Preliminary version – not for citation

The Review of Income and Wealth

https://onlinelibrary.wiley.com/doi/10.1111/roiw.12422

A Stochastic Model with Penalized Coefficients for Spatial Price Comparisons: An Application to Regional Price Indexes in Italy

José-María Montero

University of Castilla-La Mancha

Tiziana Laureti*

University of Tuscia

Román Mínguez

University of Castilla-La Mancha

and

Gema Fernández-Avilés

University of Castilla-La Mancha

**Correspondence to: Tiziana Laureti, Department of Economics, Engineering, Society and Business Organization, University of Tuscia, Via del Paradiso 47, Viterbo 01100, Italy (laureti@unitus.it).*

Abstract

This paper focuses on a new strand of research that uses stochastic approach for making spatial price comparisons. We propose a novel method to account for the presence of spatial dependencies in consumer prices and consequently in price indexes by imposing penalization

conditions on the estimation of traditional CPD models leading to the spatially-penalized country-product-dummy (SP-CPD) model, The paper proposes an appropriate estimation strategy, which enables us to simultaneously estimate all the parameters in the model, including the smoothing parameter of the penalization term instead of determining it externally. In order to estimate spatial price indexes for areas lacking in price data, we suggest applying the kriging methodology to the price indexes obtained from the SP-CPD model. This new approach is applied to official Italian CPI data for constructing regional spatial price indexes for 2014. The results show that price levels are higher in the Northern-Central regions than in the South.

Keywords: Hedonic country product dummy models; Spatial dependence; Spatial price indexes; Stochastic approach; Regional price comparisons.

JEL Classification: C21, D12, E31

1. **Introduction**

Spatial price indexes provide measures of price level differences across countries or across regions within a country and are widely used by researchers and policy-makers for comparing real income, standards of living and consumer expenditure patterns.

In all spatial price comparisons, the concept of Purchasing Power Parity (PPP) is used to measure the price level in one location compared to that in another location[1]; therefore PPPs are essentially spatial price index numbers. At international level, PPPs facilitate cross-country comparisons of Gross Domestic Product (GDP) and its major aggregates as they can be used in converting aggregates into a common currency. Likewise, sub-national PPPs allow for intra-country spatial comparisons and can serve as inputs and/or improve other inputs for estimating key economic indicators produced by countries, such as real regional price comparisons, real income dimensions and poverty estimates. The process of compiling PPPs is quite complex and is carried out in two stages[2]. First, elementary spatial price indexes are computed by aggregating, without using weights, prices of items belonging to a group of similar well-defined goods or services (called Basic Headings, BHs). In the second stage, the elementary PPPs are aggregated using expenditure weights to obtain PPPs for higher-level aggregates such as consumption, investment and GDP.

In order to improve the quality and reliability of PPP estimates, this paper focuses on methodological issues that arise when constructing spatial price indexes at the lowest level of aggregation since it is essential to obtain reliable PPPs at BH level because they are the foundations of overall comparisons (Hill and Syed, 2015). One of the main issues when constructing spatial price indexes is to capture the spatial dependence which is inherent in consumer price levels (Aten, 1996; Aten, 1997; Rao, 2004). Several researchers have found that consumer prices are more similar in geographically proximate locations, thus observing a significant positive correlation between the Law of One Price (LOP) deviations and distance

[1] Purchasing power parities of currencies are defined as the number of currency units of a country that can purchase the same basket of goods and services that can be purchased with one unit of currency of a reference currency.

[2] At international level, PPPs are compiled by the International Comparison Program (ICP), which is administered by the World Bank and overseen by the United Nations Statistical Commission with the collaboration of the OECD, EUROSTAT and other regional organizations (see Rao, 2013 for details).

(Choi and Choi, 2014, 2016; Crucini et al., 2015). This spatial effect may reflect transport costs as well as local distribution costs, which are likely to be similar in nearby locations if the distribution of goods is labour intensive and labour markets are geographically integrated (Choi and Choi, 2016). However, in spite of its theoretical attraction, as yet only a few studies have been carried out to explore the issue of spatial dependence in consumer price index construction (Aten, 1996, 1997; Rao, 2001; Biggeri et al., 2017; Majumder et al, 2017).

In order to compare consumer price levels, this paper focuses on the stochastic approach, where uncertainty and statistical ideas play central roles since index number construction is viewed as a problem of signal extraction from the messages on price changes for different commodities over space (Summers, 1973). Clements and Izan (1987), Selvanathan (1989) and Selvanathan and Rao (1994) have emphasized the versatility and usefulness of the stochastic approach which leads to familiar index-number formulae under certain circumstances (Clements et al., 2006, Diewert, 2010). Over the last two decades there has been a steady increase in research focused on the stochastic approach, which is based on the hedonic approach to price index number construction and the model proposed by Summers (1973), namely the *country product dummy* (CPD) model. This literature is still expanding and in a recent paper, Rao and Hajargasht (2016) developed CPD-based stochastic approach to international price comparisons by incorporating modern econometric tools.

With the aim of improving the CPD methodology, we propose the spatially penalized country product dummy (SP-CPD) model in order to incorporate the impact of spatial dependence on the value of spatial price indexes. In this model, spatial dependencies are introduced by penalizing the differences in the spatial price indexes of neighbouring areas (countries or regions within a country). Therefore, we focus on smoothing the spatial PPP pattern (by estimating the smoothing degree from the data) rather than on introducing spatial autocorrelated error terms on the traditional CPD specification or discovering the form of the

spatial interaction (by including spatial lags in the response and/or the covariates). It is important to note that by penalizing differences in the coefficients for neighbouring areas it is possible to consider the spatial dependencies present in the data. Moreover, this is a recommended procedure when there are few price data since it reduces the variance of the estimates. However, penalization is not equivalent to any specific type of spatial correlation structure. Please see subsection 2.2 for more details.

When using the SP-CPD model, the degree of penalization is determined by the data through the model estimation and more specifically by estimating the value of the smoothing parameter used to tune the penalizations imposed on the coefficients associated with geographical areas.

Another contribution of the paper is that it identifies an appropriate estimation strategy, based on the transformation of the SP-CPD model into a mixed model. In this way, it is possible to use maximum likelihood methods which enables us to simultaneously estimate all the parameters in the model including the smoothing parameter instead of determining it externally, as is standard practice. Moreover, in order to overcome the lack of price data at every location included in the study, we combine the SP-CPD model with a kriging strategy to estimate the spatial price indexes for the geographical areas without price data. In this way, the estimates take into account the spatial dependencies existing in the spatial price indexes provided by the SP-CPD model from a geostatistical perspective. Therefore, the SP-CPD estimates are considered to be the realization of the data generating process of "regional" price indexes and kriging is used for estimating "regional" price indexes in non-observed geographical areas (regions). Our improvements to the CPD model provide a comprehensive framework for carrying out inter- and intra-national price comparisons using data traditionally collected by National Statistical Offices (NSOs) for computing official Consumer Price Indexes (CPIs) as well as new sources of data like scanner data with detailed point-of-sale information.

With the aim of illustrating the potential of the proposed methodology and highlighting the informative results, we estimated the SP-CPD model using real data obtained from the official CPI survey carried out in Italy. Using these data, we estimated regional spatial price indexes for Italy in 2014. We referred to 7 groups of products belonging to the most important CPI product group, namely Food and non-alcoholic beverages. Kriging was used to predict PPP for Campobasso, which is the only area without price data.

The rest of the paper is structured as follows. Section 2 reviews the traditional CPD models and presents the SP-CPD model for estimating spatial consumer price indexes. Section 3 describes the CPI data used in our empirical analyses for constructing sub-national PPPs and reports the estimation results. Finally, some conclusions are drawn in Section 4.

2. Methodology

2.1 The CPD-based stochastic approach for constructing spatial price indexes

The CPD model is currently considered to be the principal aggregation method under the stochastic approach for index number construction (Rao and Hajargasht, 2010). It is widely used in the ICP at the World Bank[3] due to its ability to deal with data issues arising from variations in the quality of items across areas and from gaps in available price data for making spatial and temporal comparisons (Kokoski et al., 1999; Aten, 2006; Dikhanov et al., 2011; de Haan, and Krsinich, 2014; Biggeri et al., 2017).

In this paper, we consider the problem of making spatial comparisons of prices between R areas (regions) at elementary level, where no expenditure weights are available. However, the SP-CPD model can easily be extended to include weights for making spatial price comparisons.

[3] See World Bank (2013) for a more complete description.

According to Rao and Hajargasht (2016), the model in its multiplicative form postulates that the observed annual price of the n-th commodity in outlet k in r-th area, p_{nkr}, (n = 1, 2,…N; r = 1, 2,…, R; k = 1,…, K_{nr}) can be expressed as the product of three components: the PPP or the general price level in area r relative to reference base area BA (denoted by PPP_r^{BA}), the price level of the n-th commodity in outlet k relative to a base commodity BC (denoted by P_{nk}^{BC}) and a random disturbance term u_{nkr}.

$$p_{nkr} = PPP_r^{BA} \cdot P_{nk}^{BC} \cdot u_{nkr}, \text{ with } PPP_{BA}^{BA} = 1, \tag{1}$$

The additive form of the CPD model is obtained by taking logarithms of both sides of (1):

$$\begin{aligned} \ln p_{nkr} &= \ln PPP_r^{BA} + \ln P_{nk}^{BC} + \ln u_{nkr} \\ \ln p_{nkr} &= a_r + b_{nk} + \nu_{nkr} \end{aligned} \quad . \tag{2}$$

Model (2) can be expressed as a regression equation for each price observation corresponding to product (or commodity) n in area r in outlet k where the independent variables are dummy variables. Therefore:

$$\ln p_{nkr} = \sum_{r=1}^{R} a_r D_r + \sum_{n=1}^{N} b_n D_n^* + \nu_{nkr}, \tag{3}$$

where $D_r = 1$ for area r and 0 otherwise; $D_n^* = 1$ for product n and 0 otherwise. Obviously, restriction $a_{BA} = 1$ is imposed on a's in order to solve the normal equations, so that a_r is the difference of (fixed) effects associated with the areas with respect to the base area BA. Then, the PPP of area r with respect to a base area BA is given by $PPP_r = e^{\hat{a}_r}$.[4]

The CPD model is can be extended to include J quality characteristics of the products, Z_1, Z_2,…, Z_J, including information on outlet type and product brand *(j=1,…,J)*. Then, the hedonic CPD model is specified as:

[4] It should be noted that the CPD model assumes that the areal effect is constant for every product in the same group. In other words, interactions between these factors of the model are not considered in this specification.

$$\ln p_{nkrj} = \sum_{r=1}^{R} a_r D_r + \sum_{n=1}^{N} b_n D_n^* + \sum_{j=1}^{J} c_j Z_j + \nu_{nkrj}, \quad (4)$$

where, if the appropriate restriction is made on *c*'s, c_j can be interpreted as the difference of (fixed) effects associated with quality characteristic *j* with respect to a specific reference. Several authors have demonstrated the flexibility of this regression-based econometric methodology for constructing binary and multilateral price index numbers, since it accounts for the quality variations in cross-area price data (see Kokoski et al. 1999; Diewert, 2005; Hajargasht and Rao, 2010) and provides standard errors for the estimated parameter values and consequently for PPPs.

2.2 Spatial dependence in consumer prices and spatial penalization

This paper extends the CPD methodology for computing PPPs by taking into account the spatial effects underlying consumer price differences among geographical areas within a country. Therefore, it acknowledges the First Law of Geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970).

The literature on international and sub-national PPPs is generally based on the assumption that there is no interdependence among price movements in the various geographical areas included in the comparison. Nevertheless, empirical evidence of spatial correlation has been observed at cross-country level by Aten (1996;1997) and Rao (2001) and at sub-national level by Biggeri et al. (2017). Aten (1996) tested for spatial autocorrelation among country price relatives and found that all BHs were significantly and positively autocorrelated by at least one of the weight matrices used. Subsequently, Aten (1997) found that prices tend to be more similar in countries that are geographically close and that the spatial component provides useful insights for understanding the differences between the price relatives of tradable and

non-tradable goods[5]. Rao (2001) demonstrated the presence of spatial autocorrelation using the 1985 global comparison results from the ICP for 56 countries with eight aggregated expenditure categories. Successively, focusing on methodological issues, Rao (2004) drew attention to the fact that assuming identically and independently distributed disturbances over all countries and products is probably too restrictive and emphasized that adjustments are required for the estimates. Biggeri et al. (2017) were the first authors to explore this issue at sub-national level by estimating a CPD model using a spatial first-order autoregressive process for the error terms. With the aim of analysing the sensitivity of PPP estimates to alternative estimation procedures at international level, Majumder et al (2017;2018) recently estimated a spatial CPD model by assuming spatially correlated errors and using two neighbourhood criterions for defining alternative spatial weight matrices.

By reviewing the previously mentioned studies which involve the estimation of CPD models with a spatial structure in the errors or in the response, it appears that considering spatially correlated prices in the CPD framework may affect numerical values of PPP estimates.

Nevertheless, from a methodological point of view, if adequate adjustments are not made, biased estimates of standard errors will occur in the traditional ordinary least squares CPD model when there is spatial autocorrelation among prices. Consequently, biased t-tests and misleading indications of precision of the resulting PPPs will be obtained.

More specifically, the estimates provided by the CPD methodology are based on differences in arithmetic means of log prices, thus implicitly assuming that prices of goods and services are independent, which may not be true. Contrastingly, it is reasonable to assume that product prices are spatially autocorrelated and also exhibit spatial heterogeneity (the so-called spatial effects) especially when comparing consumer prices across areas within a country. When

[5] By estimating a spatial lag model it was assumed that price parities influence the parities of neighbouring countries or countries with strong trading relationships.

spatial dependence is present, the difference between the arithmetic mean of the log prices of products sold in region *R* and the mean of the log prices in the base region *BA* is no longer the optimal estimator of the PPP of region *R*. It is an unbiased estimator but not the estimator with the minimum variance. Therefore, as stated in Montero et al. (2015), using the arithmetic mean in the presence of spatial dependence has detrimental consequences[6].

Various approaches may be adopted to incorporate the spatial dependence inherent in consumer price levels in order to obtain efficient estimators of spatial price indexes.

As suggested by the above-mentioned authors and outlined in the introduction, from a spatial econometric perspective, a possible way to include spatial effects in a CPD model is to make adjustments that incorporate spatial autocorrelation in the error term[7], which gives rise to the SEM-CPD model. Products and services in the same area will share unobserved neighbourhood effects which will consequently lead to spatially correlated disturbances. However, one of the disadvantages of using this approach is that it is essential to specify a spatial autoregressive structure for the disturbances. The most commonly-used specification assumes a spatial first order autoregressive process for the error terms (see online Appendix A for details on the specification of the spatial error model and SEM-CPD models).

According to the geostatistical approach, the unbiased estimator with minimum variance is the difference in "kriged" means, which is a weighted mean of log prices. The specification of weights depend on the structure of the spatial dependence inherent in the prices, which is estimated with a covariance or semivariogram function (see Cressie, 1993, for details).

[6] For example, true confidence intervals have lower confidence levels (or are wider) than those obtained using arithmetic means; and the true power of the tests is lower than that obtained assuming independent prices, thus resulting in undesired rejections of the null hypothesis. Some examples can be found in Schabenberger and Gotway (2005, pp. 32–34), Cressie (2015, pp. 15–17) and Montero et al. (2015, pp.5-6).

[7] The inclusion of spatial autocorrelation in the response or/and in the explanatory variables are other possibilities.

When a large volume of spatial data is available, the so-called "big *N* case", the precision of the estimates as well as the reliability of the statistical hypotheses testing tend to improve and correcting for autocorrelation is not required. Indeed, when there is an adequate number of observed prices for each region, both the arithmetic and the kriged mean coincide with the population mean and the traditional CPD model can be used for PPP estimations. On the other hand, when the ratio between the observations and parameters in the CPD model is low, penalization provides more accurate estimates.

In the light of these considerations, a traditional CPD model should not be used for PPP estimation when spatial effects (especially spatial dependence) are present, unless a huge database is available, which is seldom the case when making spatial price comparisons. Alternatively, kriged means can be used. However, in order to estimate both the structure of the spatial correlation and the kriged means, it is essential to know the geographical location of the outlets where the product prices were collected.

The unavailability of information regarding outlet location combined with the fact that there may also be spatial dependence in the prices observed at the border areas of neighbouring regions,[8] support our idea of including spatial dependencies in the CPD model by penalizing the differences in neighbouring geographical coefficients. This penalization smooths the estimated PPPs for neighboring regions, thus capturing spatial dependence in the prices observed at the border areas of neighbouring regions. From a statistical viewpoint this results in a trade-off between the fit of the model and the roughness of the PPP variation in neighboring regions. This manipulation of the CPD model results in the SP-CPD model. As is

[8] Regions are administrative areas. Therefore, consumers move freely from one region to another. Imagine that the eastern part of region A borders on the western part of region B. Prices will be similar in these two parts; otherwise, consumers would move from one region to the other for shopping, which over the medium-term would result in similar prices.

the case with penalized models, the estimated variance is reduced even at the risk of introducing some bias with respect to traditional CPD estimates.

Besides methodological motivation, the penalization of differences in PPPs of neighbouring areas has a clear economic rationale: although regional economic theory states that the LOP does not hold across regions and that geographic price dispersion is high and persistent even within a country where trade barriers are relatively low (Crucini et al., 2015; Engel and Rogers, 2001), several researchers have observed a significant positive correlation between LOP deviations and distance (Anderson and van Wincoop, 2004; Choi and Choi, 2014). Therefore, distance can be considered as a metric for market friction, therefore consumer price difference is greater between cities or geographical areas located farther apart. As a consequence, spatial price indexes may show similar patterns for neighboring areas.

The SP-CPD model does not impose a spatial structure in the errors or in the response but allows the data to indicate how similar the PPPs are by estimating a smoothing parameter weighting the penalization term, as illustrated in the next section.

2.3 The spatially penalized CPD model

The SP-CPD approach is similar in some respects to the CPD methodology. First, it is essentially an implementation of the hedonic approach which accounts for quality variations in price data. Second, the SP-CPD approach is based on a stochastic formulation for constructing multilateral price index numbers, which is particularly advantageous as it enables us to use a range of econometric tools and techniques.

In order to introduce the SP-CPD strategy, the penalization

$$\sum_{r=2}^{R} \sum_{s \in N(r),\, s<r} \left(a_r - a_s\right)^2 \tag{5}$$

is included when estimating model (4). In the penalization term, $N(r)$ represents the number of neighbours of area *r*, and the squared differences of the coefficients a_r for all the possible combinations of neighbouring regions represent the penalty used for smoothing the spatial effects, that is for preventing drastic differences in the coefficients of neighbouring areas. It is important to note that, since the penalization itself is a restriction imposed on model (4), the traditional ANOVA restrictions are not required (for example, the coefficient for a specific area must be zero). In this way, the PPP of area *r* with respect to the reference or base area *BA* must be computed as $PPP_r = e^{a_r - a_{BA}}$. However, the SP-CPD model distinguishes itself from the hedonic CPD methodology in two important ways. First, because it includes a penalty for the differences in the PPPs of neighbouring areas and secondly due to the method used for estimating the model. More specifically, instead of using a least squares approach the penalized least squares (PLS) criterion specified as follows:

$$\min PLS(\lambda) = \sum_{nrj=1}^{NRJ} \left(\ln p_{nkr} - \ln p_{nkr} \right)^2 + \lambda \sum_{r=2}^{R} \sum_{s \in N(r),\, s<r} (a_r - a_s)^2 ,$$

It is essential to know the value of the smoothing parameter λ. The SP-CPD model is transformed into a mixed model and then the restricted maximum likelihood (REML) criterion is used. The maximum likelihood (ML) estimation is not used in this case as it does not take into account the degrees-of-freedom used when estimating the fixed effects, thus resulting in biased estimates. However, REML estimation explicitly accounts for this loss of degrees-of-freedom[9].

This reparameterization does not require prior knowledge of the smoothing parameter, nor does it need to be externally determined using procedures based on the optimization of a cross-

[9] In this respect, it is worth mentioning that in the logarithmic form of the SP-CPD model the reduction in the number of degrees-of-freedom occurs via the constraint imposed on the penalization matrix instead of the usual linear constraints used in traditional ANOVA.

validation method or an information criterion. In contrast, the smoothing parameter is estimated together with the other parameters in the SP-CPD model. This data-driven method for setting the smoothing parameter used to tune the penalty term in the SP-CPD strategy has great statistical advantage compared to traditional penalized regressions or ANOVA methods in the literature, which is what makes it a desirable procedure.

2.4 SP-CPD estimation method

In order to estimate the SP-CPD, firstly model (4) is rewritten using matrix notations:

$$\ln \mathbf{p} = \mathbf{Ma} + \mathbf{D}^{*}\mathbf{b} + \mathbf{Zc} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right),$$

Where $\mathbf{M}$ is a geographic design matrix specified as follows (see Fahrmeir et al., 2013):

$$\mathbf{M}\left(i,r\right) = \begin{cases} 1 & \text{if the observation } i \text{ corresponds to area } r \\ 0 & \text{otherwise} \end{cases},$$

$\mathbf{D}^{*}$ is a matrix dummy variables for different products and the matrix $\mathbf{Z}$ includes a set of quality characteristics; $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$ are vectors of coefficients associated with these matrices.

All the non-penalized parameters are collected in vector $\begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix}$, which corresponds to the extended matrix $\left(\mathbf{D}^{*} \quad \mathbf{Z}\right)$. As a result, model (4) is expressed as:

$$\ln \mathbf{p} = \mathbf{Ma} + \left(\mathbf{D}^{*} \quad \mathbf{Z}\right)\begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right),$$

Then, the penalization term (including the smoothing parameter) is expressed in matrix notations:

$$\lambda \sum_{r=2}^{R} \sum_{s \in N(r),\, s<r} \left(a_r - a_s\right)^2 = \lambda \mathbf{a}' \boldsymbol{\Omega} \mathbf{a}\,, \tag{6}$$

$$\text{with } \boldsymbol{\Omega}(r,s)=\begin{cases}-1 & r\neq s,\quad r \text{ and } s \text{ are neighbours}\\ 0 & r\neq s,\quad r \text{ and } s \text{ are not neighbours}\\ \left|N(r)\right| & r=s\end{cases},$$

with $\boldsymbol{\Omega}$ being the penalty matrix (whose row sums are zero). Unlike empirical applications with continuous covariates where $\boldsymbol{\Omega}$ results from the second-order difference matrix, in the case of discrete covariates we use a neighbour-based matrix. Consequently, differences between the coefficients for areas close to one another are penalized more than those corresponding to areas that are farther apart. It is important to note that the greater the difference between the parameters corresponding to neighbouring areas, the greater the increase in the PLS criterion. However, this penalization does not strictly mean that close areas must have similar PPPs. We leave the data to inform us about whether or not PPPs of close areas are similar by estimating the smoothing parameter together with the other parameters in the SP-CPD model.

For purposes of estimation, the SP-CPD model is transformed into a mixed model. In fact, a proper reparameterization of the vector of geographical coefficients **a** (these are the only coefficients that are subject to penalization) transforms models (4-5) into a mixed model in which λ is included in the covariance matrix of the random effects, so that it can be estimated jointly with the remaining parameters. It is worth noting that, according to the definition of $\boldsymbol{\Omega}(r,s)$, vector **a** follows a Markov random field (more specifically a Gauss-Markov random field, since normality is assumed), because the conditional distribution of any coefficient included, given all the others, only depends on its neighbours (see Rue and Held, 2005, for details). The connection between models with penalized coefficients and mixed models is described in the online supporting information (Appendix B). Moreover, since the above-mentioned reparameterization and the representation of model (4) as a mixed model involves a lot of technicalities, we detail the steps followed in Appendix C available online.

2.5. Kriging estimation of PPPs for areas without price data

The most promising approach to intra-country spatial price comparisons is to use CPI data, which are usually collected by NSOs in the main cities across the country. However, there may be several geographical areas of interest that are not included in such surveys, thus resulting in information loss. In this paper, kriging is used to predict spatial price indexes for areas without price data (in our case, only Campobasso).

When the variable of interest (in our case PPP) exhibits a spatial pattern (see online Appendix G where Tables G1 to G5 and Figure G2 show that PPPs estimated with the SP-CPD model exhibit a clear spatial pattern) we should determine the structure of the spatial pattern (spatial dependence) and use it to predict PPPs for unobserved locations (in our case the Molise region, represented by Campobasso). Geostatistics provides a spatial interpolation technique, kriging, which predicts the value of the PPP at an unobserved location using a weighted average of the PPPs of the regions located in a neighborhood determined by a semivariogram. This is the instrument used by geostatisticians for determining both the structure of the spatial correlation and its range. In this weighed mean, the weights of the PPPs of the observed regions usually decrease with an increase in distance of these regions from those unobserved. Roughly speaking, kriging is the spatial version of autoregressive models in time series analyses.

The fact that kriging accounts for the structure of spatial dependence represents a major advantage over other spatial interpolation techniques (inverse distance method, splines and polynomial regression, among others). Another important advantage is that kriging makes it possible to determine the accuracy of the predictions using the prediction error variance (the kriging variance) and can yield a plot of the standard deviation of the prediction errors.

In formal terms, let $X(\mathbf{s}_1), X(\mathbf{s}_2), \ldots, X(\mathbf{s}_R)$ be the PPPs of the R areas resulting from the SP-CPD model. Let them be geographically represented by their centroids, $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_R$, of the

areas where $\mathbf{s}_i$ is a pair of coordinates (longitude and latitude). Subsequently, following the geostatistical paradigm it is possible to predict the PPP of an area lacking in price data, whose centroid is represented by $\mathbf{s_0}$. In the geostatistical case, the following predictor is used:

$$\hat{X}(\mathbf{s}_0) = \sum_{i=1}^{R} \omega_i X(\mathbf{s}_i),$$

where the weights, ω_i, are obtained so that the predictor is optimal in the sense of unbiasedness and minimum variance of the prediction error. In the context of spatial autocorrelation (usually positive autocorrelation), the closer an area is to the prediction area the higher the influence of its PPP on the predicted PPP for the area without data.

In the event that the stochastic process governing the variable under study is non-stationary, as it is the case for PPPs of the Italian areas considered in this study[10], the process is assumed to have a drift rather than a constant mean, and the vector of weights used to obtain PPP of the prediction area is provided by Universal Kriging (UK) equations instead of the traditional Ordinary Kriging (OK) equations (see Montero et al., 2015, for details):

$$\begin{cases} \sum_{j=1}^{n(\mathbf{s}_0)} \omega_j \gamma_e(\mathbf{s}_i - \mathbf{s}_j) + \sum_{h=1}^{p} \upsilon_h f_h(\mathbf{s}_i) = \gamma_e(\mathbf{s}_i - \mathbf{s}_0), \quad \forall i = 1,...,n(\mathbf{s}_0) \\ \sum_{i=1}^{n(\mathbf{s}_0)} \omega_i f_h(\mathbf{s}_i) = f_h(\mathbf{s}_0), \quad \forall h = 1,...,p \end{cases},$$

where $n(\mathbf{s}_0)$ indicates the number of neighbouring areas (to the prediction area) entering in the prediction process. Since the stochastic process governing the Italian PPPs is not stationary, all of the areas are not necessarily included in this process. $\sum_{h=1}^{p} \upsilon_h f_h(\mathbf{s}_i)$ represents the local expression of the drift in the surroundings of the area whose centroid is $\mathbf{s}_i$, $\{f_h(\mathbf{s}),\ h = 1,...,p\}$

[10] The mean of consumer prices significantly differs among Italian macro-areas (North, Centre and South).

are p linearly independent known functions (more specifically, they are monomials of the coordinates), υ_h are constant coefficients obtained with moving neighbourhoods that can differ from one neighbourhood to another, and p is the number of terms employed in the drift. γ_e represents the semivariogram of the residuals obtained by subtracting the estimated value of the drift (which is not explicitly estimated) from the observed values. Finally, since the residuals are assumed to be stationary, $\mathbf{s}_i - \mathbf{s}_j$ is the distance between two areas in the neighbourhood of the prediction area, and $\mathbf{s}_i - \mathbf{s}_0$ is the distance between the prediction area and an area in its neighbourhood.

Since γ_e is generally unknown, we have addressed this problem by assuming that $\gamma_X \approx \gamma_e$, which is a reasonable assumption when the moving neighbourhoods considered in the prediction process are small, because in this case the drift can only undergo small changes.

The Universal Kriging prediction variance is given by:

$$V\left[\hat{X}(\mathbf{s}_0) - X(\mathbf{s}_0)\right] = \sum_{i=1}^{n(s_0)} \omega_i \gamma_e(\mathbf{s}_i - \mathbf{s}_0) + \sum_{h=1}^{p} \upsilon_h f_h(\mathbf{s}_0)\,.$$

3 An application to Regional Spatial Price Indexes in Italy

3.1. The importance of making inter-area price level comparisons

In countries characterized by large territorial differences in consumer preferences as well as in the quality of products and household characteristics, the calculation of sub-national PPPs acquires considerable importance. Evidence of sub-national spatial differences in consumer price levels has been found in large countries, such as Brazil (Aten, 1999), India (Deaton, 2003; Deaton and Dupriez, 2011; Coondoo et al., 2004; Majumder et al., 2015), and the United States (Koo et al., 2000; Aten, 2006), as well as in smaller countries like the United Kingdom (Wingfield et al., 2005), Germany (Roos, 2006) and Italy (Biggeri et al., 2008; Biggeri et al., 2017). Accurate measurements of price level differences are essential for assessing inequality

in the distribution of real incomes and consumption expenditures. Local or regional values of economic indicators (i.e. poverty indicators) should be adjusted for regional price differentials in order to avoid misleading regional analyses and the consequent policy implications and outcomes.

When constructing intra-country (inter-area) spatial price indexes, it is important to take spatial autocorrelation among prices into account than in the case of international comparisons. In an integrated market, cross-sub-national spillovers make the main provincial or regional economic pillars (economic growth, consumer prices, unemployment rate, population growth, etc.) strongly interdependent. This fosters market integration and promotes economic growth, which in turn expands the potential market and stimulates the mobility of production factors and the process of innovation diffusion, giving rise to new cross-sub-national spillovers (Özyurt and Dees, 2015).

Therefore, in this paper, we use the SP-CPD model to estimate Italian sub-national PPPs at regional level.

3.2. Data

In our empirical application, we use data collected for the purpose of computing Italian CPIs in 2014, which refer to capitals of 19 Italian regions. We selected seven groups of products (Table 1) belonging to the Food and non-alcoholic beverage CPI group, which accounts for 16.5 % of household final monetary consumption expenditure.

Table 1. List of groups of products (BHs), number of products and monthly price quotes

BH	Description	Num. of products in CPI survey	Num. of price quotes
1	Beef and Veal	4	16,884
2	Other meats and edible offal	2	5,520
3	Pork	2	8,424
4	Lamb, mutton and goat	1	3,552
5	Fresh or chilled fruit	73	64,655

6	Fresh or chilled vegetables other than potatoes	90	63,917
7	Fresh, chilled or frozen fish and seafood	29	55,276
	Total	201	218,228

More specifically, we considered fresh meat, all fresh fish species, all types of fresh fruit and vegetables, which make up approximately 30.3 % of the Food and non-alcoholic beverage group and 5.2 % of the entire consumption basket. We chose these products as they are comparable by definition and do not require further specifications in addition to those already present in the basket. However, the ‘‘Meat’’ group of products (Beef and veal; Pork; Lamb, mutton and goat; and Other meats and edible offal) includes various varieties that cannot be considered in the SP-CPD estimation process because they are not coded *a priori* and the data collectors usually select specific elementary items and specify the variety. Therefore, in this case we have a ‘‘loose’’ product description and weaker comparability. By choosing these groups of products the performance of SP-CPD models can be evaluated.

There is a large degree of product overlap among the 19 regional capitals considered in the 2014 CPI survey, even though the varieties available in different markets may vary reflecting distinct consumption patterns of each of the regions. Therefore, the total number of monthly price quotations in the dataset was 218,228. Starting from this detailed information, we constructed annual average prices for various products included in the CPI survey by considering the specific kind of outlet from which the prices are collected. As there are multiple quotes for all of the observations and considering the loose specification for ‘‘Meat’’ products, the annual data set contains approximately 5,000 unique individual price observations, each identified by outlet type (traditional, modern, hard discount and other), item code and geographic area. As already mentioned, the main limitation of the Italian CPI survey is the sampling design, which is limited to the regional capitals. Moreover, some of these capitals may be excluded from the CPI survey due to the quality of the price data collected. Figure E1 in online Appendix E shows the Italian regional capitals considered in the CPI

computations for 2014. Reggio Calabria, which is the regional capital of Calabria, has been replaced by Catanzaro, while Campobasso, the regional capital of Molise, was not included in the survey due to organizational issues. Since the stochastic process governing consumer prices is non-stationary, UK method was used in order to estimate the PPP for Campobasso-Molise. Online Appendix E also illustrates the socio-economic characteristics of the Italian regions and previous findings on consumer price differences across regions, while the accuracy of the kriging estimates is discussed in online Appendix F in light of the cross-validation results obtained.

3.3. Empirical results

Table 2 and Figure 1 show the estimated PPPs for 20 Italian regional capitals in 2014 (with reference to Rome=100) for seven BHs. Estimates are derived using the SP-CPD model except for Campobasso for which PPPs have been estimated with UK)[11].

On examining Figure 1 and Tables 2 and 3, we can see that the SP-CPD model confirms the large differences in price levels among major regional capitals, with higher prices observed for various BHs in most (but not all) of the Northern regions than in Rome. More specifically, price level differences underscore the well-known division between the Northern and Southern regions. Most of the towns located in the Centre, North of Italy, are ‘‘more expensive’’ than Rome for most of the BHs considered in our analysis.

[11] It is worth noting that we evaluated the extent of the bias of $\hat{\alpha}_r$ as an estimator of α_r used for obtaining sub-national PPPs. As the value of $\hat{\sigma}^2$ is small and n is very large in our case, the bias correction is negligible. Regarding the “Beef” BH and considering the cities for which we found the greatest bias corrections, we obtained 1.001499 for Ancona and 1.001362 for Naples by means of using the “less biased” estimator suggested in Appendix D.

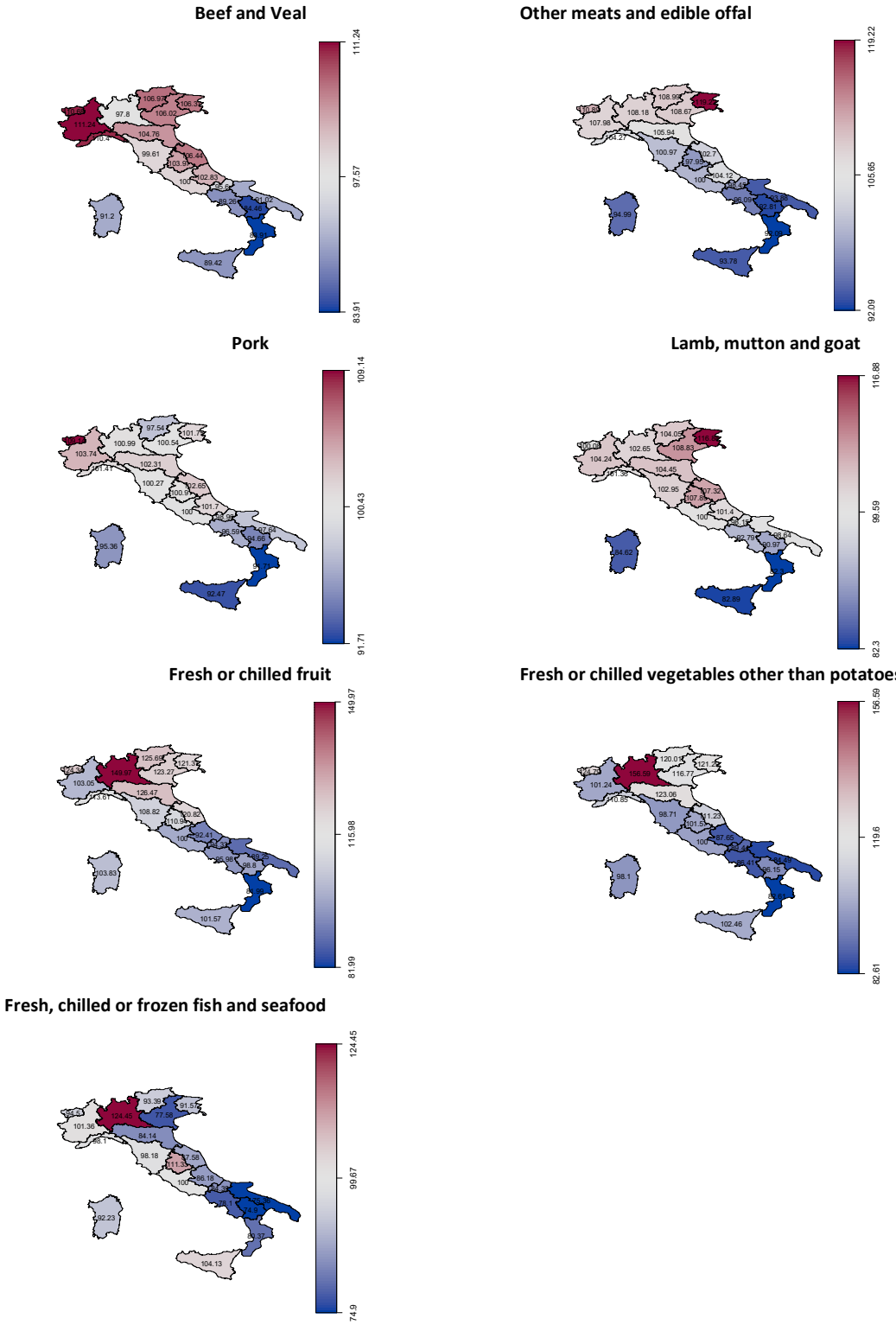
Table 2. **Sub-national PPPs for the 20 Italian regional capitals (Rome = 100) using the SP-CPD model**

	Beef and veal		**Other meats and edible offal**		**Pork**		**Lamb, mutton and goat**		**Fresh or chilled fruit**		**Fresh or chilled vegetables**		**Fresh, chilled or frozen fish and seafood**	
Region	Rome=100	S.E.	Rome=100	S.E.	Rome=100	S.E.	Rome=100	S.E.	Rome=100	S.E.	Rome=100	S.E.	Rome=100	S.E.
North														
Aosta	110.68*	7.07	110.89*	7.18	109.14*	5.84	100.08	6.84	124.34***	3.71	124.79***	4.22	94.5*	3.79
Torino	111.24**	5.85	107.98*	5.01	103.74	3.98	104.24	5.89	103.05*	2.29	101.24	2.68	101.36	3.79
Genova	110.40	5.85	104.27	5.14	101.41	4.16	101.36	5.98	113.61***	2.96	110.85***	3.49	98.1	3.78
Milano	97.80	5.03	108.18*	5.11	100.99	3.88	102.65	5.57	149.97***	3.36	156.59***	4.21	124.45***	4.37
Trento	106.97	5.92	108.99*	5.57	97.54	4.34	104.05	6.17	125.69***	3.38	120.01***	3.74	93.39*	4.17
Venezia	106.02	5.74	108.67*	5.16	100.54	3.99	108.83	6.15	123.27***	3.45	116.77***	3.72	77.58***	2.86
Trieste	106.37	6.42	119.22***	7.29	101.75	5.1	116.88	7.99	121.31***	3.13	121.29***	3.71	91.57***	3.42
Bologna	104.76	5.28	105.94	4.51	102.31	3.57	104.45	5.45	126.47***	3.27	123.06***	3.69	84.14***	3.18
Centre														
Firenze	99.61	5.17	100.97	4.45	100.27	3.63	102.95	5.55	108.82***	2.74	98.71	2.91	98.18	3.59
Ancona	106.44	5.30	102.7	4.54	102.65	3.65	107.32*	5.60	120.82***	3.56	111.23***	3.84	87.58***	3.45
Perugia	103.97	5.57	97.95	4.82	100.91	4.05	107.89*	6.35	110.94***	3.19	101.57	3.41	111.33***	4.13
South and Islands														
L'Aquila	102.83	5.43	104.12	5.28	101.7	4.03	101.4	5.69	92.41***	2.32	87.65***	2.67	86.18***	3.80
Campobasso	95.60	5.406	98.45	4.21	98.96	3.86	98.15	5.42	94.33***	2.13	89.44***	2.99	84.93	3.71
Napoli	89.26**	4.66	96.09	4.3	96.59	3.58	92.79*	5.03	95.98**	2.42	86.41***	2.71	78.1***	3.15
Potenza	84.46***	4.78	92.81	4.68	94.66*	4.00	90.97**	5.39	98.8	2.74	96.15***	2.87	74.9***	2.60
Bari	91.02**	5.40	93.88	5.02	97.64	4.11	98.64	5.86	89.25***	2.29	84.49***	2.64	75.36***	2.82
Catanzaro	83.91***	5.02	92.09	5.51	91.71**	4.32	82.30***	5.21	81.99***	2,.00	82.61***	2.31	80.37***	2.89
Palermo	89.42**	5.35	93.78	5.16	92.47**	4.34	82.89***	5.25	101.57	2.38	102.46	2.84	104.13	4.33
Cagliari	91.2**	5.13	94.99	4.67	95.36	3.92	84.62***	4.98	103.83*	2.75	98.1	3.12	92.23***	3.66
Obs.	177		74		89		42		1,643		2,018		888	
λ	2.587		3.238		5.514		0.821		0.893		1.147		0.628	
RMSE	0.169		0.132		0.136		0.083		0.175		0.231		0.189	
AIC	-594.2		-276.2		-333.7		-181.0		-5635.5		-5693.2		-2857.3	
EDF	16.92		10.84		9.90		12.82		93.28		110.24		49.12	

Notes: * 10 %, ** 5 %, *** 1 %; S.E. denotes standard error; PPP for Campobasso has been estimated with kriging. S.E. has been computed using the Delta method (Held and Sabanes Bove, 2014)

RMSE: Squared root of the mean squared error; AIC: Akaike information criterion; EDF: Effective degrees-of-freedom

Figure 1. Sub-national PPPs for the 20 Italian regional capitals (Rome = 100) using the SP-CPD model



Price differences show different spatial patterns (Table 2 and Figure 1) for different BHs. The PPP values for Beef and veal range from 83.91 for Catanzaro to 111.24 for Torino. Aosta (110.68), Genova (110.4) and Ancona (106.44) have the next highest regional PPPs. Potenza (84.46), Palermo (89.42) and Napoli (89.26) have lower prices for Beef and veal than Rome.

There are greater differences for the Fresh or chilled vegetables BH: Milan (156.59) is the most expensive regional capital while Catanzaro is the cheapest (82.61). Milan is also the regional capital with the highest prices for the Fresh or chilled fruit BH (149.97) and the Fresh, chilled or frozen fish and seafood BH (124.45).

With the aim of comparing the SP-CPD results with those obtained in a model with no penalty, we estimated a basic hedonic CPD model for the 7 BHs. We first estimated a spatial error specification of the CPD with and without geographical areas in order to check for spatial effects in consumer prices and to determine whether including areal dummies in the CPD model (4) sufficiently accounts for these spatial effects or not. As indicated in the introduction and following Biggeri et al. (2017), we used the acronym SEM-CPD for the spatial error specification of the CPD with geographical areas. Consequently, the acronym SEM-PD is used when the geographical areas (countries or regions) are not included in the model (see Online Appendix A for details).

Table 3 reports the estimation results for the "Fresh or chilled vegetables" and "Fresh, chilled or frozen fish and seafood" BHs while spatial price indexes for all of the BHs are reported in the supporting information (Online Appendix G). It is worth noting that when geographical areas are not included in the spatial specifications (SEM-PD in Table G5 of Online Appendix G) the coefficient for the spatially-dependent error term κ is very high with a value equal to 0.56 for the "Fresh or chilled vegetables" BH and 0.61 for the "Fresh, chilled or frozen fish and seafood" BH, thus indicating that a covariate with spatial effects should be used when explaining consumer prices. However, when dummies for geographical areas are included,

then $\kappa=0$, thus demonstrating the importance of geographical dummies when the spatial effects inherent in consumer prices must be captured and be present in the model. It is also important to note that the diagnostic statistics are improved (see Tables G1-G5 of Online Appendix G in the supporting information file).

Therefore, spatial effects are important and must be taken into account when estimating sub-national PPPs. However, apart from the fact that SEM-CPD practically coincides with CPD, another important reason for not using the SEM-CPD model is that we do not know exactly where product prices were collected within each region and consequently all of the regional prices observed are considered independent. This is why the geostatistical alternative using kriged means as well as the spatial econometric alternative via the SEM-CPD specification are disregarded and therefore a spatially penalized CPD model is used. The best way of accounting for these spatial effects is to add a penalization term to the CPD model so that geographical dummies are included in the set of explanatory variables and the coefficients of neighbouring geographical variables are penalized in order to obtain better (and more realistic) inferences. In this empirical application it appears that there are not only theoretical but also practical considerations in favour of using the SP-CPD model. Indeed, economic theory supports the idea that prices do not change abruptly between one region and its neighbouring regions and the penalized model may mitigate the effect of having few observations in some BHs (as is the case with the "Meat" products in our data set) and in some regions, thus reducing the standard errors of the estimates. More specifically, it is important to determine how much the PPP estimates and their standard errors differ among the various models used in this paper, namely the traditional CPD model, the SEM-CPD and the proposed SP-CPD models. With respect to the estimated PPPs, it is important to note that the CPD and SEM-CPD estimates are numerically the same. This is expected because the estimated values of κ are practically zero for every BH in SEM-CPD. However, using the PPP estimates obtained from SP-CPD

generally leads to an upward trend in the PPP estimates as obtained from traditional CPD and SEM-CPD models. As expected, larger differences between the PPPs estimated using traditional CPD regressions or SEM-CPD and SP-CPD models are observed for the BHs with a limited number of observations, as in the case of "Meat" group of product (Beef and veal, Other meats and edible offal, Pork, Lamb, mutton and goat) and high values are obtained for lambdas in SP-CPD. Moreover, a lower variability of the PPPs provided by the SP-CPD model is observed compared to the variability of the PPPs obtained by using CPD and SEM-CPD, especially for the BHs with a limited number of observations as shown in Tables 2 -3 and Tables G1-G4 in Online Appendix G.

Focusing on the standard errors of the PPP estimates, it is important to emphasize both SEM-CPD and SP-CPD provide smaller standard errors compared to those obtained using traditional CPD models. As expected, the PPP estimates for the BHs characterized by a small number of observations (BH1-Beef and Veal, BH2-Other meats and edible offal, BH3-Pork, BH4-Lamb, mutton and goat) show the lowest values of standard errors when the SP-CPD is used for carrying out multilateral comparisons. For Fresh or chilled fruit, Fresh or chilled vegetables and Fresh, chilled or frozen fish group of products (BH5, BH6, BH7 respectively), the values of standard errors provided by SEM-CPD and SP-CPD are quite similar yet slightly lower than those obtained using the traditional CPD model. Once again these results illustrate the advantages of using SP-CPD model when data are scarce.

The goodness-of-fit of the various models, reported in Tables 2-3 and Tables G1-G5 in Online Appendix G, show that the SP-CPD model performs best in terms of AIC. The lowest AIC values for all BHs are obtained when using the SP-CPD model. We note, however, that differences among the AICs of the three models (CPD, CPD-SEM and SP-CPD) are negligible when the groups of products are characterized by a large number of price observations, as in the case of BH5, BH6 and BH7. Nevertheless, these differences are non-negligible for the

other BHs (BH1, BH2, BH3, BH4) where the reduction in effective degrees-of-freedom in the SP-CPD model clearly compensates for the increase in the squared sum of errors (see Table 2 and Tables G1-G4 in Online Appendix G). Since RMSE does not take into account the degrees-of-freedom, the AIC is preferred to RMSE as a measure of goodness-of-fit.

Finally, it is clear that the SP-CPD model provides robust results even in the case of incorrect or inaccurate specification of neighbours largely due to the adjustment of the smoothing parameter which compensates for the increase in Error Sum of Squares (SSE). For checking robustness of the results, we estimated SP-CPD models using first-, second-, third- and fourth-neighbours for defining the spatial matrix omega[12]. When second-order neighbours and especially third-and fourth-order neighbours are considered, the results are practically the same because the increase in SSE due to the "erroneous" configuration of neighbours is offset by a decrease in the value of the smoothing parameter, especially in the case of BHs with large amounts of price data. In terms of AIC, the best SP-CPD specification is the one with only first-order neighbors for BHs with the fewest observations (Beef and Veal, Other meats and edible offal, Pork and Lamb, mutton and goat), while for the other BHs considered in this study, that is for BH5, BH6 and BH7, with very low smoothing parameter values, AIC is practically the same regardless of the order of neighbourhood considered.

These results clearly suggest that SP-CPD should be preferred to CPD and SEM-CPD, especially in the case of where only limited number of price observations are available which is commonplace when CPI data are used.

[12] These results are available upon request from the authors.

Table 3. Estimates and sub-national PPPs for the 20 Italian regional capitals (Rome = 100) using SEM-CPD and CPD hedonic models: ***Fresh or chilled vegetables and Fresh, chilled or frozen fish and seafood***

	Fresh or chilled vegetables				Fresh, chilled or frozen fish and seafood			
	SEM-CPD		CPD		SEM-CPD		CPD	
	PPPs	S.E.	PPPs	S.E.	PPPs	S.E.	PPPs	S.E.
North								
Aosta	124.97***	4.18	124.97***	4.30	93.67**	3.75	93.67**	3.85
Torino	100.41	2.49	100.41	2.56	100.67	3.76	100.67	3.87
Genova	110.76***	3.45	110.76***	3.54	97.54	3.76	97.54	3.86
Milano	157.75***	3.99	157.75***	4.1	125.19***	4.38	125.19***	4.5
Trento	119.35***	3.65	119.35***	3.76	92.43***	4.2	92.43***	4.32
Venezia	115.81***	3.7	115.81***	3.80	76.13***	2.77	76.13***	2.85
Trieste	121.08***	3.61	121.08***	3.72	91.07***	3.36	91.07***	3.45
Bologna	123.45***	3.69	123.45***	3.79	82.49***	3.16	82.49***	3.25
Centre								
Firenze	97.96	2.82	97.96	2.9	97.58	3.55	97.58	3.65
Ancona	111.97***	3.94	111.97***	4.05	86.28***	3.45	86.28***	3.55
Perugia	101.27	3.42	101.27	3.51	111.15***	4.09	111.15***	4.21
South and Islands								
L'Aquila	87.05***	2.61	87.05***	2.68	85.2***	3.87	85.2***	3.98
Campobasso	97.01	2.68	97.01	2.68	94,25*	2.68	94.25*	2.68
Napoli	85.73***	2.7	85.73***	2.77	76.91***	3.19	76.91***	3.28
Potenza	96.4	2.79	96.4	2.87	74.4***	2.55	74.4***	2.63
Bari	84.06***	2.6	84.06***	2.67	74.65***	2.8	74.65***	2.88
Catanzaro	82.18***	2.21	82.18***	2.27	79.68***	2.83	79.68***	2.92
Palermo	102.5	2.64	102.5	2.71	104.09	4.36	104.09	3.28
Cagliari	97.97	3.09	97.97	3.17	91.48***	3.66	97.97***	3.17
κ	0.000				0.000			
Obs.	2,018		2,018		888		888	
RMSE	0.231		0.231		0.189		0.189	
AIC	-5,690.5		-5,692.5		-2,854.6		-2,856.6	
DF	112		111		51		50	

Notes: * 10 %, ** 5 %, *** 1 %; S.E. denotes standard error; PPP for Campobasso has been estimated with kriging. S.E. has been computed using the Delta method (Held and Sabanes Bove, 2014) RMSE: Squared root of the mean squared error; AIC: Akaike information criterion; DF: Degrees-of-freedom

4 Concluding Remarks

Over the last two decades, major theoretical improvements have been made to the CPD methodology and it is now considered the principal method of aggregation under the stochastic approach to price index number construction. However, very few studies to date have focused on spatial dependence in consumer price index construction.

This paper marks a departure from previous literature on sub-national PPPs by proposing a new CPD methodology for calculating spatial price indexes according to the stochastic approach. This new CPD methodology takes spatial dependencies into account in a non-

explicit way by penalizing the differences in the PPPs of neighbouring spatial units, thus providing a smooth map of spatial price indexes. Instead of estimating the smoothing parameter externally via cross-validation or information criteria, we transform the SP-CPD model into a mixed model which allows us to simultaneously estimate all the parameters of the model. The usefulness and potential of this approach are illustrated by estimating the SP-CPD model using real data obtained from the official CPI survey carried out in Italy in 2014. We selected seven basic headings (BHs) within the Food and non-alcoholic beverages group by referring to the comparability and representativity requirements. Our results show that the differences in consumer price levels across geographical areas are not negligible and illustrate the well-known divide between the Northern-Central and Southern regions. Compared to the traditional CPD model and to the CPD model with spatially correlated errors, the SP-CPD model is particularly useful since it allows us to draw better statistical inferences especially with a limited number of price observations, which is frequently the case when CPI data are used for estimating sub-national PPPs. The results of this study provide a basis for further statistical developments in the estimation of PPPs including: the possibility of substituting the areal effects in the traditional CPD models with a spatial drift based on the spatial coordinates representing the various areas under study; and specification of spatial econometric models accounting for spatial autocorrelation and spatial heterogeneity, which also include non-parametrically-specified smooth functions of some predictor variables to account for non-linear relationships between those predictors and the response.

References

Anderson, J. E., and Van Wincoop, E., "Trade costs", *Journal of Economic Literature*, 42, (3), 691-751, 2004.

Aten, B. H., "Evidence of spatial autocorrelation in international prices", *Review of Income and Wealth*, 42 (2), 149–163, 1996.

_____,"Does space matter? International comparisons of the prices of tradables and nontradables". *International Regional Science Review*, 20(1-2), 35-52, 1997.

_____, "Cities in Brazil: An interarea price comparison", In *International and interarea comparisons of income, output, and prices* (eds A. Heston and R. Lipsey), pp. 211–229, Chicago: University of Chicago Press, 1999.

_____, "Interarea price levels: An experimental methodology", *Monthly Labor Review*, 129 (9), 47–61, 2006.

Biggeri, L., De Carli, R., and Laureti, T., "The interpretation of the PPPs: A method for measuring the factors that affect the comparisons and the integration with the CPI work at regional level", In *Proc. Joint UNECE/ILO Meeting on Consumer Price Indices*, May 8–9, Geneva, 2008.

Biggeri, L., Laureti, T., and Polidoro, F., "Computing sub-national PPPs with CPI data: an empirical analysis on Italian data using country product dummy models", *Social Indicators Research*, 131(1), 93-121, 2017.

Choi, C. Y., and Choi, H., "Does distance reflect more than transport costs?", *Economics Letters*, 125(1), 82-86, 2014.

_____ , "The role of two frictions in geographic price dispersion: When market friction meets nominal rigidity", *Journal of International Money and Finance*, 63, 1-27, 2016.

Clements, K.W., and Izan, H.Y., "The Measurement of Inflation: A Stochastic Approach", *Journal of Business and Economic Statistics*, 5, 339–350, 1987.

Clements, K. W., Izan, I. H., and Selvanathan, E. A., "Stochastic index numbers: a review", *International Statistical Review*, 74 (2), 235-270, 2006.

Coondoo, D., Majumder, A., and Ray, R., "A method of calculating regional consumer price differentials with illustrative evidence from India", *Review of Income and Wealth*, 50(1), 51–68, 2004.

Cressie, N. A. C., *Statistics for Spatial Data*. John Wiley & Sons: New York, 2015.

Crucini, M. J., Shintani, M., and Tsuruga, T., "Noisy information, distance and law of one price dynamics across US cities", *Journal of Monetary Economics*, 74, 52-66, 2015.

Deaton, A., "Prices and poverty in India, 1987–2000", *Economic and Political Weekly*, 38(4), 362–368, 2003.

Deaton, A., and Dupriez, O., "Spatial price differences within large countries". Manuscript, Princeton University, July, 2011.

de Haan, J., and Krsinich, F., "Scanner data and the treatment of quality change in nonrevisable price indexes", *Journal of Business & Economic Statistics*, 32(3), 341-358, 2014.

Dikhanov, Y., Palanyandy, C., and Capilit, E., "Subnational purchasing power parities toward integration of international comparison program and the consumer price index: The case of Philippines", ADB Economics Working Paper Series, No. 290, Mandaluyong City: Asian Development Bank, 2011.

Diewert, W. E., "Microeconomic approaches to the theory of international comparisons", NBER Technical Working Paper No. 53, Cambridge, MA: National Bureau of Economic Research, 1986.

_____., "Weighted country product dummy variable regressions and index number formulae", *Review of Income and Wealth*, 51(4), 561–570, 2005.

_____ ,"On the Stochastic approach to index numbers". In: Diewert, W.E., Balk, Bert M., Fixler, Dennis, Fox, Kevin J., Nakamura, Alice O. (Eds.), *Price and Productivity Measurement*. Trafford Press, 235–262, 2010.

Eichhorn, W., and Voeller, J., "Axiomatic foundation of price indexes and purchasing power parities, in Price Level Measurement", In *Contributions to Economic Analysis*, W.E. Diewert (Editor), North-Holland: Elsevier Science Publisher, 1990.

Engel, C., and Rogers, J. H., "Violating the Law of One Price: Should We Make a Federal Case Out of It?", *Journal of Money, Credit and Banking*, 33(1), 1-15, 2001.

Fahrmeir, L., Kneib, T., Lang, S., and Marx. B., *Regression*. Berlin: Springer-Verlag, 2013.

Hajargasht, G., and Rao, D. S., "Stochastic approach to index numbers for multilateral price comparisons and their standard errors", *Review of Income and Wealth*, 56, S32–S58, 2010.

Held, L., and Sabanes Bove, D., *Applied Statistical Inference*. Springer: Heidelberg, 2014.

Hill, R. J., and Syed, I. A., "Improving International Comparisons of Prices at Basic Heading Level: An Application to the Asia-Pacific Region", *Review of Income and Wealth*, 61(3), 515-539, 2015.

Kokoski, M., Moulton, B., and Zieschang, K., "Interarea price comparisons for heterogenous goods and several levels of commodity aggregation", In *International and interarea*

comparisons of income, output and prices (eds. A. Heston and R. Lipsey), pp. 123–166. Chicago: University of Chicago Press, 1999.

Koo, J., Phillips, K. R., and Sigalla, F. D., "Measuring regional cost of living", *Journal of Business and Economic Statistics*, 18(1), 127–136, 2000.

Majumder, A., Ray, R., and Santra, S. "Sensitivity of Purchasing Power Parity Estimates to Estimation Procedures and their Effect on Living Standards Comparisons". *Journal of Globalization and Development*, 8(1), 2017.

_____ "Sensitivity of global and regional poverty rates to alternative purchasing power parities". *Indian Growth and Development Review*, 11(1), 34-56, 2018.

Majumder, A., Ray, R., and Sinha, K., "Estimating purchasing power parities from household expenditure data using complete demand systems with application to living standards comparison: India and Vietnam", *Review of Income and Wealth*, 61(2), 302–328, 2015.

Montero, J. M., Fernández-Avilés, G., and Mateu, J., *Spatial and spatio-temporal kriging and modelling*. Chichester: Wiley, 2015.

Özyurt, S., and Dees, S., "Regional dynamics of economic performance in the EU: To what extent spatial spillovers matter?" *European Central Bank Working Paper Series*, Working Paper No. 1870, Frankfurt am Main: European Central Bank, 2015.

Rao, D. S. P., "Weighted EKS and Generalised CPD methods for aggregation at basic heading level and above basic heading level", Joint World Bank—OECD seminar on purchasing power parities. Washington, DC: Recent Advances in Methods and Applications, 2001.

_____, "The country-product-dummy method: A stochastic approach to the computation of purchasing power parities in the ICP", Paper presented at the SSHRC

conference on index numbers and productivity measurement, Vancouver, June 30–July 3, 2004.

Rao, D. S. P., and Hajargasht, G., "Stochastic approach to computation of purchasing power parities in the International Comparison Program (ICP)" *Journal of Econometrics*, 191(2), 414-425, 2016.

Roos, M., "Regional price levels in Germany", *Applied Economics*, 38(13), 1553–1566, 2006.

Rue, H., and Held, L., *Gaussian Markov Random Fields: Theory and Applications* (*Monographs on Statistics and Applied Probability*). Boca Raton: Chapman & Hall/CRC, 2005.

Schabenberger, O. and Gotway, C. A., Statistical Methods for Spatial Data Analysis Boca Raton: Chapman & Hall/CRC, 2005.

Selvanathan, E. A., "A Note on the Stochastic Approach to Index Numbers", *Journal of Business and Economic Statistics*, 7, 471–474, 1989.

Selvanathan, E. A., and Rao, D. S. P., *Index Numbers: A Stochastic Approach*. London: Macmillan, 1994.

Tobler, W., "A computer movie simulating urban growth in the Detroit region". *Economic Geography*, 46(2), 234–240, 1970.

Wingfield, D., Fenwick, D., and Smith, K., "Relative regional consumer price levels in 2004", *Economic Trends (Office for National Statistics, UK)*, 615, 36–46, 2005.

World Bank, *Measuring the Real Size of the World Economy*, World Bank, Washington DC, 2013.

Supporting information

Additional Supporting Information may be found in the online version of this paper:

Appendix A: SEM-CPD and SEM-PD models;

Appendix B: Connection between penalized models and mixed models;

Appendix C: Mixed model representation of SP-CPD model;

Appendix D: Bias adjustment for expected prices;

Appendix E: Socio-economic characteristics of Italian regions;

Appendix F: Validating the kriging estimation for Campobasso-Molise;

Appendix G: Estimates and sub-national PPPs using hedonic SEM-PD, SEM-CPD and CPD hedonic models.

Supporting information to A Stochastic Model with Penalized Coefficients for Spatial Price Comparisons: An Application to Regional Price Indexes in Italy

Appendix A: SEM-CPD and SEM-PD models

The specification of the spatial error model is (Anselin, 1988):

$$\mathbf{y} = a\,\mathbf{i}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u}: \; k\,\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ represents an $n \times 1$ vector of values of the response variable, $\mathbf{X}$ is an $n \times k$ matrix containing the observations of the explanatory variables, $\mathbf{i}_n$ is a $n \times 1$ unit vector for the intercept, α is the intercept parameter, $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression parameters and $\boldsymbol{\varepsilon}$ an $n \times 1$ vector of *iid* disturbances with $N(\mathbf{0}, \sigma^2\mathbf{I})$ distribution. $\mathbf{W}$ is a n n row-stochastic matrix of spatial weights and $\mathbf{W}\mathbf{u}$ is the spatially-dependent error vector, $\mathbf{u}$. k is the spatial parameter weighting the spatially-dependent error vector.

The spatial error specification of the CPD model (or SEM-CPD model, as it is named in Biggeri et al., 2017) refers to a spatial error specification where: (i) the response variable is the logarithm of annual price of commodities observed in the outlets of the different areas involved in the comparison; and (ii) the explanatory variables included in the model are basically commodity and area dummy variables, but also quality characteristics, such as type of outlet, product brand, etc. (see Biggeri et al., 2017 for details).

The spatial error specification of the CPD model without areas (SEM-PD) is the same as the SEM-CPD but does not include dummies for the geographical areas because its estimation is aimed at exploring the existence of spatial effects.

Appendix B: Connection between penalized models and mixed models

The connection between penalized strategies and mixed models lies in the similarity of the penalized fitting criterion to the maximization issue that produces the mixed model equations and estimates for $\boldsymbol{\delta}^*$ and $\boldsymbol{\alpha}$.

Consider the mixed model (7) in online Appendix C where both the variance-covariance matrix of the random effects and the variance-covariance matrix of the errors have been substituted with more general known matrices **G** and **R**, allowing for correlated random effects and errors, respectively:

$$\ln \mathbf{p} = \boldsymbol{\Gamma}^{*}\boldsymbol{\delta}^{*} + \boldsymbol{\psi}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R}), \quad \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G})$$

The marginal and conditional-to-$\boldsymbol{\alpha}$ distributions of $\ln\mathbf{p}$ are respectively:

$$\ln \mathbf{p} \sim N\left(\boldsymbol{\Gamma}^{*}\boldsymbol{\delta}^{*}, \mathbf{V}\right), \quad \mathbf{V} = \mathbf{R} + \boldsymbol{\psi}\mathbf{G}\boldsymbol{\psi}',$$

$$\left(\ln \mathbf{p} \middle/ \boldsymbol{\alpha}\right) \sim N\left(\boldsymbol{\Gamma}^{*}\boldsymbol{\delta}^{*} + \boldsymbol{\psi}\boldsymbol{\alpha}, \mathbf{R}\right).$$

Consequently, the likelihood and log-likelihood functions of both the fixed and random effects (proportional to the joint density of both $\ln\mathbf{p}$ and $\boldsymbol{\alpha}$) are as follows:

$$L\left(\boldsymbol{\delta}^{*}, \boldsymbol{\alpha}\right) = L\left(\ln \mathbf{p} \middle/ \boldsymbol{\alpha}\right) L\left(\boldsymbol{\alpha}\right),$$
$$\ln L\left(\boldsymbol{\delta}^{*}, \boldsymbol{\alpha}\right) = \ln L\left(\ln \mathbf{p} \middle/ \boldsymbol{\alpha}\right) + \ln L\left(\boldsymbol{\alpha}\right).$$

and, under the assumption that **R** and **G** are known,

$$\ln L\left(\boldsymbol{\delta}^{*}, \boldsymbol{\alpha}\right) = c - \frac{1}{2}\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^{*}\boldsymbol{\delta}^{*} - \boldsymbol{\psi}\boldsymbol{\alpha}\right)' \mathbf{R}^{-1}\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^{*}\boldsymbol{\delta}^{*} - \boldsymbol{\psi}\boldsymbol{\alpha}\right) - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{G}^{-1}\boldsymbol{\alpha}.$$

with *c* being a constant.

The maximization of $\ln L\left(\boldsymbol{\delta}^{*}, \boldsymbol{\alpha}\right)$ with known **R** and **G** results in the Henderson equations:

$$\hat{\boldsymbol{\alpha}} = \mathbf{G}\boldsymbol{\psi}'\mathbf{V}^{-1}\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^{*}\hat{\boldsymbol{\delta}}^{*}\right),$$

$$\hat{\boldsymbol{\delta}}^{*} = (\boldsymbol{\Gamma}^{*\prime}\mathbf{V}^{-1}\boldsymbol{\Gamma}^{*})'\boldsymbol{\Gamma}^{*\prime}\mathbf{V}^{-1} \ln \mathbf{p}.$$

However, it is important to note that the maximization of $\ln L\left(\boldsymbol{\delta}^{*}, \boldsymbol{\alpha}\right)$ with known **R** and **G** is equivalent to the minimization of a penalized model criterion with an externally determined penalization parameter λ, where the parameters included in $\boldsymbol{\alpha}$ are penalized via matrix **G**.

In the mixed model representation of P-CPD models, $\mathbf{R} = \sigma^{2}\mathbf{I}$ and $\mathbf{G} = \tau^{2}\mathbf{I}$. In such a case,

$$\ln L\left(\boldsymbol{\delta}^{*}, \boldsymbol{\alpha}\right) = c - \frac{1}{2\sigma^{2}}\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^{*}\boldsymbol{\delta}^{*} - \boldsymbol{\psi}\boldsymbol{\alpha}\right)' \left(\ln \mathbf{p} - \boldsymbol{\Gamma}^{*}\boldsymbol{\delta}^{*} - \boldsymbol{\psi}\boldsymbol{\alpha}\right) - \frac{1}{2\tau^{2}}\boldsymbol{\alpha}'\boldsymbol{\alpha}.$$

But maximizing the above log-likelihood with respect to $\boldsymbol{\delta}^*$ and $\boldsymbol{\alpha}$ is the same as maximizing:

$$\begin{aligned}\left(\ln L\left(\ln \mathbf{p}, \boldsymbol{\alpha}\right)\right)^* &= \sigma^2\left(c - \frac{1}{2\sigma^2}\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\boldsymbol{\delta}^* - \boldsymbol{\psi}\boldsymbol{\alpha}\right)'\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\boldsymbol{\delta}^* - \boldsymbol{\psi}\boldsymbol{\alpha}\right) - \frac{1}{2\tau^2}\boldsymbol{\alpha}'\boldsymbol{\alpha}\right) \\ &= c^* - \frac{1}{2}\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\boldsymbol{\delta}^* - \boldsymbol{\psi}\boldsymbol{\alpha}\right)'\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\boldsymbol{\delta}^* - \boldsymbol{\psi}\boldsymbol{\alpha}\right) - \frac{\sigma^2}{2\tau^2}\boldsymbol{\alpha}'\boldsymbol{\alpha}.\end{aligned},$$

which is equivalent to minimizing the penalized least squares criterion:

$$Q = \left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\boldsymbol{\delta}^* - \boldsymbol{\psi}\boldsymbol{\alpha}\right)'\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\boldsymbol{\delta}^* - \boldsymbol{\psi}\boldsymbol{\alpha}\right) + \frac{1}{2}\lambda\boldsymbol{\alpha}'\boldsymbol{\alpha},$$

with $\lambda = \frac{\sigma^2}{\tau^2}$.

In practice, σ^2 and τ^2 (**G** and **R** in general**)** are unknown, as is λ in penalized models.

Taking the marginal distribution of ln**p** as a starting point, the corresponding log-likelihood, up to additive constants, is:

$$\ln L\left(\boldsymbol{\delta}^*, \sigma^2, \tau^2\right) = -\frac{1}{2}\left(\ln\left|\mathbf{V}\left(\sigma^2, \tau^2\right)\right| + \left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\boldsymbol{\delta}^*\right)'\mathbf{V}^{-1}\left(\sigma^2, \tau^2\right)\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\boldsymbol{\delta}^*\right)\right).$$

The maximization of $\ln L\left(\boldsymbol{\delta}^*, \sigma^2, \tau^2\right)$ with respect to $\boldsymbol{\delta}^*$, (σ^2 and τ^2 holding fixed) results in:

$$\hat{\boldsymbol{\delta}}^*\left(\sigma^2, \tau^2\right) = \left(\boldsymbol{\Gamma}^{*\prime}\mathbf{V}^{-1}\left(\sigma^2, \tau^2\right)\boldsymbol{\Gamma}^*\right)^{-1}\boldsymbol{\Gamma}^{*\prime}\mathbf{V}^{-1}\left(\sigma^2, \tau^2\right)\ln \mathbf{p}.$$

Inserting $\hat{\boldsymbol{\delta}}^*\left(\sigma^2, \tau^2\right)$ in $\ln L\left(\boldsymbol{\delta}^*, \sigma^2, \tau^2\right)$ we have the profile log-likelihood:

$$\ln L_P\left(\sigma^2, \tau^2\right) = -\frac{1}{2}\left(\ln\left|\mathbf{V}\left(\sigma^2, \tau^2\right)\right| + \left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\hat{\boldsymbol{\delta}}^*\left(\sigma^2, \tau^2\right)\right)'\mathbf{V}^{-1}\left(\sigma^2, \tau^2\right)\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\hat{\boldsymbol{\delta}}^*\left(\sigma^2, \tau^2\right)\right)\right).$$

The maximization of $\ln L_P\left(\sigma^2, \tau^2\right)$ with respect to the variance components σ^2 and τ^2 provides the maximum likelihood estimates of these components (see Farhmeir *et al*., 2013 and the references therein).

However, the ML estimates are biased because ML does not take the loss of degrees of freedom into account due to the estimation of the fixed effects. This is why the estimation of σ^2 and τ^2 is performed using REML.

As is well known, the relationship between the logarithm of the likelihood and the logarithm of the profile likelihood is as follows (see Harville, 1974, for example):

$$\ln L_R\left(\sigma^2,\tau^2\right)=\ln L_P\left(\sigma^2,\tau^2\right)-\frac{1}{2}\ln\left|\boldsymbol{\Gamma}^{*\prime}\mathbf{V}^{-1}\left(\sigma^2,\tau^2\right)\boldsymbol{\Gamma}^{*}\right|.$$

The maximization of $L_R\left(\sigma^2,\tau^2\right)$ with respect to the variance components provides their REML estimates. For this purpose, the numerical optimization procedures included in the BayesX library can be used (Umlauf *et al.*, 2015; Belitz *et al.*, 2015).

Appendix C: Mixed model representation of SP-CPD model

A proper reparameterization of the vector of geographical coefficients **a** is as follows:

$$\mathbf{a}=\mathbf{X}\boldsymbol{\delta}+\mathbf{Q}\boldsymbol{\alpha},$$

where the matrices **X** and **Q** are chosen in order to rewrite the penalty $\lambda\mathbf{a}'\boldsymbol{\Omega}\mathbf{a}$ as $\lambda\boldsymbol{\alpha}'\boldsymbol{\alpha}$, so that the penalization, whatever the value of $\boldsymbol{\Omega}$, is only expressed in terms of a vector of independent and identically distributed random effects $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha}\sim\mathrm{N}\left(\mathbf{0},\mathbf{G}=\tau^2\mathbf{I}\right)$); $\boldsymbol{\delta}$ can then be considered as a vector of fixed effects. This enables us to separate the density into a non-informative distribution for the fixed effects and a non-singular normal distribution with a normalized density for the random effects, regardless of the penalization approach. The above objective requires **X** and **Q** to satisfy the following conditions:

(i) $\left(\mathbf{X},\mathbf{Q}\right)$ is full rank yielding a one-to-one transformation.

(ii) $\mathbf{X}'\boldsymbol{\Omega}=\mathbf{0}$, that is, $\boldsymbol{\delta}$ is not penalized.

(iii) $\mathbf{Q}'\boldsymbol{\Omega}\mathbf{Q}=\mathbf{I}$, so that $\boldsymbol{\alpha}$ is a vector of independent and identically distributed random effects.

In effect, the satisfaction of conditions (i)-(iii) implies that:

$$\lambda \mathbf{a}'\boldsymbol{\Omega}\mathbf{a} = \lambda\left(\mathbf{X}\boldsymbol{\delta}+\mathbf{Q}\boldsymbol{\alpha}\right)'\boldsymbol{\Omega}\left(\mathbf{X}\boldsymbol{\delta}+\mathbf{Q}\boldsymbol{\alpha}\right)$$
$$= \lambda\left(\boldsymbol{\delta}'\overbrace{\mathbf{X}'\boldsymbol{\Omega}}^{\mathbf{0}}\mathbf{X}\boldsymbol{\delta} + \boldsymbol{\delta}'\overbrace{\mathbf{X}'\boldsymbol{\Omega}}^{\mathbf{0}}\mathbf{Q}\boldsymbol{\alpha} + \overbrace{\boldsymbol{\alpha}'\mathbf{Q}'\boldsymbol{\Omega}\mathbf{X}\boldsymbol{\delta}}^{\boldsymbol{\delta}'\overbrace{\mathbf{X}'\boldsymbol{\Omega}}^{\mathbf{0}}\mathbf{Q}\boldsymbol{\alpha}} + \boldsymbol{\alpha}'\overbrace{\mathbf{Q}'\boldsymbol{\Omega}\mathbf{Q}}^{\mathbf{I}}\boldsymbol{\alpha}\right) = \lambda\boldsymbol{\alpha}'\boldsymbol{\alpha}$$

As a consequence of the above reparameterization, model (4) can be rewritten as[13]:

$$\ln\mathbf{p} = \mathbf{M}\left(\mathbf{X}\boldsymbol{\delta}+\mathbf{Q}\boldsymbol{\alpha}\right)+\begin{pmatrix}\mathbf{D}^* & \mathbf{Z}\end{pmatrix}\begin{pmatrix}\mathbf{b}\\ \mathbf{c}\end{pmatrix}+\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}\sim N\left(\mathbf{0},\sigma^2\mathbf{I}\right),\quad \boldsymbol{\alpha}\sim N\left(\mathbf{0},\tau^2\mathbf{I}\right),\quad \lambda = \sigma^2/\tau^2$$
$$\ln\mathbf{p} = \boldsymbol{\Gamma}\boldsymbol{\delta}+\boldsymbol{\psi}\boldsymbol{\alpha}+\begin{pmatrix}\mathbf{D}^* & \mathbf{Z}\end{pmatrix}\begin{pmatrix}\mathbf{b}\\ \mathbf{c}\end{pmatrix}+\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}\sim N\left(\mathbf{0},\sigma^2\mathbf{I}\right),\quad \boldsymbol{\alpha}\sim N\left(\mathbf{0},\tau^2\mathbf{I}\right),\quad \lambda = \sigma^2/\tau^2$$

where matrix $\begin{pmatrix}\mathbf{D}^* & \mathbf{Z}\end{pmatrix}$ can be inserted into an extended matrix $\boldsymbol{\Gamma}$ (denoted with an asterisk), so that $\begin{pmatrix}\mathbf{b}\\ \mathbf{c}\end{pmatrix}$ extends $\boldsymbol{\delta}$ (its extension is also denoted with an asterisk). In such a case,

$$\ln\mathbf{p} = \boldsymbol{\Gamma}^*\boldsymbol{\delta}^*+\boldsymbol{\psi}\boldsymbol{\alpha}+\boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon}\sim N\left(\mathbf{0},\sigma^2\mathbf{I}\right),\quad \boldsymbol{\alpha}\sim N\left(\mathbf{0},\tau^2\mathbf{I}\right),\quad \lambda = \sigma^2/\tau^2 \qquad (7)$$

$\boldsymbol{\Gamma}^*$ can also be extended with a unitary column corresponding to the intercept.

The only pending question is the determination of matrices $\mathbf{X}$ and $\mathbf{Q}$. With respect to $\mathbf{X}$, its orthogonality to $\boldsymbol{\Omega}$ can be achieved by using a basis of the null space of $\boldsymbol{\Omega}$ for the columns of $\mathbf{X}$. As for $\mathbf{Q}$, it can be obtained from the spectral decomposition of $\boldsymbol{\Omega}$. According to this decomposition, $\boldsymbol{\Omega} = \mathbf{P}\boldsymbol{\Lambda}_+\mathbf{P}'$, with $\boldsymbol{\Lambda}_+$ being a diagonal matrix whose diagonal elements are the non-null eigenvalues of $\boldsymbol{\Omega}$, and $\mathbf{P}$ being a matrix whose columns are the orthonormal eigenvectors corresponding to those non-null eigenvalues.

Then $\mathbf{Q}$ can be defined as $\mathbf{L}\left(\mathbf{L}'\mathbf{L}\right)^{-1}$, with $\mathbf{L} = \mathbf{P}\boldsymbol{\Lambda}_+^{1/2}$. In doing so, $\boldsymbol{\Omega} = \mathbf{P}\boldsymbol{\Lambda}_+\mathbf{P}' = \mathbf{L}\mathbf{L}'$ and $\mathbf{Q}'\boldsymbol{\Omega}\mathbf{Q} = \left(\mathbf{L}'\mathbf{L}\right)^{-1}\mathbf{L}'\boldsymbol{\Omega}\mathbf{L}\left(\mathbf{L}'\mathbf{L}\right)^{-1} = \left(\mathbf{L}'\mathbf{L}\right)^{-1}\mathbf{L}'\mathbf{L}\mathbf{L}'\mathbf{L}\left(\mathbf{L}'\mathbf{L}\right)^{-1} = \mathbf{I}$.

[13] In Appendix B, in this supporting information material, it can be seen that $\lambda = \sigma^2/\tau^2$.

Once the SP-CPD model is expressed as a mixed model, all the parameters of the model, that is, $\boldsymbol{\delta}^*, \sigma^2$ and λ can be estimated by REML (see Montero et al., 2012, for details). The expression of the restricted log-likelihood of σ^2 and λ, concentrated on $\boldsymbol{\delta}^*$, is as follows:

$$\log L_R(\sigma^2, \lambda) = -\frac{1}{2}\left\{\log|\mathbf{V}| + \log\left|\boldsymbol{\Gamma}^{*\prime}\mathbf{V}^{-1}\boldsymbol{\Gamma}^*\right| + (\ln \mathbf{p} - \boldsymbol{\Gamma}^*\hat{\boldsymbol{\delta}}^*)^T \mathbf{V}^{-1}(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\hat{\boldsymbol{\delta}}^*)\right\},$$

with $\mathbf{V} = \tau^2 \boldsymbol{\psi}\boldsymbol{\psi}' + \sigma^2 \mathbf{I}_n$ and $\hat{\boldsymbol{\delta}}^* = (\boldsymbol{\Gamma}^{*\prime}\mathbf{V}^{-1}\boldsymbol{\Gamma}^*)'\boldsymbol{\Gamma}^{*\prime}\mathbf{V}^{-1} \ln \mathbf{p}$.

σ^2 and λ are estimated using the optimization procedures implemented in the BayesX library. In addition, using the Henderson equations, the random effects can also be estimated, $\hat{\boldsymbol{\alpha}} = \hat{\tau}^2 \boldsymbol{\psi}'\hat{\mathbf{V}}^{-1}\left(\ln \mathbf{p} - \boldsymbol{\Gamma}^*\hat{\boldsymbol{\delta}}^*\right)$, with $\tau^2 = \frac{\sigma^2}{\lambda}$, which allows for the estimation of the vector of coefficients **a** in models (4) and (5) in the main text, so that all the parameters in that model have been estimated at the same time.

Appendix D: Bias adjustment for expected prices

This appendix offers some considerations on the expected prices (e^{a_r} in the case of a CPD model with only the areal factor). It is important to note that, given λ, model (4) in the main text is a fixed-effects model, as is model (2) in the main text. It is well known that in log-regression models, $\ln \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with errors following a Gaussian distribution, for a set of values $\mathbf{x} = \left(1, \ x_1, \ \ldots \ x_p\right)'$ the *r*th moment with respect to the origin is given by $M_r = e^{r\mathbf{x}'\boldsymbol{\beta} + \frac{r^2\sigma^2}{2}}$, $r = 0,1,2,\ldots$. Consequently, the ML estimator of such a moment is $\hat{M}_r = e^{r\mathbf{x}'\hat{\boldsymbol{\beta}} + \frac{r^2\hat{\sigma}^2}{2}}$, $r = 0,1,2,\ldots$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are the MV estimators of $\boldsymbol{\beta}$ and σ^2. As stated

in El-Shaarawi and Viveros (1997), typically $\hat{M}_r$ will exhibit some bias as an estimator of M_r . This bias usually depends on σ^2 and could be substantial, particularly when σ^2 is large. Under a lognormal regression model, El-Shaarawi and Viveros (1997) proposed the following "less biased" estimator of M_r :

$$\hat{\hat{M}}_r = e^{r\mathbf{x}'\hat{\boldsymbol{\beta}}+\frac{r^2\hat{\sigma}^2}{2}-\frac{r^2\hat{\sigma}^2}{2}\left[\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}+\frac{r^2\hat{\sigma}^2}{2(n-p)}+\frac{r^4\hat{\sigma}^4}{3(n-p)^2}\right]} = e^{r\mathbf{x}'\hat{\boldsymbol{\beta}}-\frac{r^2\hat{\sigma}^2}{2}\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}+\frac{r^2\hat{\sigma}^2}{2}\left[1+\frac{\hat{\sigma}^2}{2(n-p)}+\frac{r^2\hat{\sigma}^4}{3(n-p)^2}\right]}, \quad r=0,1,2,\ldots,$$

which, in the case of r =1, reduces to:

$$\hat{\hat{M}}_1 = E(\mathbf{y}) = e^{\mathbf{x}'\hat{\boldsymbol{\beta}}-\frac{\hat{\sigma}^2}{2}\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}+\frac{\hat{\sigma}^2}{2}\left[1+\frac{\hat{\sigma}^2}{2(n-p)}+\frac{\hat{\sigma}^4}{3(n-p)^2}\right]}.$$

In our case, where $\mathbf{y}=\mathbf{p}$, $\mathbf{X}=(\mathbf{M}\,|\,\mathbf{D}^*\,|\,\mathbf{Z})$ and $\boldsymbol{\beta}=(\mathbf{a}\,|\,\mathbf{b}\,|\,\mathbf{c})$, given the small value of $\hat{\sigma}^2$ and the large value of *n* the bias correction is negligible.

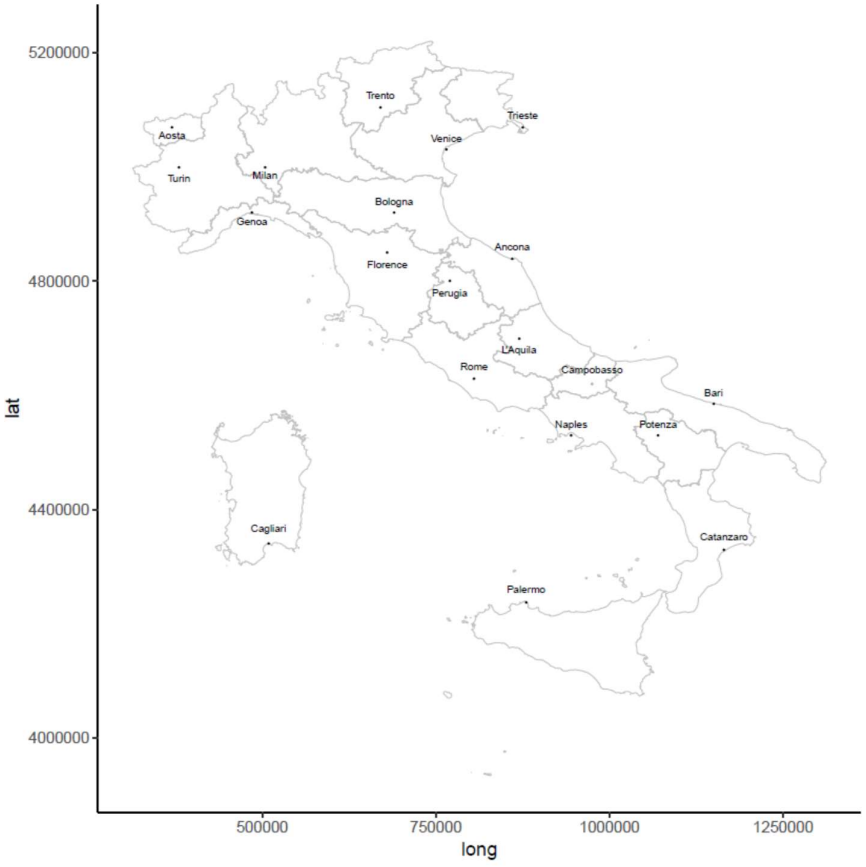
Online Appendix E: Socio-economic characteristics of Italian regions

The second territorial level of the Nomenclature of territorial units for statistics (Nuts2) divides Italy into 19 regions and the two autonomous provinces of Trento and Bolzano, which make up the Trentino-Alto Adige region. More specifically, Italy is subdivided into the following 20 regions, each of which has its own regional capital (specified in brackets): Aosta Valley (Aosta), Piedmont (Turin), Liguria (Genoa), Lombardy (Milan), Adige Trentino-Alto Adige (Trento), Veneto (Venice), Friuli-Venezia Giulia (Trieste), Emilia-Romagna (Bologna), Tuscany (Florence), Umbria (Perugia), Marche (Ancona), Lazio (Rome), Abruzzo (L'Aquila), Molise (Campobasso), Campania (Naples), Apulia (Bari), Basilicata (Potenza), Calabria

(Catanzaro), Sicily (Palermo), Sardinia (Cagliari). Figure E1 shows the regional capitals considered in CPI computations for 2014.

The Italian regions vary widely in terms of population, surface area and socio-economic characteristics. Italy has a dualistic economy with all Southern regions attaining a lower level of per capita income on average than the Centre-Northern regions. Compared to other OECD countries, there is a broad spectrum of regional differences in income and jobs in Italy: in 2013, 50% of households in the South and the Islands earned less than 20,188 euro (approximately 1,682 euro per month), while in terms of national averages, half of all households reported a net income below or equal to 24,310 euro per year (approximately 2,026 euro per month).

Figure E1. Italian regional capitals



In 2014, unemployment rates ranged from 3.9% in Trentino Alto-Adige to 15.4% in Calabria. In the same year, the South and the Islands were the areas of the country with the highest risk of poverty or social exclusion, affecting a little less than half the population. The worst poverty rates were observed in Calabria and Basilicata (almost a third of households), while the lowest rate was found in Trentino-Alto Adige. In 2015, households in Northern Italy spent more than households in the South and in the Islands. In fact, the highest expenditure was observed in the North-West (2,836.32 euros per month), approximately 1,000 euros more than the average expenditure in the Islands (1,891.78 per month).

Regarding consumer price differences across regions, it is worth noting that Italy is one of the few countries that have carried out official experimental sub-national PPP computations. For many years, Istat has focused on the issue of comparing consumer prices across the various geographical areas and in particular across the 20 Italian regions (Istat, 2008; 2010). The results of the latest experiment carried out by Istat in 2009 showed significant differences in the level of consumer prices across the regional capitals (Istat, 2010). Consumer price levels in the Northern cities are generally higher than those in the Centre and especially in Southern Italy. Bolzano (105.6) and Milano (104.7) showed the highest prices compared with the Italian average (100) while the least expensive city proved to be Napoli (93.8).

The sub-national PPP results obtained from these analyses have encouraged Istat to go ahead with the project of regularly producing spatial indices of consumer prices at regional level.

Appendix F: Validating the kriging estimation for Campobasso

Following Montero et al. (2015), in the kriging process, a valid semivariogram model has been chosen and used, and some other assumptions have been made (for example about the

stationarity of the stochastic process behind the data). However, these assumptions should be validated, otherwise, the results obtained from the kriging process could lead to erroneous conclusions.

With the use of statistical tests being discarded for the reasons given in Montero et al. (2015), the most widely-used procedure, especially with small data sets, is to perform a cross-validation (CV), or leave-one-out process.

The CV process consists of:

(i) Obtaining the kriging prediction $\hat{X}(\mathbf{s}_0)$ at each sample point (in our case, region) $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_R$, as if the sample values at these points were unknown, from the observations at the $R-1$ remaining points (or from a set of neighbouring observations, as in the case of universal kriging). The prediction variance at each sample point, $\hat{\sigma}^2(\mathbf{s}_i)$, is also calculated.

(ii) Calculating the following diagnostic statistics from the results obtained in (i):

- The mean prediction error: $ME = \frac{1}{R}\sum_{i=1}^{R}\left(X(\mathbf{s}_i) - \hat{X}(\mathbf{s}_i)\right)$
- The mean squared prediction error: $MSE = \frac{1}{R}\sum_{i=1}^{R}\left(X(\mathbf{s}_i) - \hat{X}(\mathbf{s}_i)\right)^2$
- The mean squared standardized prediction error: $MSSDE = \frac{1}{R}\sum_{i=1}^{R}\left(\frac{X(\mathbf{s}_i) - \hat{X}(\mathbf{s}_i)}{\hat{\sigma}(\mathbf{s}_i))}\right)^2$

If the assumptions made, especially the semivariogram choice, are valid, ME should be approximately 0, which is indicative of non-systematic prediction errors (anyway, regardless of which semivariogram is chosen, kriging predictions are unbiased and ME is expected to tend to 0), MSE should be small, and MSSDE should be approximately 1 (which indicates that prediction errors are compatible with the corresponding kriging prediction variances).

Obviously, results obtained from the application of a kriging procedure meeting the above requirements (especially the third one), are expected to be highly accurate (relative to the corresponding kriging variance).

Based on the PPPs obtained from the SP-CPD for the 18 regional capitals with available price data and the 7 BHs considered, a CV procedure was performed using universal kriging with linear trend in the coordinates, a selection of neighbouring PPPs (not all of them) and a Gaussian semivariogram. The diagnostic descriptive statistics obtained from this CV procedure are shown in Table F1.

Table F1. **Kriging validation. Cross-validation results for PPPs provided by the SP-CPD model**

	Beef and veal	Other meats and edible offal	Pork	Lamb, mutton and goat	Fresh or chilled fruit	Fresh or chilled vegetables	Fresh, chilled or frozen fish and seafood
ME	-0.865	0.008	-0.277	-0.731	0.128	0.120	-0.218
MSE	45.768	14.378	11.286	35.836	55.478	91.115	133.745
MSSDE	1.017	1.002	1.024	1.287	0.950	0.964	0.986

As can be seen, ME, MSE and MSSDE meet the requisites for the kriging procedure performed to provide good estimations and, in this sense, it can be stated that the kriging PPP estimate for Campobasso is highly reliable. It is important to note that MSE for Fresh or chilled vegetables and Fresh, chilled or frozen fish and seafood is not as good as for the other BHs. However, even in these two cases (where the square root of MSE is around 10%) the results of the diagnostic descriptive statistics can be considered very good because MSE does not take into account the "isolation effect" resulting from the fact that only 18 PPPs are provided by SP-CPD. If the isolation effect is considered, or, in other words, if the focus moves from MSE to MSSDE, the results are excellent for Beef and veal, Other meats and edible offal and Pork,

roughly as good as for Fresh or chilled vegetables and Fresh, chilled or frozen fish and seafood, and very good for Lamb, mutton and goat.

Finally, a word of caution. As stated in Cressie (1993), CV cannot prove that the correct semivariogram model has been selected (or that other assumptions made are valid). It can only confirm that the assumptions are not incorrect (there is no reason for rejecting them).

Appendix G: Estimates of sub-national PPPs using SEM-CPD and CPD hedonic models

Table G1. Estimates of sub-national PPPs for the 20 Italian regional capitals (Rome = 100) using SEM-CPD and CPD hedonic models: *Beef and Veal and Other meats and edible offal*

	Beef and Veal				Other meats and edible offal			
	SEM-CPD		CPD		SEM-CPD		CPD	
	PPPs	S.E.	PPPs	S.E.	PPPs	S.E.	PPPs	S.E.
North								
Aosta	106.11	8.70	106.11	9.36	109.82	10.34	109.82	12.33
Torino	112.88*	8.83	112.88*	9.49	105.91	8.81	105.91	10.5
Genova	109.97	8.35	109.97	8.98	98.54	8.56	98.54	10.22
Milano	87.17**	6.62	87.17**	7.12	105.49	9.31	105.49	11.1
Trento	105.44	8.00	105.44	8.61	106.09	8.52	106.09	10.16
Venezia	104.50	8.57	104.5	9.22	98.88	8.59	98.88	10.25
Trieste	102.51	8.02	102.51	8.62	123.31**	10.72	123.31**	12.78
Bologna	99.92	8.19	99.92	8.81	104.13	9.05	104.13	10.8
Centre								
Firenze	87.76**	7.20	87.76*	7.74	92.51	8.04	92.51	9.59
Ancona	107.31	8.15	107.31	8.76	103.22	9.11	103.22	10.87
Perugia	101.62	7.82	101.62	8.41	87.28**	7.68	87.28*	9.16
South and Islands								
L'Aquila	100.66	7.64	100.66	8.22	113.95	11.00	113.95	13.12
Campobasso	91.12**	7.62	91.12***	7.62	97.62	7.62	97.62	7.62
Napoli	81.18***	6.66	81.18***	7.16	92.63	8.05	92.63	9.6
Potenza	77.92***	6.39	77.92**	6.87	86.32**	7.50	86.32*	8.95
Bari	89.13*	7.91	89.13*	8.51	81.61**	8.78	81.61**	10.47
Catanzaro	78.77***	6.46	78.77***	6.95	85.23*	9.16	85.23*	10.93
Palermo	87.10**	7.14	87.10**	7.68	90.38	7.85	90.38	9.37
Cagliari	86.14**	7.06	86.14**	7.60	87.51*	7.61	87.51*	9.07
κ	0.000				0.000			
Obs.	177		177		74		74	
RMSE	0.164		0.164		0.123		0.123	
AIC	-586		-588		-260.3		-262.3	
DF	26		25		24		23	

* 10 %, ** 5 %, *** 1 %; S.E. denotes standard error; PPP for Campobasso has been estimated with kriging

S.E. has been computed using the Delta method (Held and Sabanes Bove, 2014)

RMSE: Squared root of the mean squared error; AIC: Akaike information criterion; DF: Degrees of freedom

Table G2. Estimates of sub-national PPPs for the 20 Italian regional capitals (Rome = 100) using SEM-CPD and CPD hedonic models: *Pork and Other meats and edible offal*

	Pork				Lamb, mutton and goat			
	SEM-CPD		CPD		SEM-CPD		CPD	
	PPPs	S.E.	PPPs	S.E.	PPPs	S.E.	PPPs	S.E.
North								
Aosta	111.10	10.07	111.1	11.6	95.59	7.11	95.59	10.06
Torino	100.09	8.40	100.09	9.68	104.73	7.79	104.73	11.02
Genova	93.61	8.48	93.61	9.77	95.52	7.11	95.52	10.05
Milano	96.28	8.08	96.28	9.31	97.03	6.75	97.03	9.55
Trento	87.65**	7.36	87.65*	8.48	100.35	6.98	100.35	9.88
Venezia	94.91	8.60	94.91	9.91	108.93	8.1	108.93	11.46
Trieste	98.14	8.24	98.14	9.50	116.89**	8.7	116.89*	12.30
Bologna	102.96	9.33	102.96	10.75	101.29	7.53	101.29	10.66
Centre								
Firenze	87.75*	7.95	87.75*	9.16	97.62	7.26	97.62	10.27
Ancona	105.33	8.84	105.33	10.19	109.99*	7.65	109.99	10.82
Perugia	95.88	8.29	95.88	9.55	110.45	8.22	110.45	11.62
South and Islands								
L'Aquila	100.14	8.41	100.14	9.69	98.37	6.84	98.37	9.68
Campobasso	96.34	7.99	96.34	7.99	96.45	7.99	96.45	7.99
Napoli	87.14*	7.89	87.14*	9.10	87.21**	6.49	87.21*	9.17
Potenza	87.45*	7.92	87.45*	9.13	88.34**	6.57	88.34*	9.29
Bari	96.38	8.73	96.38	10.06	101.77	7.57	101.77	10.71
Catanzaro	82.38***	7.46	82.38**	8.60	76.50***	5.69	76.5***	8.05
Palermo	85.08**	7.71	85.08**	8.88	80.07***	5.96	80.07***	8.42
Cagliari	86.89**	7.87	86.89*	9.07	74.54***	5.54	74.54***	7.84
κ	0.000				0.000			
Obs.	89		89		42		42	
RMSE	0.128		0.128		0.074		0.074	
AIC	-315.8		317.8		-170.3		-172.3	
DF	24		23		23		22	

* 10 %, ** 5 %, *** 1 %; S.E. denotes standard error; PPP for Campobasso has been estimated with kriging

S.E. has been computed using the Delta method (Held and Sabanes Bove, 2014)

RMSE: Squared root of the mean squared error; AIC: Akaike information criterion; DF: Degrees of freedom

Table G3. Estimates of sub-national PPPs for the 20 Italian regional capitals (Rome = 100) using SEM-CPD and CPD hedonic models: *Fresh or chilled fruit and Fresh or chilled vegetables*

