EXTENDED ABSTRACT

Title: 'New estimates for Total Factor Productivity across Spanish regions using spatial dependence'

Authors and e-mails: Marco Palomeque (marco.palomeque@uah.es)
Andrés Maroto Sánchez (andres.maroto@uam.es)

Department: Análisis Económico

University: Universidad Autónoma de Madrid y Universidad de Alcalá

Subject area: Competitividad, eficiencia y productividad

Abstract:

Improvements in Total Factor Productivity are always associated to economic development. For this reason, the study of TFP is a central topic in economics. More recently, emphasis shifted towards explaining regional inequalities. This is especially important for Spain, whose regions are performing poorly since the mid-90s. To date, every analysis of the productivity of Spanish regions have neglected the important role of geography. To fulfil this lack of research, a measure of Total Factor Productivity (TFP) is derived based on the estimation of Spatial Lag Models for a panel of the Autonomous Communities (1991-2020) and Provinces (1995-2018) of Spain, using data from the National Institute of Statistics (INE). Subsequently, the effects of human capital, knowledge capital and public capital on productivity are explored following a similar methodology. After controlling for spatial autocorrelation, all variables have a positive significant effect on the regional levels of TFP. Furthermore, the results suggest that the spillovers of knowledge capital and public capital are important determinants of the disparities in TFP across Spanish regions. Lastly, the distance range within which productivity spillovers have their major influence is investigated. Particularly, most of the diffusion of productivity takes place between bordering regions.

Keywords: *Total Factor Productivity, Spanish regions, Spatial Lag Model, Spillovers*
JEL codes: C21, C23, C31, 047, R11
(Working progress paper)

1. Introduction

The study of Total Factor Productivity (TFP) is a key issue for economic researchers and policy makers. The primary reason is that TFP enhancements are always associated to processes of economic growth. More approaches that are recent have placed special emphasis on explaining inequalities within a country. This has been favoured by the availability of new and more reliable regional data, as well as the rising concern of authorities in reducing the gap. The problem is particularly critical for Spain since the country is going through decades of low productivity growth relative to other EU countries, problem that also remains widespread across regions. Although this fact is highlighted by multiple authors (De la Fuente, 2002; Cuadrado and Maroto, 2012), none of them considers the statistical influence of the spatial patterns that are present in the data. To fulfil such lack of empirical research, this paper addresses the estimation of TFP levels and the analysis of their determinants for the Spanish regions, explicitly controlling for the presence of spatial dependence.

According to the spatial econometrics literature, the omission of the spatial relationships existing within the sample seriously affects the consistency of the estimates. Ignoring the impact of spillovers thus has severe implications for the analysis of TFP: the effects of a region's intrinsic characteristics might be overestimated, since the contribution from the outside factors is neglected. This is the case for the Spanish regions. A glance at the geographical distribution of the variables involved in the production process reveals a strong tendency towards clustering. Moreover, the potential determinants of the remaining TFP gap show a similar agglomeration pattern. Accordingly, an important part of the regional TFP and income level would be originated from the outside.

Within this context, the objective of this paper is twofold: First, to derive a consistent measure of TFP based on the estimation of the production functions using spatial econometrics techniques. The research hypothesis is that controlling for spatial autocorrelation when estimating TFP raises the consistency of the measurement. Secondly, we disentangle the role played by human, knowledge and public capital in explaining TFP disparities across Spanish regions. In particular, two research hypotheses will be contrasted: i) Productivity is the main actor explaining regional differences in the Spanish economy; and ii) Capital spillovers account for more the combined contribution from inputs in the production process.

The empirical procedure is as follows. First, the production functions are estimated through Spatial Lag Models using data on value added, productive capital, labour units and geographic distances between regions for the Spanish provinces, corresponding to the *pre-crisis* years. Previously, the underlying presence of spatial dependence patterns in the data is revealed using spatial analysis techniques. Then using the estimation of TFP, and a purely numeric index as dependent variables, the effects of human capital, knowledge capital and public capital on regional productivity levels are investigated. Since spatial dependence is identified in the variables selected as determinants of TFP, again the estimation method requires controlling for it.

The key finding is that spillovers account for roughly two times the combined contribution from capital and labour in the production process, once spatial autocorrelation is properly controlled. Therefore, geography has a major influence in determining income and productivity disparities across the Spanish territory. Moreover, the main factors that have a positive impact on regional productivity levels are human capital and particularly, knowledge capital. The latter, measured in terms of patents, plays the most significant role in determining the productivity of a region, especially when its effects arise from the outside. Besides, human capital also has a positive significant effect, although it is not spatially dependent. Furthermore, the results suggest that productivity spillovers could have almost no influence over areas located more than 400 km away. Indeed, the results suggest that most of the effect from productivity spillovers arise between adjacent regions.

The structure of this paper is outlined below. Section 2 reviews the economic literature about the study of TFP and its determinants at the country and regional levels and introduces some concepts in spatial econometrics. Subsequently, the methodology applied for measuring TFP is described in section 3. Section 4 addresses the descriptive analysis of the variables, how the data is spatially distributed and other statistical issues. Then, section 5 discusses the results from the production function estimation and the derived TFP measures. Section 6 provides the description of the data on the determinants, and the estimates of its role in explain regional differences in TFP. Finally, section 7 summarizes the conclusions extracted from the whole analysis.

2. Literature review

The measure of productivity[1] is a central question in empirical economics since productivity developments are always present in any process of economic growth. Indeed, a significant part of the income gap among countries and regions responds to differences in TFP, both in growth rates and levels (Hall and Jones, 1996 and 1999; Klenow and Rodriguez-Clare, 2005; Caselli, 2005; Gehrinzer et al., 2013). Focusing on the Spanish case, several works point out that the level of productivity, and mainly TFP, in Spain remains behind the levels of the majority of European countries and the USA (Domenech, 2008; Cuadrado and Maroto, 2012). This low performance of TFP is observed at the regional level too (De la Fuente, 2002; González-Páramo and Martínez-López, 2003).

In fact, the study of productivity at regional level has gained importance through the improvements in the availability and quality of regional data, and the increasing interest in policies directed towards reducing regional gaps (ONS, 2007). Ascari and Di Cosmo (2004) explore the determinants of TFP in the Italian regions. Using dynamic and static panel data estimation techniques, they show that R&D expenditures and human capital have a noteworthy influence over the regional levels of TFP, unlike other factors, such as ICT expenditure that is scarcely related to TFP.

