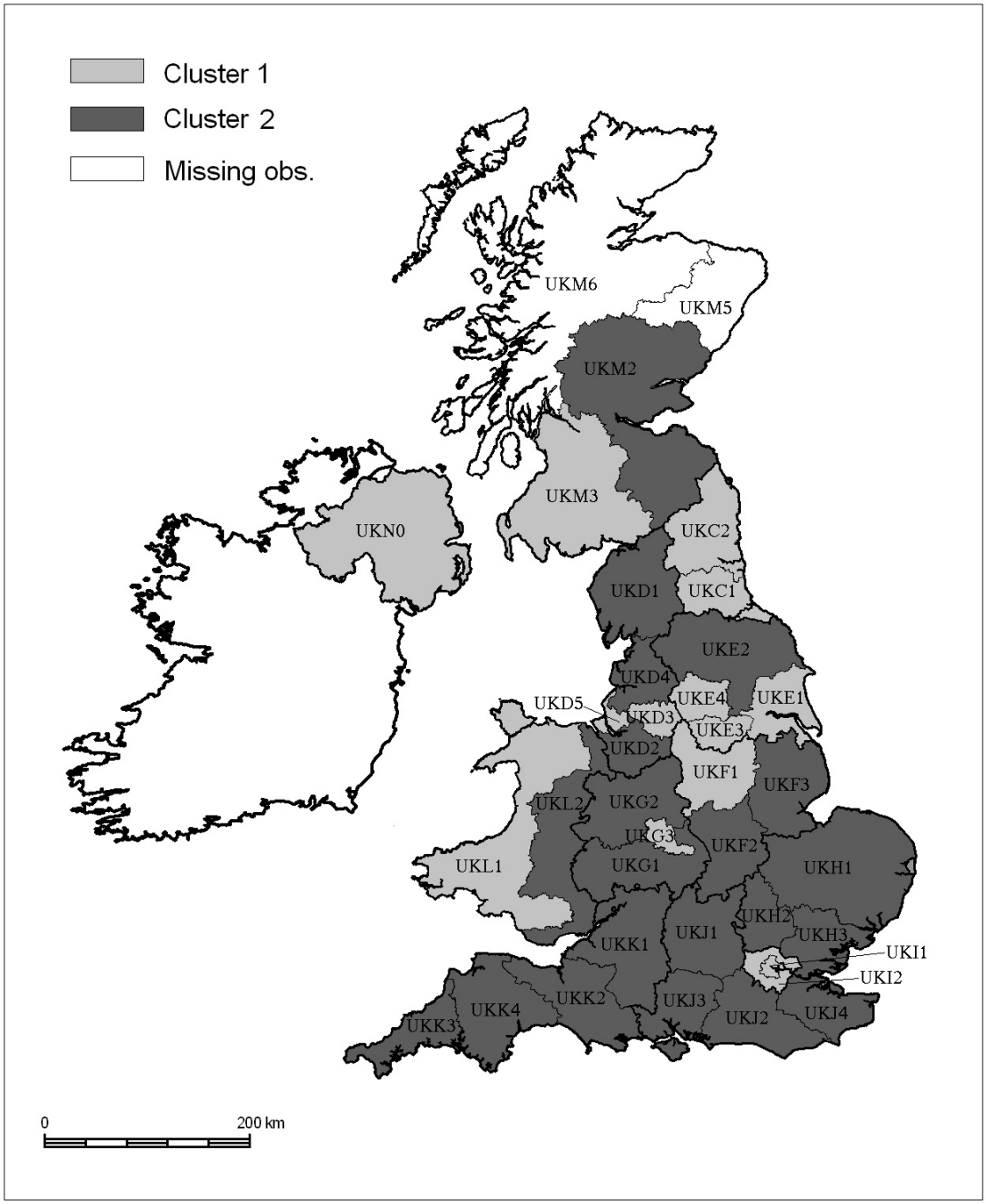


Figure 3: