Temperature interpolation by local information: The example of France

Daniel Joly, Thierry Brossard, Hervé Cardot, Jean Cavailhès, Mohamed Hilal, Pierre Wavresky

Temperature interpolation based on local information: the example of France

Daniel Joly,[a]* Thierry Brossard,[a] Hervé Cardot,[b] Jean Cavailhes,[c] Mohamed Hilal[c] and Pierre Wavresky[c]

[a] *Laboratoire ThéMA, CNRS and Université de Franche-Comté, Besançon, France*
[b] *Institut de Mathématiques de Bourgogne, Université de Bourgogne, Dijon, France*
[c] *Laboratoire CESAER, INRA, Dijon, France*

ABSTRACT: Methods of interpolation, whether based on regressions or on kriging, are global methods in which all the available data for a given study area are used. But the quality of results is affected when the study area is spatially very heterogeneous. To overcome this difficulty, a method of local interpolation is proposed and tested here with temperature in France. Starting from a set of weather stations spread across the country and digitized as 250 m-sided cells, the method consists in modelling local spatial variations in temperature by considering each point of the grid and the n weather stations that are its nearest neighbours. The procedure entails a series of steps: recognition of the n stations closest to the cell to be evaluated and subdivision of the study area into polygons defined by a neighbourhood rule, elaboration of a local model by multiple regression for each polygon, and application of the parameter estimate from the regression to obtain a predicted value of temperature at each point of the polygon under consideration.

These results are compared with results from three global interpolation methods: (1) regression, (2) ordinary kriging, and (3) regression with kriging of residuals. We then develop the original results from local interpolation such as mapping of the coefficients of determination and of the parameter estimate related to altitude and to distance to the sea. These developments highlight the processes that dictate the spatial variation of climate.

KEY WORDS interpolation; temperature; France

1. Introduction

Interpolation is a way of reconstructing continuous fields from variables measured at point locations. This is no straightforward operation and one of the main difficulties is to select the method that provides the best estimates. Two families of methods have come to stand out for the quality of their results: the methods of kriging and regression. Given the statistical constraints associated with them, these methods are not interchangeable and do not yield optimal results in all cases. Kriging is better suited when variables are strongly spatially autocorrelated, where, for temperatures, say, a gentle topography engenders regular thermal gradients. Conversely, regression yields better results where, again for the example of temperatures, their spatial variation is dictated by prominent relief. The two methods may be concatenated and results are often improved by kriging the residuals of a regression. The criteria for choosing from among these possibilities are not always obvious even when the geographical sectors in question appear homogeneous.

Matters are further complicated where plains and mountains lie side by side over areas of some size (Joly *et al*., 2010). This is the case in a multidisciplinary research project to estimate the 'price of climate' across France. Continuous climatic information across the entire country was required so that economic and climatic data could be matched. As climatic data are sporadic, they had to be interpolated by relating the response variables, those for climate, and the explanatory variables (latitude, longitude, and environmental data on relief and land cover). To investigate this issue, we tested three global methods of interpolation: regression, kriging, and regression followed by kriging of residuals. Then we compared the findings with results from a fourth method based on local interpolation. The results of the experiment are presented here.

Section 2 describes the data used. Section 3 presents the main features of the four methods. The principle behind the local interpolation method is described in detail from a set of data on sunshine duration; this variable was chosen because it was recorded at just 111 weather stations, thus simplifying our exposition. In Section 4, the entire approach is applied to temperature measured at 651 stations to compare the findings from all four methods. The standard deviation of residuals is used

* Correspondence to: Daniel Joly, ThéMA, CNRS, Université de Franche-Comté, Besançon, France. E-mail: daniel.joly@univ-fcomte.fr

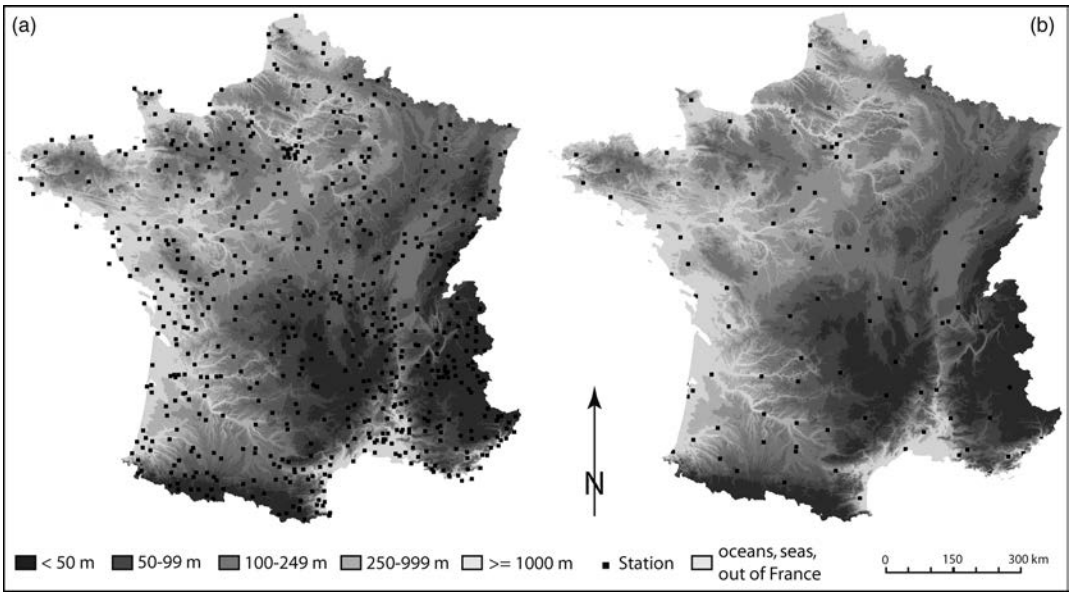

Figure 1. (a) Location of the 651 weather stations in France recording temperatures. (b) Location of the 111 stations recording duration-of-sunshine.

for this. Section 5 relates to the specific developments of the local interpolation method, i.e. the mapping of the coefficient of determination, the Pearson correlation coefficient, and the parameter estimate for each of the explanatory variables included in a geographical information system (GIS). This additional information is a fundamental contribution to climatology that is specifically interested in the study of local climate. This provides insight into the factors behind the spatial distribution of climatic phenomena.

2. Study area and data

2.1. Climatic data

The temperature data were taken from Météo-France, the national meteorological office in charge of the observation network countrywide. 'Normal' temperatures were computed on the basis of records from 1971 to 2000. The methodological tests presented here relate solely to average monthly temperatures collected at 651 stations (Figure 1). The 111 stations of the network that recorded duration-of-sunshine are used in Section 3 only (for describing the local interpolation method).

2.2. Environmental data

2.2.1. Acquisition

Environmental data are used as explanatory variables in modelling (Arnaud and Emery, 2000). The operation requires the formation of a spatially referenced database. To this end, two sources of information were mobilized to produce the data required in the raster format at 250 m resolution.

Land cover information was taken from the European Corine Land Cover database whose initial vectorial data were rasterized. Additional information was derived from this source. First, a vegetation index was constructed from land cover types to which a standard index value was attributed (5 for densely built, city centres, airports, etc.; 250 for compact forest). This index provided an approximation for the abundance of biomass in the vicinity of the points in question (Joly, 2007). Next, distances to the main types of land cover (distance to forest, distance to nearest ocean or sea in a logarithm form) were calculated.