Finally, according to Magrini (2004) the models in regional studies that treat regions in the same way as countries might be incorrectly specified. The reason is that economic barriers are virtually inexistent at the regional level, fact that calls for the inclusion of geographical effects in the model. Recently, there is a rise[2] in the number of analyses of regional inequality or convergence that implements the methods from the field of Spatial Econometrics (Anselin, 2010). For instance, convergence in GDP per capita for a panel of European regions can be analysed using models that include the spatial lag of the dependent variable or the spatially lagged error term (Fingleton and Lopez-Bazo, 2006). Paas and Vahi (2012) apply a similar framework to investigate regional inequality and growth using a sample of EU countries and regions at the NUTS3 level. Their main finding is that growth spillovers are not likely to transcend country boundaries, but rather to act between regions. Moreover, Arbia (2006) suggests that using different administrative aggregation levels in the analysis may severely alter the

1 See Maroto (2012) for a review of the literature on productivity and economic growth, and Maroto (2013) for a revision of the literature on productivity and regional and territorial developments.

2 Advances in the field of spatial econometrics respond to two reasons: in first place, the growing interest in modeling spatial interactions between heterogeneous agents and; second, because of the generalization of the geographic information system (GIS), which allows to analyze and manage distinct kinds of spatial data (Anselin, 2001).

results[3]. Di Liberto and Usai (2013) approach the problem of spatial dependence through the inclusion of a contiguity weights matrix in the model. According to Arbia and Fingleton (2006), among other common applications, spatial regression is used to study land prices or education inequalities.

However, the studies in spatial econometrics that analyse regional inequality in TFP and its determinants are still scarce. The work that resembles the most to this paper is Dettori et al. (2012). Using spatial lags models, they estimate a traditional Cobb-Douglas production function in order to obtain a measure of TFP for a panel of 199 European regions during the period 1985-2006. Then, with the help of a similar model for a cross section of year 2004, they found evidence on the significant role of the 'intangible assets' (human capital, technological capital, knowledge capital and social capital) in explaining TFP disparities across regions. Moreover, Bronzini and Piselli (2009) test the long run relationship between the levels of TFP, human capital, R&D capital, and public infrastructures for a panel of the Italian regions from 1980 to 2001, including spatial effects and departing from a standard Cobb-Douglas production function. They find a positive relation between TFP and the three forms of capital.

Moreover, LeSage and Fischer (2009) build a TFP index as the ratio of gross value added over the combined stock of capital and labour in order to estimate the effect of knowledge capital, expressed in patents, using Spatial Lags models. They found significant effects of knowledge capital, which are positively related to spatial proximity. Recently, Bos et al. (2014) use Spatial Lag Model to estimate productivity spillovers between Indian manufacturing firms including data on headquarters location. Finally, Caliendo and Parro (2015) estimate TFP levels across sectors and regions for the US economy, to show the important effects of spatial dependence on welfare and productivity. Particularly, they found that interregional and inter-sectorial trade linkages have an important role in the diffusion of changes in productivity.

3. Methodology: Measuring TFP

Productivity is commonly defined as the quotient between output and inputs. Usually, the choice of the optimal expression for productivity depends on the purpose of the researcher and the availability of data. In its simplest expression, output is related to one input, commonly labour. Thus, all workers are incorrectly assumed to share the

[3] The production functions in section 5 are estimated at two different scales to add robustness to the research.

same characteristics, affecting the overall analysis of efficiency. When both capital and labour are included in the denominator, the resulting relation is known as[4] Multi-Factor Productivity (MFP) or TFP (ONS, 2007). When interpreted most broadly, the concept of TFP refers to the contribution from all unobservable inputs to the production process, or simply as a residual (Comín, 2006).

In the present study, the problem of measuring TFP is faced through two different approaches. The first method consists of a numeric TFP index built under the assumption of constant returns to scale using value added, capital stock and labour. This is a common assumption in the in the literature on TFP, especially in papers exploring its determinants (Hall and Jones, 1999; Ascari and Di Cosmo, 2005; among others). Moreover, this calculation is used in the present study in order to investigate the spatial interactions of productivity and particularly, to analyse the determinants of TFP in presence of spatial autocorrelation (section 6).

The relationship between the variables within the traditional neoclassical framework is defined by equation 1:

$$VA_{rt} = TFP_{rt}\, F(K_{rt}, L_{rt}) \qquad (1)$$

Where VA_{rt} is the value added at constant prices for the region *r* in the year *t*, K_{rt} is the stock of productive capital at constant prices and, L_{rt} is labour units. In order to compute a TFP index, the last expression takes the form of a standard Cobb-Douglas production function (equation 2):

$$VA_{rt} = TFP_{rt}\, K_{rt}^{\alpha_r}\, L_{rt}^{1-\alpha_r} \qquad (2)$$

Now, α_r is the elasticity of output (*GVA*) to changes in the workforce in the region *r* and it is assumed to be constant. Leaving the TFP term alone on one side, the final expression for productivity (equation 3) is obtained:

$$TFP_{rt} = \frac{VA_{rt}}{K_{rt}^{\alpha_r}\, L_{rt}^{1-\alpha_r}} \qquad (3)$$

For the 17 AACC, data on the share of labour rents in the GVA of each territory can be found in the database from De la Fuente (2015a). However, for the 50 Provinces there is no data for such variable. According to Ascari and Di Cosmo (2005), the share of the labour rents in the GVA can be obtained through the equation 4 below:

[4] Both terms are often used without discrimination in empirical studies although TFP should refer to when all inputs related to the production process are considered while MFP should refer to approaches that introduce more than one input (*partial productivity*) into the estimation.

$$\alpha_{rt} = \frac{\frac{CSE_{rt}}{AE_{rt}} L_{rt}}{GVA_{rt}} \qquad (4)$$

Where CAE_{rt} and AE_{rt} are the compensation of salaried employees and the total number of salaried employees, respectively; and GVA_{rt} is the gross value added. Thus, the quotient $\frac{CSE_{rt}}{AE_{rt}}$ is equivalent to the average wage of salaried employees. Then, it is used as a proxy for the average wage of all workers and multiplied by L_{rt} in order to compute the total labour rent of the region r at time t. To obtain the participation of labour in output, the last expression is divided by value added. Finally, α_r is calculated as the average value of α_{rt} due to the high variability of this variable[5].

The second approach is regression based. This original method has two advantages: i) not imposing any a priori restrictions in the coefficients of the production function inputs –capital and labour– and ii) allowing for including spatial components. Additionally, it is also broadly used in recent studies of TFP (Bronzini and Pisselli, 2006; Di Liberto and Usai, 2013; Gehringer et al., 2013; etc.). Currently, there are only a few examples in the literature applying this methodology to the analysis of TFP and its determinants (Dettori et al., 2012; Marijke et al., 2014).

The main novelty of the present study is the estimation of the production function for the Spanish regions considering the spatial interactions between the variables. For this purpose, three models have been estimated. Firstly, a traditional *ordinary least squares* model (OLS), serves as benchmarking and workhorse for the spatial models below. It is an OLS panel data regression including regional fixed effects[6] and year dummies. Taking logs in both sides of the production function, the resulting fixed effects OLS model is described by equation 5:

$$va_{rt} = \alpha_r + \beta_1 k_{rt} + \beta_2 l_{rt} + dummies_t + \varepsilon_{rt} \qquad (5)$$

Where the terms va_{rt}, k_{rt} and l_{rt} are the natural logs of the per capita values of the value added, capital and labour. For now, the presence of regional time invariant effects is assumed. Fixed effects are a reasonable intuition in view of the low variability of the TFP level of the regions, as it is shown in the descriptive analysis (section 4). In addition, this will be supported by the results from the Hausman test. The year dummies are included in order to control for temporary shocks affecting all the regions, although they are removed for the spatial models at the AACC level of aggregation level due to their low joint significance.

[5] Many authors choose a constant value, usually 0.3 (Hall and Jones, 1999; Klenow and Rodriguez-Clare, 2004).

[6] The inclusion of time invariant effects is common in the literature on regional economics (Ascari and Di Cosmo, 2004; Bronzini and Pisselli, 2006; Dettori et al., 2010, among others)

The spatial interaction patterns existing between the variables in the production function are investigated in the next section (4) by visualizing the distribution of the variables in the map and testing for spatial autocorrelation. In presence of spatial autocorrelation, the previous estimation has an endogeneity problem caused by the omission of spillovers arising from neighbouring regions. The implementation of the Spatial Lag models permits to overcome this issue. In the second model (equation 6), the spatially lagged dependent variable is included as an additional independent variable. Equation 6 describes the basic Spatial Lag Model, also known as the *Spatial Autoregressive model* (SAR):

$$va_{rt} = \alpha_r + \lambda W va_{rt} + \beta_1 k_{rt} + \beta_2 l_{rt} + dummies_t + \varepsilon_{rt} \qquad (6)$$

Where the term Wva_{rt} is the dependent variable pre-multiplied by the distance-based weights matrix W. This term can be interpreted as the average of the values of the dependent variable in the neighbouring locations (Anselin, 2001). The spatial autoregressive parameter for the dependent variable is indicated by λ. The parameters of the model should be estimated by maximum likelihood (LeSage, 1999). However, in view of the results of the Moran I test (section 4), which shows that the three variables in the production function are spatially correlated, a third specification of the production function is estimated through a *Spatial Durbin Model* (SDM). Testing the joint significance of the coefficients of the spatially lagged independent variable is a functional criterion to choose between the SAR model and the SDM model (Belotti et al., 2013). Subsequently, the production function regression is defined by equation 7 below:

$$va_{rt} = \alpha_r + \lambda W va_{rt} + \beta_1 k_{rt} + \beta_2 l_{rt} + \gamma_1 W k_{rt} + \gamma_2 W k_{rt} + dummies_t + \varepsilon_{rt} \qquad (7)$$

Where the parameters vectors for the spatial autoregressive independent variables are indicated by γ. In what respect to the weight matrix W, it consists of $N \times N$ elements corresponding to the distances between the centroids of the geometric areas described by the coordinates of the boundaries of regions r and s. The weights w_{rs} in the matrix are calculated through the squares of the inverse distances. The reasoning for this is that a variable observed in a certain area of the geographical space is more closely related to its outcome in surrounding locations than in distant locations. As it is common in the literature, the matrix is then row-standardized to have all the elements in a row summing to unity (Anselin, 2003). Moreover, the row-standardization facilitates interpretation, because the spatial lag of a variable represents the weighted average of the variable in neighbouring regions, and locates between 0 and 1. Finally, additional weights matrices are built using the distance ranges 0-400 km and 400-800 in order to analyse the scope of productivity spillovers in terms of distance, as well as a contiguity matrix[7], which permits to explore the cross-border effects (section 6).

4. Data and some introductory issues on Spanish regional productivity

The measurement of productivity is carried for provinces, which are equivalent to NUTS3[8]. According to this division, there are 50 Provinces in total. Geospatial data about Spanish territories is extracted in *shapefile* format from the GADM geographic database of global administrative areas[9]. The Autonomous Cities of Ceuta y Melilla are excluded from the study due to the presence of missing values negatively affecting the quality of the results[10]. In addition, both cities are located overseas and the squared inverse distances might wrongly capture such relation. However, the islands – Balearic and Canary – are maintained because their population is enough relative to the rest of regions, but also present similar levels of economic activity to be relevant for this study[11].

In what respect to the measurement of productivity, statistical data about output and factor inputs is collected from the Spanish statistics office (INE) and covers the period 1995-2020, which is the longest series available for provinces. There is only one exception, which is the data for the stock of productive capital measured in constant 2005 euros, provided by BBVA Foundation and IVIE (2020). Output data consist of the nominal gross value added (GVA) at basic prices/factor costs, also converted to constant 2005 euros using the GVA deflator calculated by ADF (2015a). At this level of aggregation, the number of employed people is the only variable on labour considered.

[7] The elements of the contiguity matrix take value 1 if a pair of regions shares the same border and 0 otherwise. This matrix is also row-standardized as it is common in the literature (LeSage, 1999).
[8] NUTS stands for the National Territorial Units for Statistics, a commonly used system to subdivide the area of the EU. The number 3 indicates the smallest level of aggregation.
[9] GADM database provides the coordinates of the regions limits at every level of aggregation. In particular, this file contains the geographical coordinates of the regions' boundaries that are required for mapping economic data and for building the distance-based weights matrices used in the econometric analysis. Using this file facilitates the management of spatial data through different statistical packages such as Stata. See www.**gadm**.org/.
[10] Neither are representative of the overall performance of Spanish regions because of their low population. Moreover, these are not taken into account in some examples in the literature on regional economics (De La Fuente, 2002).
[11] Since much of the effect of spillovers takes place across de border (section 6), and these regions are located relatively far away, their spatial influence might be negligible and not alter the reliability of the estimates. The calculated centroid of Balearic Islands is located at 244 km from Tarragona, its closest region. Distances larger than 1000 km separate the Canary Islands.

Finally, data on the compensation of employees from the INE serves to compute α_i (equation 4).

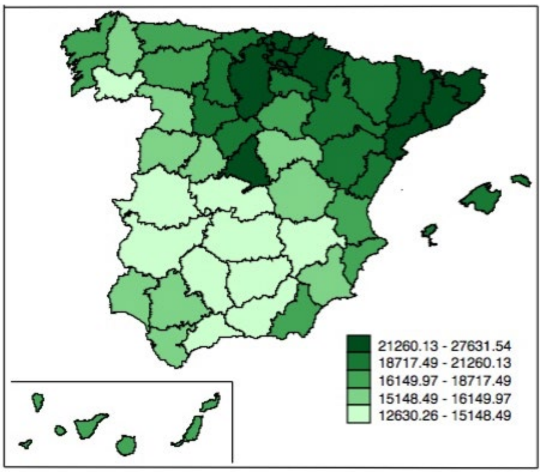
The graphical representation of the variables in the map serves as the first approach to the identification of spatial patterns in the data. Figure 1 displays the level in 2018 of value added, capital, labour and the TFP index (eq. 3) across the map of the Spanish Provinces, respectively. The colour intensity increases with the value of the variable[12]. All the variables are expressed in per capita terms[13]. In terms of value added, the richest regions are those located in the area around the River Ebro[14] (Northeast), Madrid (Centre), and Balearic Islands (East). The poorest regions are grouped in the Centre, South and West[15]. The distribution of productive capital is fairly similar but it appears somewhat more disperse around the centre but it is still concentrated in the Northern regions. The lowest levels of capital are mainly found in the South (Andalusia and Murcia). Employment, on the other hand, is distributed almost identically to value added, but even more concentrated in the East (Balearic Islands, Murcia and Valencian Community). Concerning the TFP index, similar values are slightly more disperse, but still some clusters are identifiable. The biggest groups are located in the North (Asturias, Basque Country, and Navarre) and Centre (Madrid, Toledo and their surrounding regions).

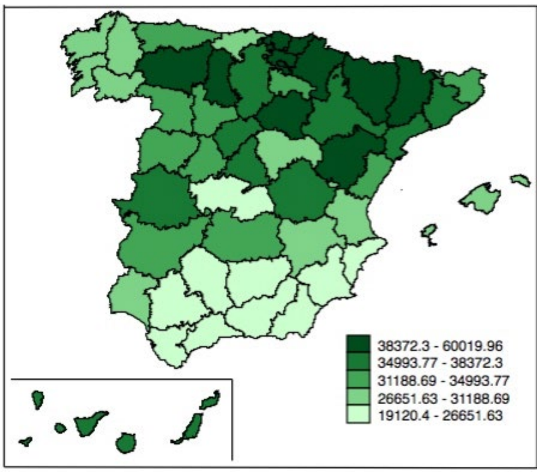
Figure 1. Value added in constant prices, stock of productive capital and TFP. NUTS3, 2018

Value added

Productive capital

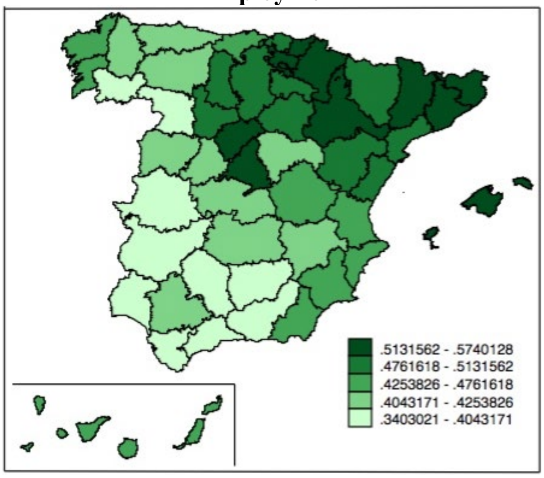
[12] In the case of the 17 AACC, the colours in the map reflect the quartile distribution. For the 50 Provinces, the variables are mapped accordingly to its quintile distribution.
[13] The source of the population series is the Spanish statistics office (INE).
[14] Area consisting of Cantabria, East part of Castile and León, La Rioja, Basque Country, Navarre, Aragon and Catalonia.
[15] Specifically, in the AACC of Andalusia, South of Castile-La Mancha and Extremadura

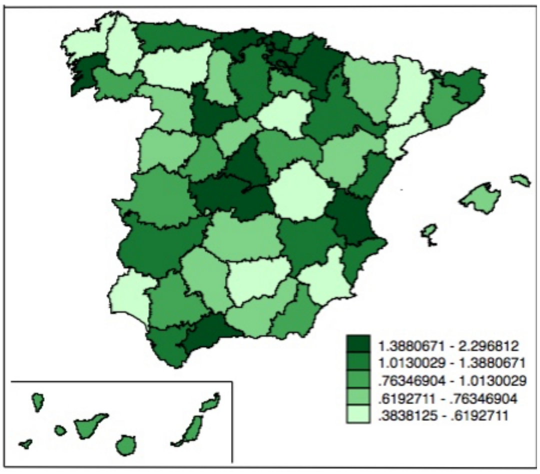




Employment

TFP index





Note: VA, K and labour are expressed in per capita terms.
Source: Based on data from the INE (2016).

To conclude, the maps of the Provinces suggest that spatial dependence patterns among the variables are more noticeable when studied at the lowest level of aggregation. The Moran I test serves to check the intuition extracted from the maps above. The I statistic is defined in equation 8:

$$I = \frac{N}{\Sigma_r \Sigma_s w_{rs}} \frac{\Sigma_r \Sigma_s w_{rs}(z_r - \bar{z})(z_s - \bar{z})}{\Sigma_r (z_r - \bar{z})^2} \qquad (8)$$

Where N is the number of observations, z is the variable of interest, $\bar{z}$ is its average and, w_{rs} is the element of the weights matrix W. The latest reflects the distance between region r and s. The elements in the diagonal take value 0 since $i = j$. As it is explained in section 3, the weights matrix is row-normalized and consists of the squared inverse distances. The expected value of the statistic is equal to $E(I) = \frac{-1}{N-1}$. In the case that $I > E(I)$, then there exists positive spatial autocorrelation. When $I < E(I)$, the spatial autocorrelation is negative (Griffith, 2009). Since the test is designed to study purely spatial autocorrelation, the results can be applied to every period. Thus, the reported I statistics is simply the average of all periods.

Table 1 below exhibits the Moran I test results. In the case of Provinces the spatial autocorrelation is statistically strong. The I statistics locate between 0.4 and 0.5, and all of them are significant at 1% level. Regarding the TFP, the spatial autocorrelation is positive. Although the Moran I test (*p* value= 0.315) fails to reject the null of no spatial autocorrelation, all the index components are spatially autocorrelated, which might cause the formation of clusters observed in the maps above.

Table 1. Moran I test: value added, capital, labour and TFP.

	I	E(I)	SD(I)	z score	p-value	CD	p-value
Value added	0.49929	-0.02	0.059	8.80145	0.000	25.233	0.000
Productive capital	0.45257	-0.02	0.059	8.00969	0.000	34.830	0.000
Workers	0.44521	-0.02	0.059	7.88499	0.000	18.998	0.000
Total Factor Productivity	0.00836	-0.02	0.059	0.48063	0.315	5.539	0.000

Note: VA, K and L are expressed in per capita terms. All the variables are in log form.
Source: Author's estimations.

To check for the robustness of the previous results, the Pesaran (2004) cross-section dependence (CD) test is applied[16]. The two methods are not related because the CD test does not account for physical distances. The CD statistic is built from the correlation of the residuals from a OLS regression of the variable on two of its own lags, for a pair of regions *r* and s, where $r \neq s$. The regressions include an intercept and a linear trend. The statistic is normally distributed under the null (cross-section independence). Equation 9 below describes the CD statistic:

$$CD = \sqrt{\frac{2T}{N-1} \sum_{r=1}^{N} \sum_{s=r+1}^{N-1} \widehat{corr}(u_r, u_s)} \qquad (9)$$

Where N is the number of observations and T the number of periods. The terms u_r and u_s are the residuals of the OLS regression. All the variables report high values of the CD statistic at the two levels of aggregation and are significant at 1% level. Therefore, the CD test indicates that there is correlation between pairs of observations located in different areas. Moreover, the results from the Moran I test suggest that spatial effects might cause it.

[16] See Dettori et al. (2010), among others.

5. Results:

5.1. Regional TFP estimates

In order to measure the regional levels of TFP, the production function is estimated by means of the models described in section 3. The results are listed in Table 2 below. In first place, the OLS fixed effects model defined by equation 5 is used as the basic framework (Column I). In this model, the dependent variable is the regional value added. Apart from the year dummies, the only independent variables are the conventional inputs - capital and labour. Due to the underlying endogeneity of the model (section 3) caused by the omission of spillovers, the next specification (equation 6) includes the spatially lagged dependent variable in the list of regressors (Column II). Lastly, the SDM from equation 7 is estimated (Column III). In this case, the model includes also spatial lags of the inputs. The joint significance of these terms indicates that the latter specification fits the data better than the SAR model, as explained in section 3. At the end of each column, the results of the Moran I test for residual spatial autocorrelation and the Hausman test are reported. All the variables are expressed in per capita terms.

In all the cases, the estimated coefficients for the inputs –capital and labour– are similar. In the case of the 50 Provinces (1995-2018), the OLS regression (IV) reports similar coefficients, although slightly above, to those obtained in the analogous model for the AACC. The parameters estimated, both significant at 1% level, are 0.1307 for capital and 0.3305 for labour. Now, the SDM (V) can produce consistent estimates of the production function since the spatial lags of the independent variables are jointly significant. The coefficients of the inputs are 0.1383 for capital and 0.3465 for labour, only slightly higher than those estimated by the SAR model (II). Again, both are significant at 1% level. Moreover, the estimated parameter for the spatially lagged dependent variable in the SDM (V) is close (0.5195) to the one obtained for the AACC (0.5367) and significant at 1 % level. Moreover, the coefficient estimated for the spatially lagged capital is relatively high (0.3120) and significant at 1% level, which might explain the low contribution from the capital accumulated within the boundaries of the region. However, the estimated parameter for the spatial lag of labour is small (-0.0229) and significant only at 10% level. However, the term is maintained without affecting the rest of parameters, once checked that all the spatial lags of the independent variable are jointly significant. Its sign might indicate some negative relationship, although weakly significant, between the value added and employment in neighbouring regions. Lastly, the R^2 of the OLS regression (IV) is 0.91, while that of the SDM is 0.92.

Table 2. Estimates of the production function

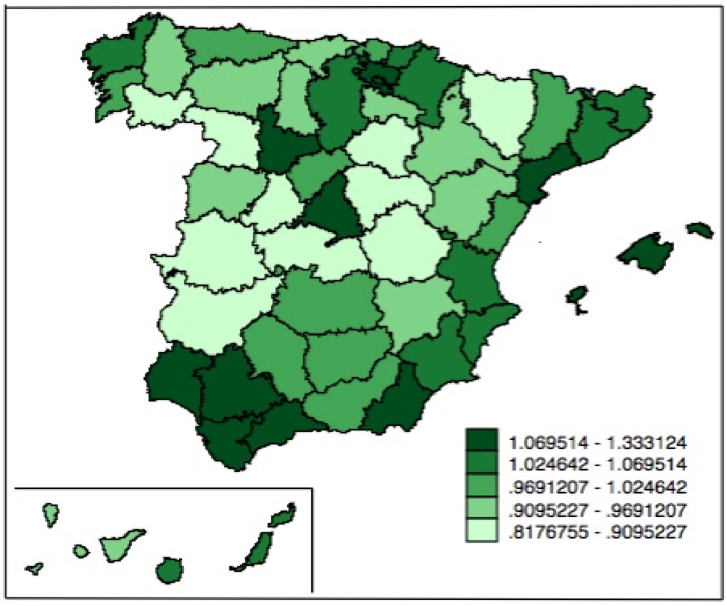
Variables/Model	(I) FE	(II) SAR	(III) SDM
Capital	0.1307*** (0.0191)	0.1434*** (0.0255)	0.1383*** (0.0201)
Workers	0.3305*** (0.0249)	0.2772*** (0.0366)	0.3465*** (0.0260)
Spatially lagged terms			
Value Added		0.5367*** (0.0528)	0.5195*** (0.0661)
Capital			0.3120*** (0.0633)
Workers			-0.0229* (0.0708)
Fixed Effects			
Regional	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes
Moran I test	0.388	-0.033	0.019
p value	*0.000*	*0.367*	*0.246*
Hausman Test	68.53	13.56	27.33
p value	*0.000*	*0.003*	*0.073*
R^2	0.91	0.91	0.92
No. obs.	700	700	700

Source: Author's estimations

In what respect to the presence of regional fixed effects, the results from the Hausman test support the previous assumption about the dynamics of the TFP. The Hausman test indicates that the fixed effects are significant in all the regressions. Overall, this result supports the interpretation of the TFP as a fixed component in the regressions. Finally, the Moran I test serves to check for the remaining presence of spatial autocorrelation in the fixed component of the residuals. In the case of the OLS regressions (I), the null of no spatial autocorrelation is always rejected. Therefore, the OLS estimators are not consistent. However, for the residuals from the Spatial Lag models (II and III) the null is not rejected, which implies that spatial autocorrelation is successfully removed from the TFP measure. As one can expect, the *I* statistic is much higher at the NUTS3 level, suggesting that the presence of spatial autocorrelation increases as the scale of the sample decreases.

Considering the results discussed above, the SDM produce the most consistent estimates of the production function. Because of the spatial interactions, it is possible to observe multiple changes in the spatial distribution of TFP in comparison to the index calculated excluding the spatial effects (section 4). In Figure 2 below, the map of Spain reports the quartile and quintile distribution of TFP levels for the NUTS3 regions. After taking into account the spatial effects, the map for the Spanish provinces is somewhat different with respect to the map in Figure 1. The distribution of similar values of TFP becomes more heterogeneous. Now, some of the Southern Provinces (Almeria, Cadiz, Malaga, Seville and Huelva) move to the highest quintile of the distribution. On the other hand, most elements in the bottom quintile (Badajoz, Caceres, Toledo, Avila, Soria and Cuenca) are now distributed across the middle part, close to the most efficient region (Madrid).

Figure 2. Regional distribution of TFP , NUTS3 1995-2018



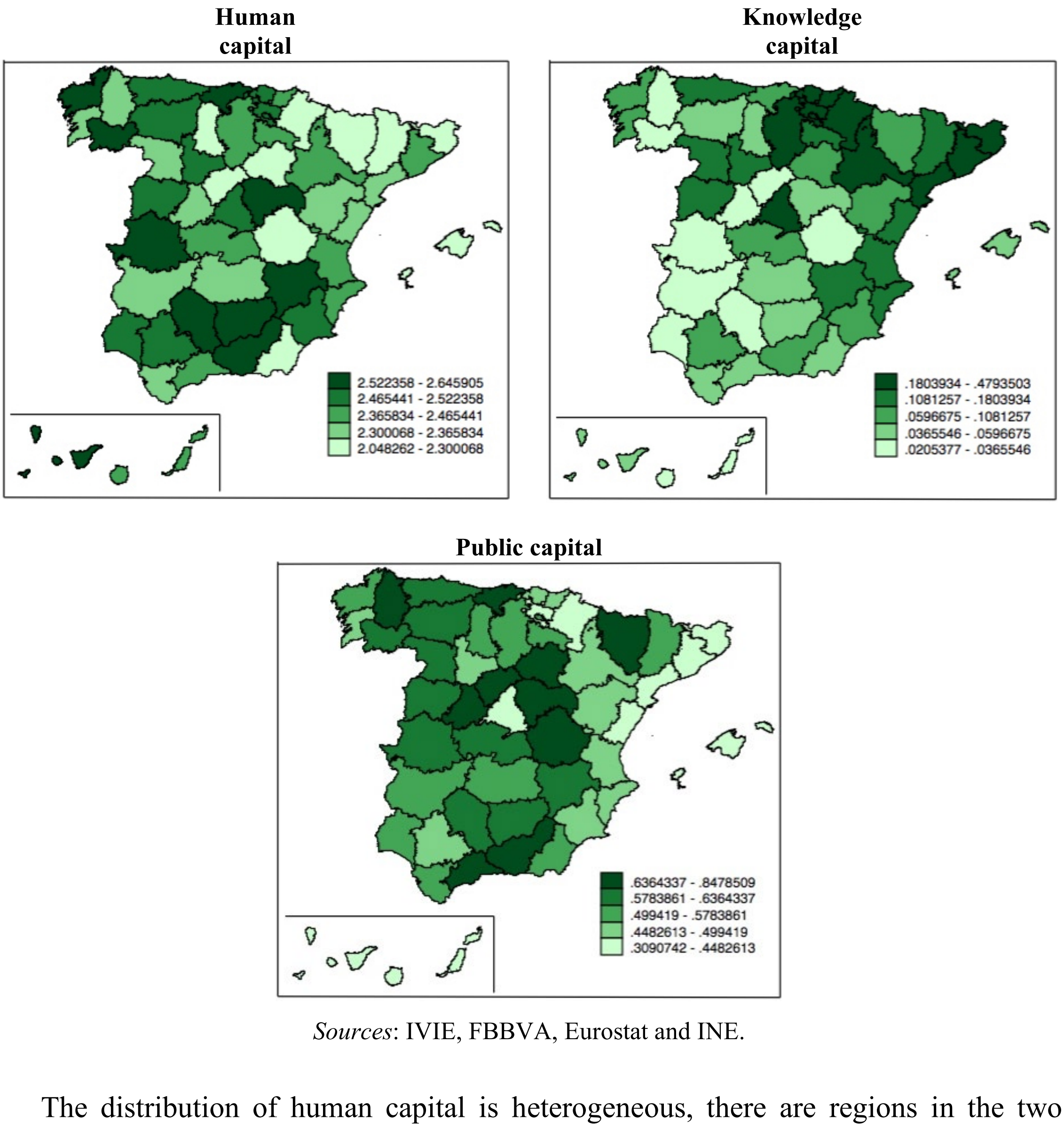
Source: Author's estimations

5.2 Geographical distribution and spatial analysis of the variables

In order to approach the question from a further level of detail, the analysis of TFP determinants is restricted to the smallest level of aggregation (NUTS3). Moreover, previous results indicate that spatial autocorrelation is clearly more significant at this level. The choice of variables is based on the literature reviewed in section 2. First, the source of the data on the total amount of human capital of the workforce for the period 1995-2018 is the Bancaja Foundation and the IVIE (2015). Then, human capital is expressed as the average per worker using the previously described labour series (section 4). The variable units are equivalent workers; the reference is a 20 years old employee without studies or only primary schooling. Besides, the number of patents produced in a year is used as a proxy for knowledge capital. The source of the number of patents per regions is Eurostat. Then, the stock of patents is constructed as the sum of the total patents in the previous five years and divided by the number of workers. Finally, data on the net stock of public capital in constant 2005 prices is found in the database of BBVA Foundation and IVIE (2015b). It is expressed as the ratio to the real stock of capital using data on productive capital (section 4). All the variables are converted to natural logs.

To investigate the presence of spatial dependence in the variables introduced as potential determinants of TFP, the first step is to represent the quintile distribution of the variables in the Spanish map (Figure 3).

Figure 3. Distribution of human capital, knowledge capital and public capital, 2018.



Sources: IVIE, FBBVA, Eurostat and INE.

The distribution of human capital is heterogeneous, there are regions in the two highest quintiles dispersed across the area composed by Galicia, Asturias, Cantabria, Basque Country, Castile and León, Madrid, Castile - La Mancha and Andalusia, among others. Surprisingly, regions in the East, located in the AACC of Catalonia (Lleida and Gerona), Balearic Islands and Navarre present values of human capital per worker below the average. On the other hand, knowledge capital is distributed similarly to value added (section 4). Thus, the highest values are found in the Northern regions, mainly in those situated within the area formed by the Basque Country and Navarre; the coastal Catalonian regions (Gerona, Barcelona and Tarragona) and; the Centre (Madrid). Contrarily, low stocks of this variable are located in the West and Centre, particularly in the regions within Galicia, Castile and León, Extremadura, Castile – La Mancha and Andalusia. However, the distribution of the ratio of public capital to total capital stock presents a somewhat different pattern. Some of the best performing regions (located in the AACC of Madrid, Catalonia and the Basque Country), in terms of value added or productivity, have low values of public capital relative to the total, probably because much of the private stock of capital concentrates in these areas. This is not the case for the majority of regions, where high values are homogeneously distributed, explaining thus the high presence of regions in the top quintile, located around regions with low values of this variables. This is particularly noticeable in the case of Madrid (Centre).

Formally testing for spatial dependence in the selected variables confirms the intuition extracted from the maps. The table 3 below reports the results from the Moran I test for knowledge capital, public capital and human capital. For the first two variables, the I statistics are 0.2940 and 0.1796, respectively. Both are significant at 1% level, indicating the presence of spatial autocorrelation. However, the test fails to reject the null in the case of human capital, although the I statistic (0.0415) is almost significant at the 10% level (p value is 0.1445). To summarize, the visual analysis of the maps and the Moran I test suggest that a spatial dependence pattern is clear in the case of knowledge capital and public capital, but is weak, if exists, for human capital.

Table 3. Moran I test of human capital, public capital and knowledge capital

	I	E(I)	SD(I)	z score	p-value
Human capital	0.0415	-0.02	0.058	1.0603	0.1445
Patents	0.2940	-0.02	0.058	5.5100	0.0000
Public capital	0.1796	-0.02	0.058	3.4421	0.0002

Note: The variables are in per capita terms and transformed to natural logs.
The weights matrix used is row-standardized.
Source: Author's estimations.

The departing point is the regression of TFP obtained in the previous sector, on two of the variables described above -human capital and knowledge capital- for the year 2018. Due to the fixed feature of the variable, the analysis of the determinants requires to change to a cross-section context. For this model, it is assumed that the spatial dependence is totally removed in the production function regression, so that there is no need for including spatial terms, which is limited by the small number of observations (50)[17]. For the same reason, public capital is excluded from the cross-section OLS regression. However, the problem with such model is that it does not enable to control for spatial effects, which exists accordingly to the Moran I test results. In order to be able to estimate Spatial Lag models, a panel data approach is adopted for the period 1995-2008. The dependent variable in the panel case is now the TFP index derived from equation 3.

The empirical procedure for the panel data sample is outlined here. First, a fixed effects regression is run using human capital, knowledge capital and public capital as covariates for the period 1995-2008. Then, a SDM is estimated including spatial lags of the three forms of capital. However, the spatial lags of the independent variables are not jointly significant. Therefore, the term of the spatially lagged human capital is removed[18]. The resulting model is described by equation 10:

$$tfp_{rt} = \alpha_r + \lambda W tfp_{rt} + \beta_1 hc_{rt} + \beta_2 kc_{rt} + \beta_3 pc_{rt} + \gamma_1 W kc_{rt} + \gamma_2 W pc_{rt} + u_{rt} \quad (10)$$

Where tfp_{rt} reflects the TFP index (equation 3); W is the weights matrix described before and; the term $Wtfp_{rt}$ is the spatially lagged dependent variable, included as proxy for the effects of productivity spillovers. The variable hc_{rt} is human capital per worker, while kc_{rt} is the knowledge capital stock and, pc_{rt} the ratio of public capital over the total stock of capital in each region. To form the spatial lags of both variables, these are pre-multiplied by the matrix W. All data is transformed to natural logs. Regional fixed effects and time dummies are included in all the models, assumptions supported by the results from the Hausman test and the joint significance test, respectively. Table 4 below reports the results from the analysis of TFP determinants.

[17] The sample is too small to implement Spatial Lag models to the analysis of the determinants of TFP. Similar works, within the cross-section context, account for almost 200 regions (EU). See Dettori et al., (2010).

[18] The spatial lags of knowledge capital and public capital are jointly significant at 1% level. The Moran I test in Table 4 indicates that human capital is the only variable that is not spatially correlated.

Table 4. Analysis of TFP determinants for the Spanish NUTS3 regions

Dependent variable	**Estimated TFP (2018)**	**TFP index (1995-2018)**				
Variables/Model	**(I) OLS**	**(II) FE**	**(III) SDM**	**(IV) SDM**	**(V) SDM**	**(VI) SDM**
Human capital	0.2894 (0.1985)	0.1213*** (0.0312)	0.1256*** (0.0298)	0.1231*** (0.0299)	0.0942*** (0.0309)	0.1321*** (0.0285)
Knowledge capital	0.3108 *** (0.1082)	0.1618*** (0.0474)	0.1029** (0.0497)	0.1099** (0.0495)	0.2009*** (0.0465)	0.0511 (0.0478)
Public capital		0.3735*** (0.0320)	0.3966*** (0.0312)	0.3973*** (0.0312)	0.4152*** (0.0324)	0.4267*** (0.0307)
Spatial terms						
TFP index			0.1516* (0.0868)	0.1445** (0.0709)	-0.2597* (0.1341)	0.0998* (0.0579)
Knowledge capital			0.3826*** (0.1162)	0.3307*** (0.1000)	0.1154 (0.1902)	0.4999*** (0.0723)
Public capital			-0.5088*** (0.0944)	-0.4369*** (0.0765)	0.5161*** (0.1510)	-0.3492*** (0.0546)
Distances (km)	-	-	*All*	*<400*	*>400*	*Binary*
Fixed effects						
Regional	*No*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Year dummies	*No*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Moran I test *p value*	0.336 (0.000)	0.183 (0.000)	0.008 (0.315)	0.004 (0.344)	0.008 (0.320)	-0.042 (0.351)
R^2	0.12	0.32	0.36	0.35	0.33	0.39
No. obs.	50	700	700	700	700	700

Note: All the variables are transformed to logs.
The parenthesis indicates standard errors.
The Moran I test is computed following the methodology described in section 4.
***(1%), **(5%), *(10%).
Source: Author's estimations.

In column I above, the dependent variable of the regression is the residual from the production function estimated in section 5 based on the SDM of equation 7. The coefficients in the OLS regression for human capital and knowledge capital are 0.2894 and 0.3108, respectively. Only the second is significant at 1%. The R^2 is only 0.12. The model is not able to explain most of the regional differences without including the spatial terms. Furthermore, the Moran I test indicates the residual presence of spatial dependence at 1% level of significance. The remaining columns correspond to the panel data regression. At the right hand, the TFP index derived from equation 3 is regressed on human capital, knowledge capital and public capital for a panel of Spanish regions during the period (1995-2018) without including spatial terms (II). All the covariates have an important impact and their coefficients are significant at 5% level. The major contribution is from public capital (0.3735). On the other hand, the coefficients of human capital and knowledge capital are 0.1213 and 0.1618, respectively. The I statistic obtained after applying the Moran I test on the residuals is 0.183, and the p value is 0.000. Thus, again there is presence of remaining spatial autocorrelation. Now, the R^2 is more than double that obtained in the first regression 0.32.

In the SDM (III), the coefficients of human capital, knowledge capital and public capital are significant at 1% level. The parameters estimated for human capital (0.1256) and public capital (0.3966) are very similar to those obtained in the last regression (II). Only the coefficient of knowledge capital (0.1029) is well below the one estimated in regression II. Regarding the spatial terms, the parameter of the spatial lag of the dependent variable is 0.1516, being significant at 10% level, which suggests that productivity spillovers have a certain role in explaining disparities across Spanish regions. The parameter estimated for the spatial lag of the knowledge capital is 0.3826 and is significant at 1% level. This result suggests that knowledge capital spillovers have a strong positive effect, accounting for almost four times the contribution from the variable within the region boundaries, explaining the reduced effect of the accumulated patents within the region.

In the case of public capital, the parameter obtained is -0.5088 and is also significant at 1% level. Its sign suggests that public capital spillovers have a negative effect on TFP, while the impact of the public capital within the region boundaries is positive as shown above. Such relationship is anticipated in the subsection 5.2., where the visual analysis of the map for public capital in Figure 3 reveals a uniform distribution of high values of public capital that contrasts with the relatively low values of this variable displayed by those among the richest in terms of income and productivity, that are geographically separated[19]. This result might be reflecting the stablished regional economic policy, which aims to reduce the gap by favouring those areas that lack of resources to develop. To conclude, spatial dependence is successfully controlled in the SDM according to Moran I test results. The I statistic of the residuals is 0.008 (p value is 0.315). Besides, the R^2 is the highest of the six regressions in table 6 (0.36).

5.3. Scope of productivity spillovers in terms of distance

In order to assess the scope of spillovers, zeroes can substitute the elements in the weights matrix except those elements included within a certain range according to

[19] Different proxies for public capital were tried, such as the variable in per worker/capita terms, achieving similar results.

different distances[20]. The elements of the weights matrix used in next regression (IV) take value 0 for distances larger than 400 km. On the other hand, the weights included in the regression of column V take value 0 for distances smaller than 400 km. Furthermore, a binary weights matrix is introduced in the last regression (VI), so that the pairs of regions that are strictly contiguous take the value 1, and 0 otherwise. This weight allows to control solely for cross-border spillovers. Although the choices of weights are purely arbitrary, the purpose is to explore the scope of spillovers by trying the different matrices described above and identifying important changes in the estimations.

In the three cases, the coefficients estimated for the (not spatially lagged) determinants are fairly similar to those obtained with the full distances. In what respect to the spatial terms, the parameters estimated by the SDM using only distances smaller than 400 km (IV) are only slightly below those in the previous estimation (III) and all of them are significant. However, the results obtained through the matrix built with distances larger than 400 km show important distortions. The most important is the change in the sign of productivity spillovers, which indicates that its impact is positive. The sign of public capital also changes. The spatial lag of knowledge capital, however, keeps the same sign but losses its significance. Moreover, the two variables that present spatial dependence – public capital and knowledge capital – now present higher coefficients for their values observed within the regions. All these facts imply that when the model does not reflect the spatial pattern of spillovers adequately, the contribution from the (not spatially lagged) determinants is overestimated.

Finally, the overall impact of cross-border spillovers is slightly behind the effect observed within the 400-km range, with only one exception, knowledge capital. Now the stock of patents produced within the region has no significant effect on the level of TFP. However, the coefficient for knowledge diffusion is now 0.4999, the highest of the six regressions, and it is significant at 1% level. In the last regression (VI), the R^2 is the maximum obtained in the analysis of TFP determinants (0.39). The model, however, omits spillovers arising from regions that are not contiguous, but have a significant effect in productivity levels. Besides, in the three regressions the Moran I test fails to reject the null of no spatial autocorrelation, which indicates that the spatial interactions between the variables are successfully controlled. To conclude, the results above suggest that spillovers have their major effects inside the 0-400 km range, particularly

[20] This approach is adopted from Dettori et al., (2012)

most of them take place across the border. The inclusion of distances outside this range is the only case in which extreme changes are observed in the spatial parameters.

6. Final remarks

The investigation above confirms that spatial interactions play a critical role in explaining disparities in per capita income and productivity across the Spanish regions. The main implication from this paper is that the previous empirical analyses of productivity that ignore the impact of geography are likely to overestimate the effect of the region intrinsic characteristics. Moreover, the high significance achieved by the coefficients of the spatially lagged variables suggests that the omission of these terms in the regression might result in inconsistent estimators for either the production function, or the effect of TFP determinants.

The estimates of the production function are subject to strong and significant spatial patterns, particularly at the smallest level of administrative aggregation. The estimated contribution of value added spillovers to regional output is greater than the combined effect of the inputs –capital and labour– accumulated within a region. Moreover, the effect of the stock of capital in neighbouring areas is nearly twice as large the contribution of capital located within the region boundaries. The main implication of this point is that the factors situated outside a certain region are responsible of more than a half of its production. Regarding the analysis of the determinants, the results suggest that productivity spillovers have a certain role in explaining inequalities across the Spanish regions. In addition, the diffusion of knowledge is revealed as the main determinant. To conclude, the addition of spatial interactions to the empirical analysis of productivity is crucial to understand the existing differences across the Spanish territory.

This paper is the first approach to the analysis of regional productivity in Spain using Spatial Lag models. However, there are still some important topics that need for further research. The main challenge is to carry out a similar analysis but disaggregated across sectors. It would provide a valuable insight about the possible changes in the behaviour of spatial interactions in relation to the different economic activities, which is still totally unexplored. Furthermore, the present study takes into account three important determinants of TFP, but much of the regional differences in TFP are still unexplained. The methods presented in this master thesis allow the researcher to estimate the impact of different factors such as international trade, R&D expenditure or social capital. Lastly, one important question that is left unresolved is whether the current spatial patterns are affected by the economic crisis. In such case, the sample would necessarily cover the latest years resulting in the reduction of the number of observations due to the length of the series available.