	Fresh or chilled fruit				Fresh or chilled vegetables			
	SEM-CPD		CPD		SEM-CPD		CPD	
	PPPs	S.E.	PPPs	S.E.	PPPs	S.E.	PPPs	S.E.
North								
Aosta	124.96***	3.70	124.96***	3.81	124.97***	4.18	124.97***	4.30
Torino	102.66	2.16	102.66	2.22	100.41	2.49	100.41	2.56
Genova	113.83***	2.92	113.83***	3.00	110.76***	3.45	110.76***	3.54
Milano	151.04***	3.17	151.04***	3.27	157.75***	3.99	157.75***	4.1
Trento	125.63***	3.33	125.63***	3.42	119.35***	3.65	119.35***	3.76
Venezia	123.03***	3.46	123.03***	3.56	115.81***	3.7	115.81***	3.80
Trieste	121.47***	3.04	121.47***	3.13	121.08***	3.61	121.08***	3.72
Bologna	127.23***	3.27	127.23***	3.37	123.45***	3.69	123.45***	3.79
Centre								
Firenze	108.73***	2.68	108.73***	2.75	97.96	2.82	97.96	2.9
Ancona	122.40***	3.69	122.4***	3.79	111.97***	3.94	111.97***	4.05
Perugia	111.18***	3.02	111.18***	3.29	101.27	3.42	101.27	3.51
South and Islands								
L'Aquila	92.19***	2.25	92.19***	2.31	87.05***	2.61	87.05***	2.68
Campobasso	93.82**	2.61	93.82**	2.61	97.01	2.68	97.01	2.68
Napoli	96.06*	2.41	96.06**	2.48	85.73***	2.7	85.73***	2.77
Potenza	99.40	2.70	99.40	2.78	96.4	2.79	96.4	2.87
Bari	89.14***	2.26	89.14***	2.33	84.06***	2.6	84.06***	2.67
Catanzaro	81.80***	1.93	81.80***	1.98	82.18***	2.21	82.18***	2.27
Palermo	101.86	2.24	101.86	2.31	102.5	2.64	102.5	2.71
Cagliari	104.15**	2.71	104.15**	2.79	97.97	3.09	97.97	3.17
κ	0.000				0.000			
Obs.	1,673		1,673		2,018		2,018	
RMSE	0.175		0.175		0.231		0.231	
AIC	-5,632.9		-5,634.9		-5,690.5		-5,692.5	
DF	95		94		112		111	

* 10 %, ** 5 %, *** 1 %; S.E. denotes standard error; PPP for Campobasso has been estimated with kriging

S.E. has been computed using the Delta method (Held and Sabanes Bove, 2014)

RMSE: Squared root of the mean squared error; AIC: Akaike information criterion; DF: Degrees of freedom

Table G4. Estimates of sub-national PPPs for the 20 Italian regional capitals (Rome = 100) using SEM-CPD and CPD hedonic models: *Fresh, chilled or frozen fish and seafood*

	Fresh, chilled or frozen fish and seafood			
	SEM-CPD		CPD	
	PPPs	S.E.	PPPs	S.E.
North				
Aosta	93.67**	3.75	93.67**	3.85
Torino	100.67	3.76	100.67	3.87
Genova	97.54	3.76	97.54	3.86
Milano	125.19***	4.38	125.19***	4.5
Trento	92.43***	4.2	92.43***	4.32
Venezia	76.13***	2.77	76.13***	2.85
Trieste	91.07***	3.36	91.07***	3.45
Bologna	82.49***	3.16	82.49***	3.25
Centre				
Firenze	97.58	3.55	97.58	3.65
Ancona	86.28***	3.45	86.28***	3.55
Perugia	111.15***	4.09	111.15***	4.21
South and Islands				
L'Aquila	85.2***	3.87	85.2***	3.98
Campobasso	94,25*	2.68	94.25*	2.68
Napoli	76.91***	3.19	76.91***	3.28
Potenza	74.4***	2.55	74.4***	2.63
Bari	74.65***	2.8	74.65***	2.88
Catanzaro	79.68***	2.83	79.68***	2.92
Palermo	104.09	4.36	104.09	3.28
Cagliari	91.48***	3.66	74.4***	2.63
κ	0.000			
Obs.	888		888	
RMSE	0.189		0.189	
AIC	-2,854.6		-2,856.6	
DF	51		50	

* 10 %, ** 5 %, *** 1 %; S.E. denotes standard error; PPP for Campobasso has been estimated with kriging

S.E. has been computed using the Delta method (Held and Sabanes Bove, 2014)

RMSE: Squared root of the mean squared error; AIC: Akaike information criterion; DF: Degrees of freedom

Table G5. SEM-PD: Spatial autoregressive parameter and goodness of fit

	Beef and Veal	Other meats and edible offal	Pork	Lamb, mutton and goat	Fresh or chilled fruit	Fresh or chilled vegetables	Fresh, chilled or frozen fish and seafood
κ	0.609	0.629	0.552	0.707	0.656	0.557	0.610
Obs.	177	74	89	42	1,673	2,018	888
RMSE	0.187	0.147	0.142	0.110	0.219	0.275	0.236
AIC	-575.7	-269.9	-332.9	-173.8	-4,918.2	-5,027.6	-,.2
DF	8	6	6	5	77	94	33

RMSE: Squared root of the mean squared error; AIC: Akaike information criterion; DF: Degrees of freedom

Figure G1. Estimates of sub-national PPPs for the 20 Italian regional capitals (Rome = 100) using CPD hedonic model.

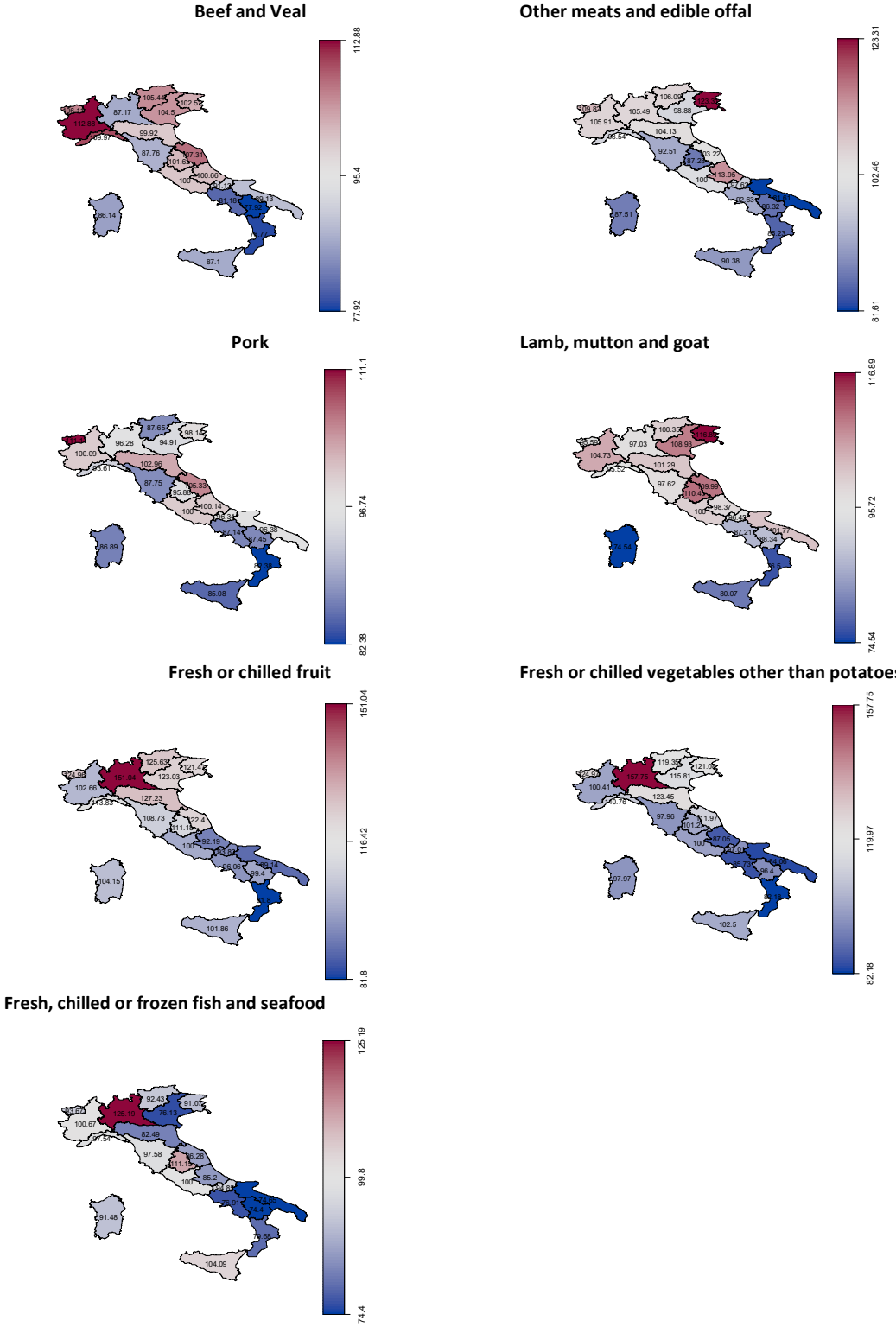
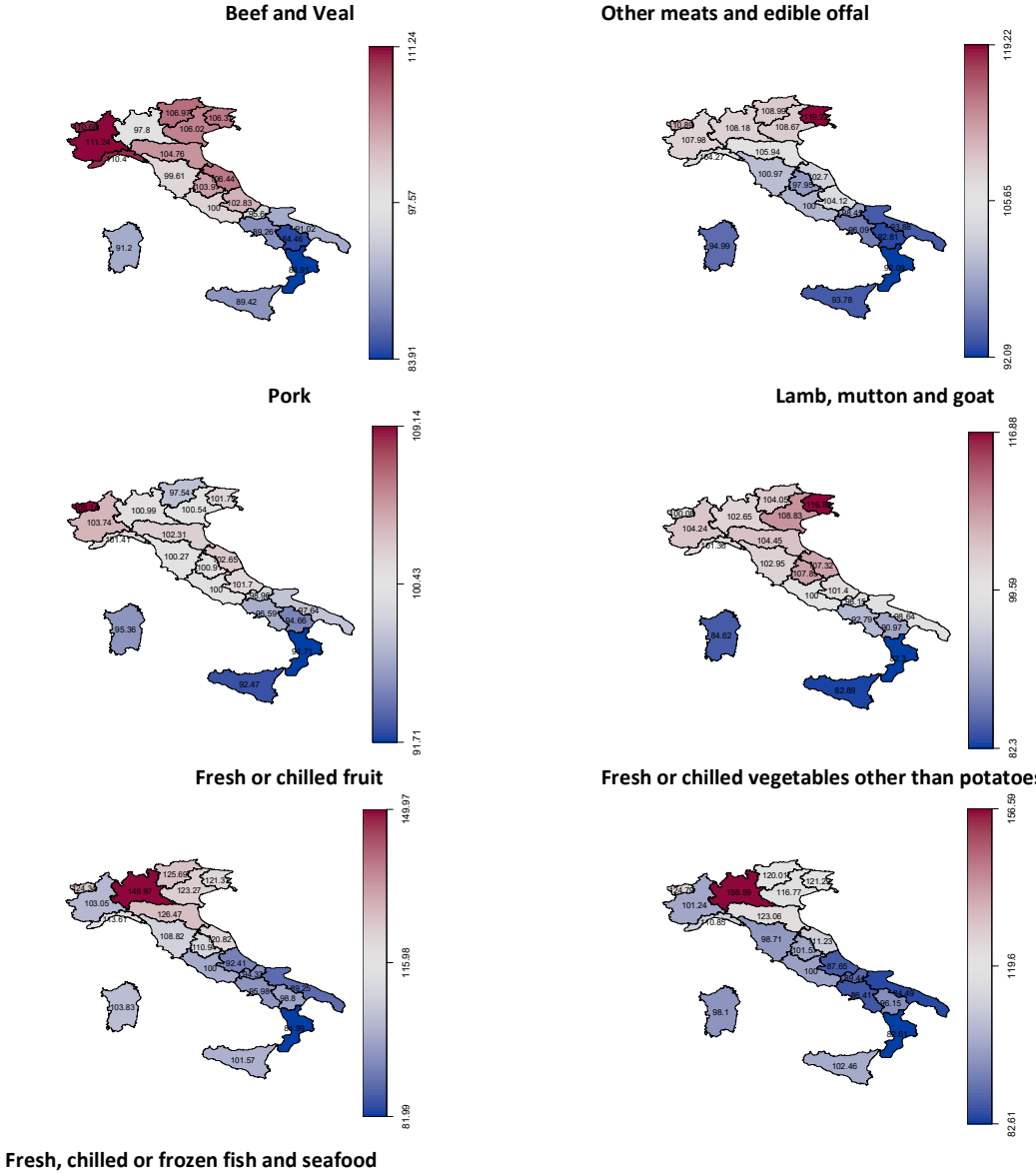
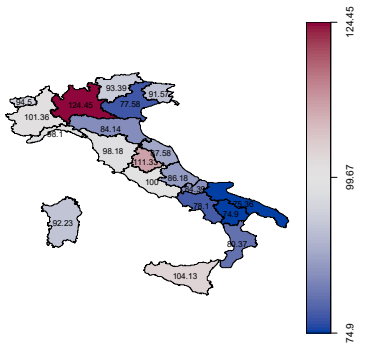


Figure G2. Estimates of sub-national PPPs for the 20 Italian regional capitals (Rome = 100) using SP-CPD hedonic model





Online supporting information: References

Anselin, L., Spatial econometrics: Methods and models. Kluwer: Boston, 1988.

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2015). Bayes-X software for Bayesian inference in structured additive regression models. Version 3.0. Available at http//www.bayesx.org.

Biggeri, L., Laureti, T., and Polidoro, F., "Computing sub-national PPPs with CPI data: an empirical analysis on Italian data using country product dummy models", Social Indicators Research, 131(1), 93-121, 2017.

Cressie, N. A. C. (1993) *Statistics for Spatial Data*. Chichester: Wiley.

El-Shaarawi, A. H. and Viveros, R (1997) Inference about the Mean in Log-Regression with Environmental Applications. Environmetrics, 8, 569–582.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. Biometrika, 61(2), 383–385.

Istat (2008). Le differenze nel livello dei prezzi tra i capoluoghi delle regioni italiane per alcune tipologie di beni Anno 2006, Roma: ISTAT (in Italian).

Istat (2010). La differenza nel livello dei prezzi al consumo tra i capoluoghi delle regioni italiane, Anno 2009, Roma: ISTAT(in Italian).

Montero, J. M., Fernández-Avilés, G., and Mateu, J. (2015) *Spatial and spatio-temporal kriging and modelling*. Chichester: Wiley.

Montero, J. M., Mínguez, R., and Durbán, M., "SAR models with nonparametric spatial trends. A P-spline approach", *Estadística Española*, 54(177), 89–111, 2012.

Umlauf, N., Adler, D., Kneib, T., Lang, S., and Zeileis, A. (2015). Structured Additive Regression Models: An R Interface to BayesX. Journal of Statistical Software, 63(21), 1–46 (Available from http://www.jstatsoft.org/v63/i21/).