Topographic information was taken from the digital elevation model (DEM) distributed by the *Institut Géographique National* (IGN). New variables can be derived from this source by procedures based on cartographic algebra and trigonometry. In addition to altitude (Figure 2), these included slope angle, slope orientation, topographic ruggedness indicative of irregular relief (it may be zero for flat land or slopes that form perfect straight lines), enclosure or exposure index (a narrow valley bottom takes a negative value, whereas a high point has a positive value), and theoretical solar radiation calculated for the equinox with allowance for topographic masks up to 5 km around each point.

All told, the base was made up of three layers derived from Corine Land Cover and six from the DEM, all of these variables being candidates to explain the spatial variation of temperature. The whole of the area analysed contained 8 704 283 cells.

2.2.2. Collinearity processing

Some explanatory variables displayed marked covariation. The plainest examples were the Pearson correlation coefficients between altitude and ruggedness (0.62), altitude and slope angle (0.69), altitude and vegetation index (0.38), and, to a lesser extent, altitude and distance to the sea (0.23). This would not have mattered had our objective been solely to estimate temperatures so as to

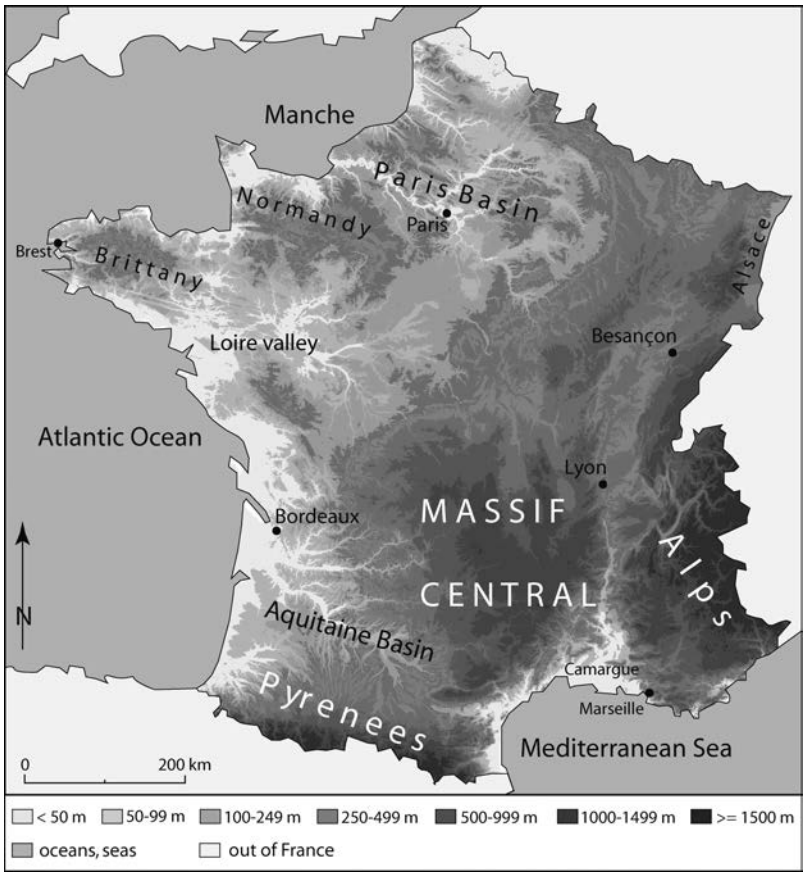

Figure 2. Study area.

interpolate them. However, the occurrence of collinearity may affect estimations of the model's parameters (Gunst, 1983). When multicollinearity exists, the variances of some of the estimated regression coefficients may become very large, leading to unstable and potentially misleading estimates of the regression equation.

Statistical collinearity therefore possesses serious difficulties for interpreting results. For example, the fact that collinearity may change the sign of a regression coefficient is troublesome when studying the eigen-effect of a variable X_j on Y. Insofar as we were looking to compare the respective influences of several variables on monthly temperatures through Pearson correlation coefficients, it seemed preferable to eliminate the influence of altitude on the foregoing variables. There are several ways to limit the collinearity of explanatory variables:

- Partial least squares (PLS) regression: a series of estimators is obtained by considering residuals to be a new dependent variable (Wold *et al.*, 1984; Helland, 1990).
- Stepwise regressions (Hocking, 1976): by limiting the number of explanatory variables depending on their partial correlation coefficients with the response variable, any collinearities are reduced.

We chose the first way to eliminate collinearity between altitude and most of the other explanatory variables. Note that the new latent variables obtained with the PLS procedure are linear combinations of the initial variables. Consequently, one can easily write the coefficients of regression in terms of the original variables that have been selected in the PLS procedure (Gunst, 1983). After applying the resulting transformation, one final collinearity remained between slope angle and ruggedness ($r = 0.49$). This should be kept in mind when interpreting results later. All these factors may generate bias that is hard to control for. The stepwise method for selecting significant explanatory variables included in the regression model also tends to reduce collinearity.

3. Interpolation methods

Remember that climatic observation data are collected by weather stations and by regions. They are plotted in a two-dimensional space (altitude is considered as an attribute of the pixels and not as a dimension). When integrated in a GIS, these data are characterized by geographical attributes known for the whole territory and stored as layers of data (Mitas and Mitasova, 1999).

3.1. Global regression and kriging

By global we mean that the calculations pertain to all available stations. The first two interpolation methods we use are a statistical method based on regressions (Cressie, 1993; Joly *et al.*, 2003), and a probabilistic method, ordinary kriging (Matheron, 1970; Courault and

Monestiez, 1999; Baillargeon, 2005). It is worth going over their respective advantages and drawbacks.

3.1.1. Regression

In short, the linear model mathematically expresses the relation between a statistical variable, called the response or dependent variable Y (each of the 12 monthly temperature values), and p explanatory or independent variables $X_1, \ldots, X_p$ (the environmental variables recorded in the GIS: altitude, slope angle, etc.). We denote n as the number of data samples considered, y_i the ith observation of variable Y, and x_i^j as that of variable X_j. To simplify, we assume that these variables are centred and reduced:

$$\sum_{i=1}^{n} y_i = 0 = n$$

$$\forall j = 1, \ldots, p \qquad \sum_{i=1}^{n} x_i^j = 0 \qquad \sum_{i=1}^{n} \left(x_i^j\right)^2 = n \quad (1)$$

The linear model is defined by the equation:

$$\boldsymbol{Y} = X\beta + \varepsilon \quad (2)$$

in which

- $\boldsymbol{Y}$ is the vector $(y_1, y_2, \ldots, y_n)$ of the n observed values of the response variable Y.
- $\boldsymbol{X}$ is the data matrix with n rows and p columns (from 1 to p) being defined by the vectors j $(x_1^j, x_2^j, \ldots, x_n^j)$.
- $\beta = (\beta_1, \ldots, \beta_p)$ is the vector of regression coefficients.
- $\varepsilon$is the vector of residuals $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ defined by an independent sample of residual variable ε of variance σ^2.

The regressions disregard distance between stations. This is why they are effective above all in heterogeneous sectors where large deviations may occur over short distances. Now, in homogeneous sectors, distance has great explanatory power because both dependant and explanatory variables are characterized by regular gradients. In this case, it is kriging that should be preferred.

3.1.2. Kriging

Interpolation by kriging is underpinned by semi-variogram analysis. This is an unbiased, optimal linear estimation method using the structural properties of the semi-variogram to determine whether the distribution of the parameter(s) under study is regionalized (i.e. has a spatial structure), random, or periodic. The semi-variogram theoretical model involves semi-variance $\gamma(h)$, which is a function of the sampling interval (h); the equation for the n (h) points a_i and a_j, which are $h = |a_i - a_j|$ apart, is given by:

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} (a_i - a_j)^2 \quad (3)$$

The use of the semi-variogram is supported by the hypothesis stating that mathematical expectation exists and takes the same value at all points a, that the covariance function is finite, and that it is a function solely of distance h between observations (second-order stationarity). We used a linear function to adjust the semi-variogram.

3.1.3. Limits of regression and kriging

The two methods being complementary, it is a worthwhile solution to associate them in a single procedure: regression of climatic variables on environmental variables and then kriging of the residuals thus obtained. The process is analogous to kriging with external drift (Goovaerts, 1997; Wackernagel, 2003).

When applied globally, regression and kriging yield good results when the analysis is confined to a climatologically consistent area. However, results are poorer when interpolation is over far larger and heterogeneous zones like the whole of France. The technique works but the statistics are disrupted by discordant constraints. Apart from the contrast between plains and mountains, the space is subjected to separate climatic systems. Now, these systems operate, if not autonomously, at least largely independent of each other, with the result that the processes causing spatial variations of climate do not function in the same way everywhere. This is why, as the explanatory spatial factors differ from one system to another, the 'global' statistic produces a 'scrambled' general model that is not really satisfactory anywhere.

3.2. Local regression

One way to overcome this difficulty is to address interpolation locally as in the 'local regression' method (Cleveland and Devlin, 1988), also known as 'kernel regression' (Wand and Jones, 1995). Local regression consists in modelling the variable of interest using polynomials whose explanatory variables are the station coordinates (Baillargeon, 2005). This is much like calculating local trend surfaces together with data weighting by distance; accordingly, Fotheringham *et al*. (2002) named the method 'geographically weighted regression' (Loader, 2004).

From local regression, our method (Joly *et al*., 2008) uses the principle of processing by proximity of data but it differs in the choice of explanatory variables because these pertain to the geographical setting (altitude, slope angle, vegetation, etc.). Therefore, weighting by distance is not required. Given its aim and to differentiate it from 'local regression', the method proposed is termed 'local interpolation'. It involves the following three stages:

- Identification of the n stations closest to the estimation point and division of the territory into polygons bounded by a neighbourhood rule
- Analysis by multiple regression for each polygon
- Application of coefficients to the cells (or pixels) making up each polygon (interpolation)

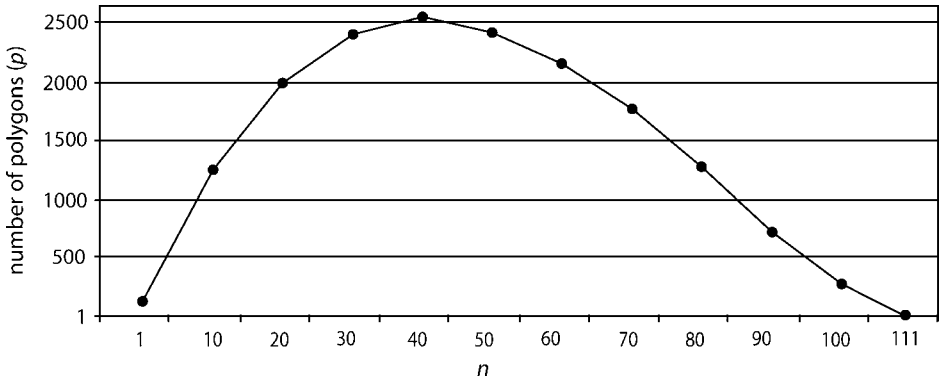

Figure 3. Number of polygons p by parameter n for 111 duration-of-sunshine recording stations.

There are two possible ways to define the neighbourhood. The first is based on the criterion of distance: all the stations within a fixed perimeter (10, 50, 100 km, etc.) are clustered around the point to be interpolated. In such cases, the number of stations varies with the density of the network. This becomes troublesome for further statistical analyses when, the set distance being too small and the network being too loose-knit, there are too few stations within the relevant area. This is what happens in our application where, even with a 50-km radius, there are instances where only four stations are taken into account.

An alternative approach is to set the number of stations and look for 20, 30, 100, n stations around the cell to be interpolated. In this case, it is the surface area within which the n stations are located that varies with the density of the network. In our application, an extreme case shows one must go up to 88 km to enclose 20 stations and up to 178 km to enclose 100 stations. Despite this drawback, it is this alternative approach that was chosen because it ensures a sufficient statistical basis for calculation in all cases.

3.2.1. Choice of parameter n and grid spacing

Parameter n determines the area over which information is to be collected to solve the regressions:

- Where n is low (say, $n = 20$), the area over which stations are recruited is small. The advantage is that the 20 stations being close to each other, they are climatologically highly consistent. The downside, though, is that the statistic is unreliable. The significance levels are quite low and correlatively the number of explanatory variables picked out in the regression equation may well be low.
- Where n is high (say, 100), the statistic is reinforced but the area over which stations are recruited becomes very broad. This means there is a greater likelihood of it including different climatic zones, which may be a problem for the coherence of the models.

3.2.2. Definition of polygons

In the procedure for defining neighbourhoods and in view of the density of stations, it is very likely that two neighbouring cells are associated with the same n stations. It follows that the clustering within the same spatial unit (a polygon) of all cells connected to the same n stations is an advantage. Rather than needlessly repeating the same regression calculations for all the cells making up a polygon, they may be performed once per polygon, the associated coefficients and constants being valid for all the cells belonging to the same spatial unit.

To address this issue, we take the example of the network that records duration-of-sunshine, as there are fewer stations (111) than for temperature records and the demonstration is easier to follow. With $n = 1$, each of the 111 polygons (1 per station) covers a mean area of 4900 km^2 (within a circle of radius 40 km). But when n increases, the outline of the polygons changes and their number, p, increases. With $n = 2$, p shifts to 290, then to 1273 for $n = 10$, and so on, as depicted in Figure 3. The number of polygons p peaks at $n = 41$ ($p = 2545$), and then tails off to the end of the process with $n = 111$ ($p = 1$).

3.2.3. Polygon pattern

Figure 4 shows that the division of the territory is akin to Thiessen polygons (De Berg *et al*., 2000) when $n = 1$. Each polygon has four to six faces. Even if they are of similar sizes, differences arise locally because of deviations in regional density and the uneven location of weather stations throughout France. With $n = 20$, the position changes. The peripheral polygons expand while the central ones shrink. This contrast is even starker with $n = 41$ and continues when $n = 100$. A hundred or so very small central polygons stand against 150 sometimes enormous ones covering the remainder of the country. This transformation of polygons can be explained by an edge effect. The search for nearby stations is in one direction only when on the periphery of the space, as available stations are all located towards the interior.

3.2.4. Variation of the number of polygons with the measurement network density

Table I shows how the number of polygons varies with n and with the number of stations making up the network in question (duration-of-sunshine and temperatures, with

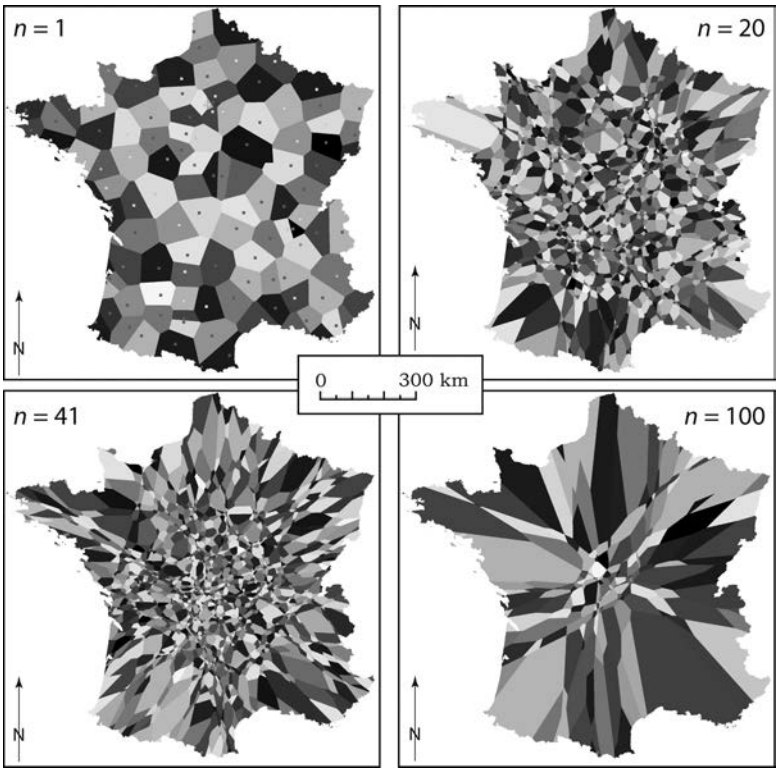

Figure 4. Grid patterns obtained for four values of n.

Table I. Comparison of the number of polygons calculated for four different n values and applied to the network of stations recording duration-of-sunshine and temperature.

	$n = 1$	$n = 20$	$n = 30$	$n = 50$	$n = 100$
Sunshine network	111	2025	2407	2428	264
Temperature network	651	18 225	25 618	38 121	60 513

111 and 651 stations, respectively). As the number of polygons affects computation time, n should be adjusted carefully. For this reason, we now turn to temperatures for which the measurement network is denser and engenders a much larger number of polygons for processing.

3.2.5. Regression analyses

The choice of n is constrained by statistical considerations. Insofar as the climatic variable (temperature in our case) is explained by a potential of nine explanatory variables (three from Corine Land Cover and six from the DEM), n should not fall below a certain value that would make the results of the regressions non-significant. Even if significance testing can evaluate the relevance of the variables used in the case of small samples, it is accepted that one should avoid having too few individuals in regression calculations. Considering all of the constraints on implementing the procedure (grid spacing, computation time, proportion of stations common to two adjoining polygons), we were careful not to take n below 20 or to set it too high, even if this is favourable from a strictly statistical viewpoint (Loader, 2004). This is why, in the tests presented, we make n vary between 20 and 100.

The analysis procedure, carried out once for a global procedure but reiterated for each polygon in the case of a local analysis, involves two phases:

- A simple linear regression between the mean monthly temperature (response variable) and the nine explanatory variables recorded in the GIS (Section 2.2.1) can be used to calculate the Pearson coefficient of correlation (r) associated with each explanatory variable. Then r can be used to identify the predictors that are significant at the 5% level.
- These candidate predictors are then systematically integrated into multiple regressions by ascending stepwise selection, two by two, then three by three, and so on, until the combination that groups all of them is reached. The combination yielding the greatest R^2 value is selected. The parameters of the multiple regression are stored for the validation process and for the interpolation to come.

3.3. Validation

The quality of the estimations is evaluated by the so-called 'leave-one-out cross-validation', the principle of which is to calculate the parameters of the regression model using all the dataset except for one weather station and to reiterate for each station (Plutowski *et al.*, 1994). Multiple regressions (whose explanatory variables are those of the combination just identified) are conducted as

Table II. The Pearson correlation coefficient (×100) of variables used in multiple regressions of monthly temperatures.

	January	February	March	April	May	June	July	August	September	October	November	December
Altitude	−72	−73	−80	−79	−75	−67	−56	−59	−66	−71	−72	−72
Ruggedness	–	19	–	–	–	–	–	24	–	18	21	–
Slope angle	20	–	18	16	18	13	–	18	–	–	–	20
Distance to sea	−50	−44	−32	−21	–	–	−15	−19	−31	−40	−49	−50
Distance to forest	–	–	–	21	–	11	–	–	–	–	–	–
Vegetation index	–	–	–	–	–	–	12	–	–	–	–	–

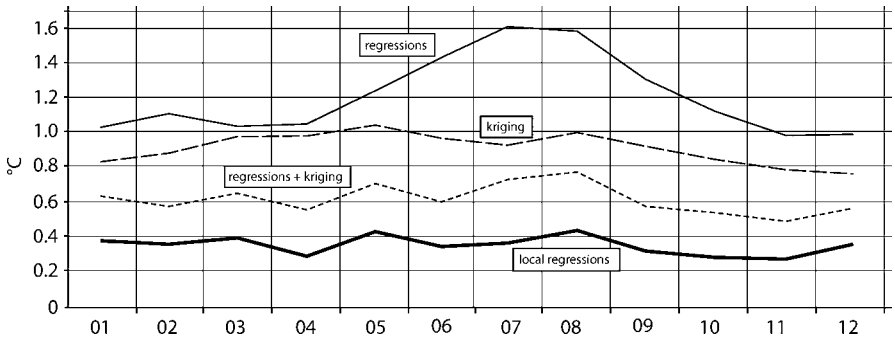

Figure 5. Standard deviation of residuals from the four approaches by month.

many times as there are stations. In the case of global regression, for each iteration, one station is rejected; thus, it is not involved in determining the parameters, and its predicted value is computed from the estimated equation. The difference between the predicted and the observed temperature for this station is computed. For local regression, the same process is applied. Each of the 641 stations is related to the polygon of which the gravity centre is the nearest. The parameters are estimated using $n-1$ stations and the cross-validation test is repeated 651 times as for global regression.

The interpolation itself is then performed for each of the 8 704 283 × 250 m-sided cells in France. Obviously there are as many analyses as there are monthly temperatures to estimate.

4. Comparison of the four interpolation methods

Global interpolation requires just a single regression, whereas local interpolation requires as many regressions as there are polygons ($p = 25\,618$ for $n = 30$). The coefficients from the analysis/analyses are then applied to each of the 8 704 283 × 250 m-sided cells in France to reconstruct the continuous temperature field.

4.1. Quality of estimations

The quality of interpolation is evaluated from the value of residuals given by cross-validation: extreme values, frequency of high values, and standard deviation values which have been calculated for each of the four types of analysis. All of these indications and measurements of dispersion exhibit wide deviations from one method to another. These deviations, which move in the same direction, are coherent and can be used to classify the types of analysis in terms of the quality of estimations produced.

4.1.1. Regression method

Stepwise selection of predictors in the spatial variation model of each of the 12 mean monthly temperatures leads to combining three to four among the nine available explanatory variables (Table II). Altitude is invariably selected; depending on the month in question, the Pearson correlation coefficient r varies from −0.80 (March) to −0.56 (July). Mean winter temperatures are therefore more sensitive to altitude than mean summer temperatures. Distance to the sea (reduced by its covariation with altitude) follows an analogous pattern; it is not an explanatory factor in summer time. The influence of the other variables is less marked. Vegetation influences temperature between April and July at the time of most intense growth. Ruggedness and slope angle are complementary. They invariably (except for July and September) occur alternately (the two being associated in August alone) in the temperature estimation function.

The value of residuals is not the same throughout the year. Figure 5 shows that the value of the standard deviation is lowest in winter (1 °C) and that it rises to 1.6 °C in summer. Sporadically, some stations display very high estimation errors (Figure 6(a)). For negative residuals, the maximum (in absolute values) is −4.7 °C; the highest positive residual is 4.9 °C. Lastly, 1.9% of residuals exceed 3 °C in absolute value, whereas low residuals (<1 °C) make up 60% of the population.

4.1.2. Kriging

Kriging provides better results with standard deviations close to 0.9 °C (0.8 °C in December, 1 C in April).

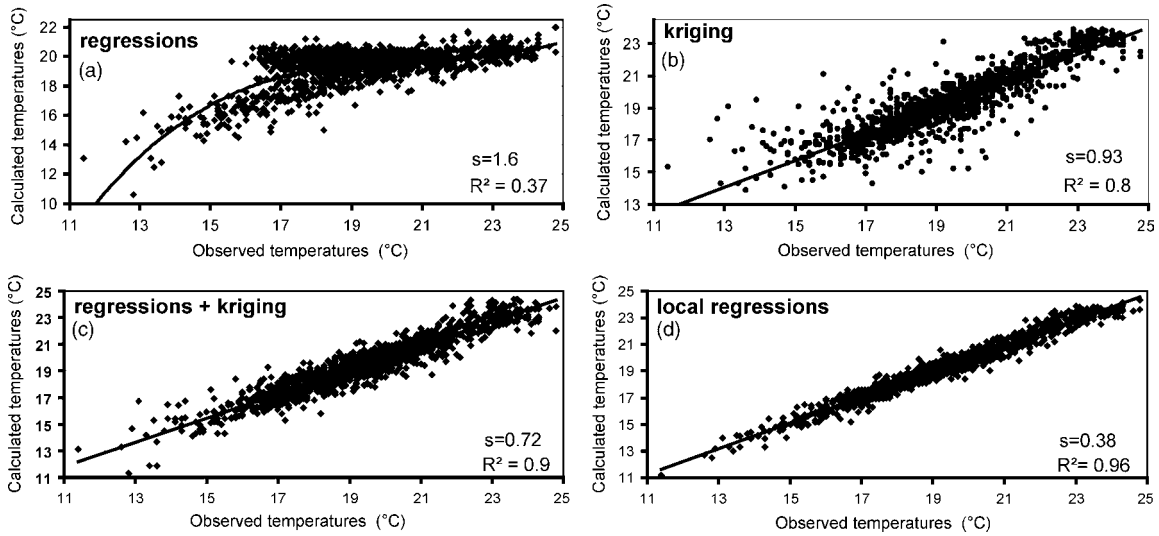

Figure 6. Scatter diagram between the observed and the predicted values resulting from a cross-validation applied to the four methods (July).

Sporadically (Figure 6(b)), large deviations appear also with kriging. The extreme residuals lie at −6.9 and +4.9 °C. The frequency of residuals greater than 3 °C in absolute value is 1.4%; the frequency of low residuals (<1 °C) is 83%.

4.1.3. *Regression then kriging*

The approach concatenating regressions than kriging further improves results. The Moran index of residuals from the regression stage (0.88) indicates strong residual autocorrelation. Moran's spatial autocorrelation coefficient is an indicator relating local covariance, calculated between neighbouring points, to overall variance (Anselin, 1988).

After kriging, the standard deviation of residuals falls to 0.72 °C with, here again, a better estimation of winter temperatures (0.5 °C) than summer temperatures (0.75 °C). The extreme values lie at −3.7 and 3.5 °C. Lastly, the frequency of high residuals (>3 °C in absolute value) becomes insignificant (0.1%), whereas the frequency of low residuals (<1 °C) exceeds 90%.

4.1.4. *Local regression*

It is the local regression method (tested here with $n = 30$ which generates 25 618 successive regressions) that proves best. The standard deviation of residuals is less than 0.3 °C for all months (except May and August) and sometimes verges on 0.3 °C (April, October, November). Residuals are never less than −2 °C or more than 1.9 °C. The frequency of values between −1 and 1 °C is now 98%, while R^2 reaches 0.96 (Figure 6(d)).

4.2. Influence of n on the standard deviation of residuals

The previous operation allowed us to measure the standard deviation of residuals from four methods. The local regression method was parametered with $n = 30$, but other values are possible. To assess the influence of this parameter on the quality of estimations, we conducted a series of six analyses modifying n from $n = 20$ to $n = 100$.

The best results are with the lowest value of n (20). In this case, standard deviations range from 0.25 (November) to 0.39 °C (August). With $n = 30$, the standard deviations are consistently greater than about 0.05 °C. Thereafter, this trend continues so that the standard deviations obtained with $n = 100$ are almost 0.17 °C higher than with $n = 20$. There is nothing to indicate that this trend might reverse. As standard deviation values tend to increase with n, it is conceivable that when this parameter becomes very high, the standard deviations are very close to those given by the global method (regression + kriging). In this case, when $n = 651$, the technique becomes identical to the global regression method because the number p of polygon is 1 (= one regression analysis) for the whole of study area.

This assessment is logical in spatial terms. With $n = 20$, the 20 stations included in the regressions are spread over a limited area with consistent environmental and climatic characteristics, so the estimations are excellent. But as n increases, the regressions include stations that are ever further from the initial core of 20, which tends to diversify the dataset, making it increasingly like that of France as a whole. The quality of correlations is affected and the standard deviations increase. Although these comments plead for a low n value, it should be stated that the level of significance diminishes with n (there are fewer degrees of freedom). Even if the accuracy of the estimates increases as n decreases, we did not apply a value for n lower than 20, even if that would be technically possible. It is not obvious at which value of n we would expect the skill to be highest, but it seems unlikely that it would be for very small n. This needs further investigation.

4.3. Interpolations

Given the constraints listed in Section 3.2, the local regression method was developed for interpolations that,

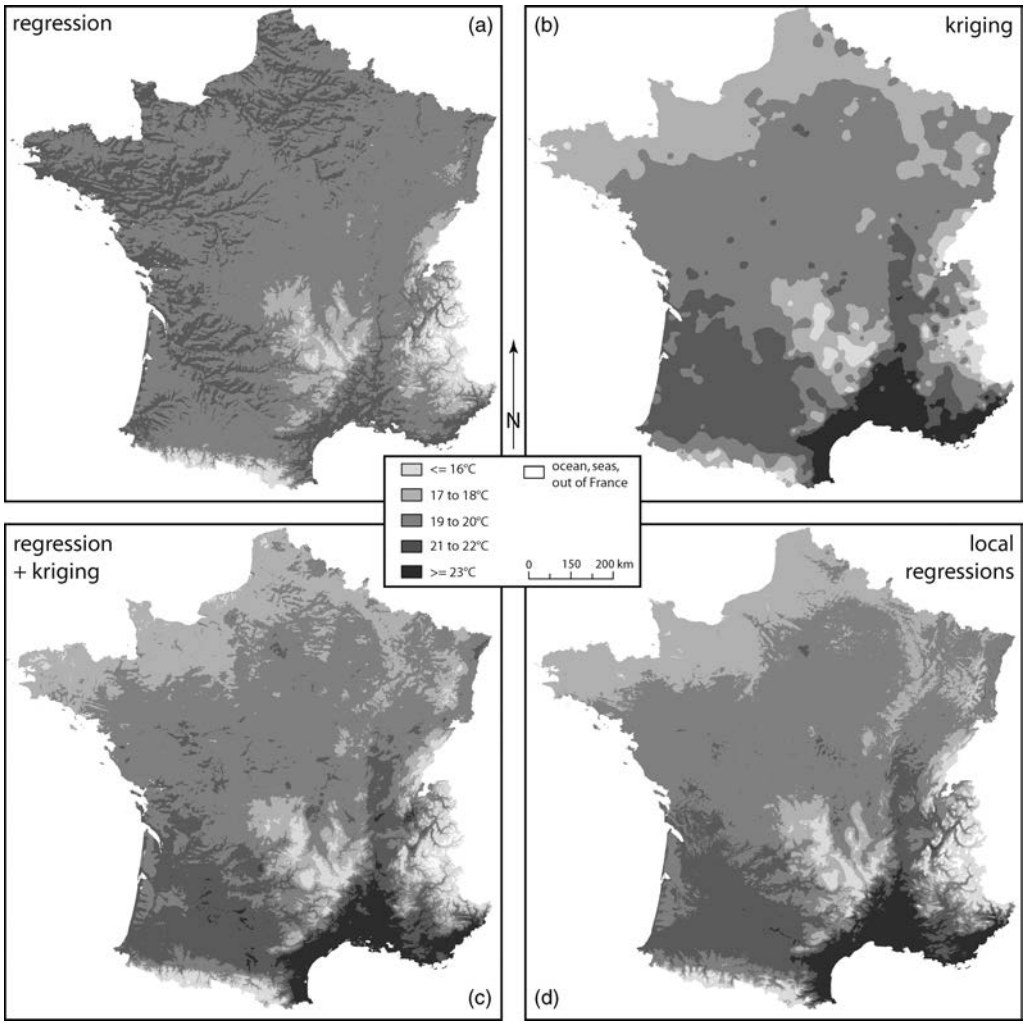

Figure 7. July temperatures predicted by the four methods.

with $n = 30$, follow a middle road between low values that are statistically unreliable and higher values that are ill-suited to the problem at hand (geographically too large an area over which to recruit weather stations, lengthy computation times).

4.3.1. Interpolation of July temperature

The results from each of the four methods are mapped (Figure 7). The regression function from the tests involves these parameters:

$$\text{July temperature} = f\ (\text{altitude, distance to the sea, vegetation index}) \quad (4)$$

The resulting equation is:

$$\text{Temp (July)} = 20.32 + (-0.0033 \times \text{alt.}) + (0.0007 \times \text{dist. sea}) + (0.0049 \times \text{veg. index}) \quad (5)$$

The map (Figure 7(a)) brings out the structuring effect of altitude: the cold mountain ranges contrast with the remainder of the country that is composed of plain and plateau. The temperature falls by 0.33 °C per 100 m. Distance to the sea is also a differentiating factor. Besançon, a semi-continental town in eastern France, contrasts with Brest, at the western tip of the country, where temperatures are 2.6 °C higher. Lastly, the vegetation index reveals that barren zones (city centres, rock) are 1.2 °C warmer than vegetation-covered areas (forests).

Kriging yields a map (Figure 7(b)) where mountains and highlands exhibit temperatures of less than 16 °C. By contrast, the Mediterranean rim and its northward extension along the Rhône Valley between Marseille and Lyon have values exceeding 22 °C. The map generated by combined regression and kriging of residuals (Figure 7(c)) clearly shows the separate effects of each model. Kriging tends to smooth distributions, whereas regression indicates the effect of topography in the fine contrasts where local variability is expressed. Lastly, Figure 7(d) is the result of local interpolation on the basis of 25 618 multiple regressions. Other factors such as land cover are also perceptible (e.g. the pocket of heat due to the influence of the Paris urban area on temperatures). Local topographic effects stand out clearly here. It should be specified that kriging of residuals was omitted for the reasons that will be given in Section 4.3.3.

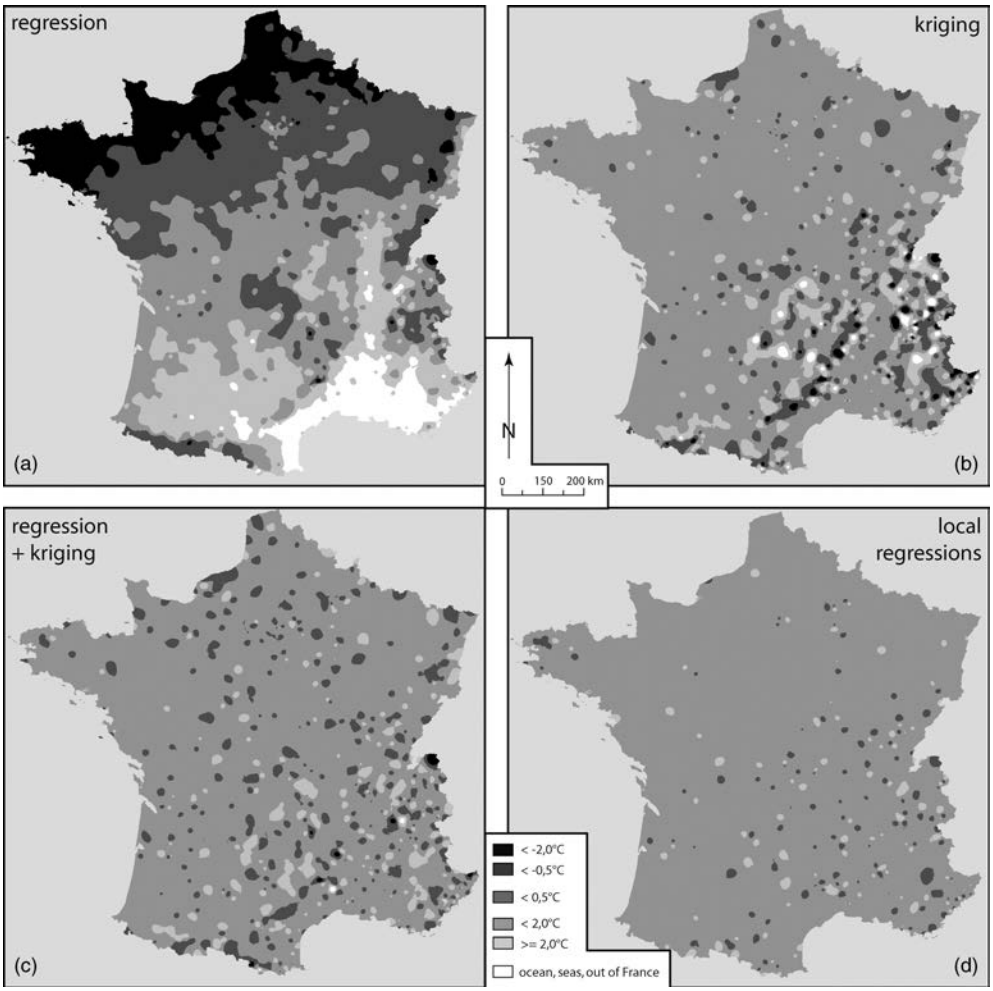

Figure 8. Residuals from the four methods (July temperatures).

4.3.2. *Mapping residuals*

The residuals obtained by cross-validation are interpolated by kriging so as to provide a continuous field of estimation errors for the four models.

The map for the global regression method (Figure 8(a)) shows that the extreme values are at the north (negative residuals) and south (positive residuals) of France. Kriging (Figure 8(b)) yields better results as the mid greys (residuals between -0.5 and $+0.5\,^{\circ}$C) occupy a much larger part of the map. The contrast values are denser in the Centre and south-east, leaving just a scattering of spots over the remainder of the territory, revealing nugget effects. The residuals from the last two methods are even smaller and, above all, are less and less concentrated in coherent zones. The local regression method gives the impression of a background made up of very weak residuals overall, dotted with small spots where values exhibit large variations (Figure 8(d)).

4.3.3. *Autocorrelation of residuals*

At the end of the phase of local analysis by regression, the question arises as to whether there is still any autocorrelation of residuals that could be kriged. Figure 9 shows above all that the autocorrelation is not evenly distributed across the territory. Such a situation is troublesome as kriging theory stipulates that it cannot be used properly unless the same spatial distribution model is observed everywhere; but this is not the case. Accordingly, kriging calculated globally from such a heterogeneous situation makes very little improvement on local regression and sometimes even entails an increase in the value of residuals. In July, kriging deteriorates the results obtained at the end of the regression stage. The standard deviation of residuals shifts from 0.35 to 0.41 °C.

5. Mapping coefficients

The local regression method leads to segmentation of the space into a large number of polygons of various sizes. For each polygon, applying simple regressions between temperature and the nine explanatory variables first yields the Pearson correlation coefficient and the parameter estimate related to each of them; then a multiple regression is performed yielding R^2. Each of these coefficients provides specific information about the local climate characteristics. Mapping them provides insight into climatic processes at local scale. We illustrate this with January and July temperatures.

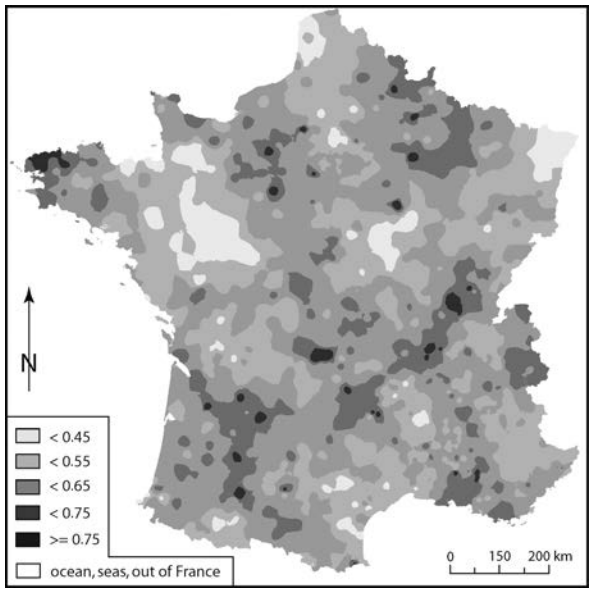

Figure 9. Spatial variation of the Moran's coefficient of residuals.

As it was just said, there is only one coefficient for each polygon. These values have been plotted at the centre of each polygon and then interpolated by kriging to produce the maps.

5.1. Spatial variation of the coefficient of determination related to altitude

The coefficient of determination R^2 is a measure of the level of explanation of variance. Altitude is a powerful explanatory factor for spatial variation in temperature and probably plays a key role in this distribution (Figure. 10(a)). The case of the low-altitude Paris region can be explained here primarily by the influence of the pocket of urban heartland, one further factor behind the spatial variation in temperature (Joly, 2007); the density of weather stations around Paris is probably also instrumental in raising the value of R^2.

Many parts of the east and south of the Paris Basin, the Atlantic coast, the Aquitaine Basin, and the Camargue and Marseille area exhibit R^2 values less than 0.5. These are low-lying regions with only slight changes in altitude. Under these circumstances, topography has a marginal effect. And yet other factors, such as distance to the sea, could supersede altitude as an explanatory factor and so cause an increased R^2; but this is not so. For example, the flatland Camargue area confirms this hypothesis in that the influence of the sea is less marked because the prevailing mistral is a continental wind.

5.2. Spatial variation of the parameter estimate related to altitude

The value of the parameter estimate provides information about the vertical temperature gradient. The global value for July (-0.0033) corresponds to a drop in temperature of 0.33 °C per 100 m. The sectors with constants between -0.002 and -0.005 °C, the range of values centred on the global value for July, cover more than 75% of the territory (Figure 10(b)). They are all superimposed on zones characterized by high r values (especially mountain areas).

The zones with values less than -0.006 °C are subject to very high altitudinal thermal gradients (0.7, 0.8 °C per 100 m). They are spread over many spots in the Paris Basin and Brittany in particular. These marked gradients may be explained by exaggerated thermal variations between overheated low zones and much cooler, windier, (even moderately) higher zones. This hypothesis is probably valid for high gradients located slightly inland from the Atlantic coast.

The parameter estimate values less than 0.002 indicate that altitudinal gradients are almost nonexistent locally (<0.2 °C per 100 m). Most of the areas concerned are sectors where the R^2 value is low too. These are probably sectors where altitude variations are too low to engender any significant variations in temperature.

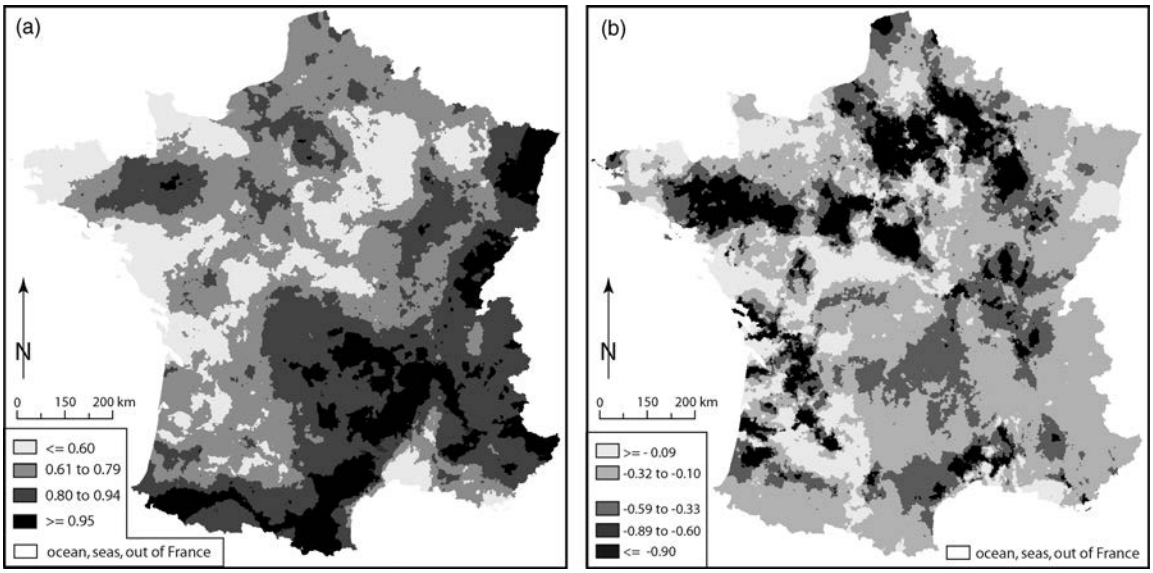

Figure 10. Spatial variation of two coefficients related to altitude (July); (a) R^2 values (simple regression); (b) Parameter estimate.

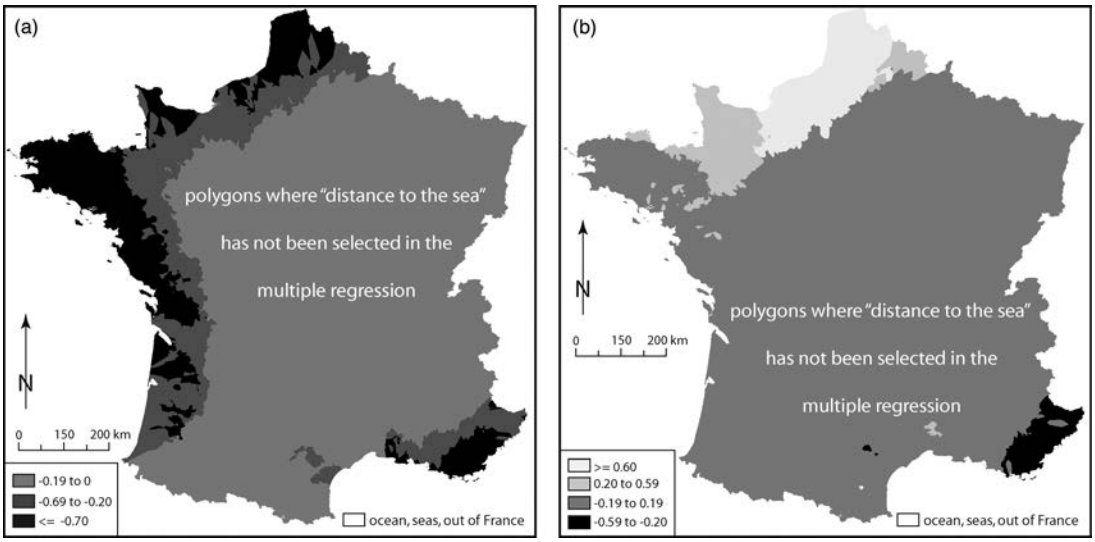

Figure 11. Spatial variation of the Pearson coefficient of correlation related to distance to the sea; only the polygons having selected 'distance to the sea' in the multiple regression are represented. (a) January and (b) July.

5.3. Spatial variation of the Pearson correlation coefficient related to distance to the sea

The spatial variation of the Pearson correlation coefficient related to distance to the sea shows how far inland the influence of the sea extends. The highest values ($r < -0.70$) for January (Figure 11(a)) occur along the North Sea and the Atlantic Ocean (especially in Brittany) and along the Mediterranean Sea east of Marseille. There, the warming up of the air is the maximum close to the sea and then decreases. Beyond a distance of 150 km, the influence of the sea is no more discernible. Along the Rhône Valley and especially in Camargue, the coefficient is greater than -0.19 because of the *Mistral*, a strong continental wind blowing from North to South. West of the Rhône Valley, the coefficient is greater than -0.70 because of another local wind blowing from West to East. There, the air from the land is blown down to the sea so that its influence is reduced.

In July (Figure 11(b)), the spatial pattern of the Pearson coefficient of correlation related to distance to the sea is different from the previous one except east of the Rhône valley (Côte d'Azur). Maybe, some collinearity with altitude remains there. The coefficient is positive along the coast of North Sea. There, the sea has a cooler effect on temperature which increases towards inland.

6. Conclusion

Many interpolation methods are used in climatology but all have their limits so that the quality of estimations produced varies greatly with context. The choice of a single option is generally insufficient and the combined implementation of complementary methods such as regression and kriging improves results. This study also shows the value of a local interpolation approach which is based on an analysis of the n stations closest to each cell to be estimated. The n parameter is important as it determines the speed of execution of the computation and the quality of the interpolations. For a given n value, the cells depending on the same 'nearest neighbours' are part of the same polygon. Several thousand polygons are defined on this basis. A statistical correlation analysis is conducted for each polygon: identification of significant estimators, choice of multiple regression formula, and evaluation of the residual of the estimation by cross-validation.

The example used for this demonstration pertains to the mean monthly temperatures. The results are plain: the standard deviation of residuals from the local regression method (0.4 °C) is 0.2 °C lower than that from the method combining both global multiple regression and kriging of residuals, which in turn is 0.2 °C lower than that from kriging alone. This rank order is maintained whichever month is considered. The highest residuals from kriging are found mostly in the zones where topography is contrasted (mountain ranges and contact between mountains and piedmont), whereas with the local regression method they have been greatly reduced with no preferential distribution. As with global methods, this method also allows other variables than the one of interest to be mapped. Several values are available for each regression performed (one per polygon) and may be mapped as shown in the three examples: mapping of R^2, of r, and of the parameter estimate related to temperature and altitude. We could likewise have mapped the estimated parameters of the other nine explanatory variables available in the GIS. In the same way, the results of analyses for temperature for all the other months of the year can be analysed and mapped. It would then be possible to monitor the changes over time in R^2, in the coefficient of regression, etc., all year round.

Many other variables, if recorded at a large number of stations, could also be processed by local analyses: number of days when certain temperatures (<-5 °C, $+30$ °C) are exceeded, levels and number of days of precipitation for each month of the year (Joly *et al.*, 2009), etc.

Tests of monthly rainfall show that local interpolations also yield better results than global methods. Similarly, it would be interesting to monitor daily the deformation of spatial heterogeneity of regressions as a function of changes in synoptical situations.

In this way, it becomes possible to study climate and its behaviour and variations thereof at higher spatio-temporal resolutions. This approach may provide a diagnosis of factors behind spatio-temporal variations in climate. Such results could be material for a climate atlas: an atlas of means and frequencies, but also and above all an atlas of the spatial workings of climate.

Acknowledgement

This research was funded by grant 0001724 (2 November 2006) from 'MEEDDM'. The views stated in this publication are those of the authors and do not represent views of the MEEDDM

References

Anselin L. 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers: Dordrecht.

Arnaud M, Emery X. 2000. *Estimation et Interpolation Spatiale : Méthodes Déterministes et Méthodes Géostatistiques*. Hermès: Paris; 221.

Baillargeon S. 2005. *Le krigeage : Revue de la théorie et Application à l'interpolation Spatiale de données de Précipitations*. Mémoire de la faculté des sciences et de génie de l'université Laval: Québec; 128. Available from http://www.theses.ulaval.ca/2005/22636/22636.pdf [accessed April 2005].

Cleveland W, Devlin S. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**: 596–610.

Cressie N. 1993. *Statistics for Spatial Data*, Revised edn. Wiley: NewYork, 900.

Courault D, Monestiez P. 1999. Spatial interpolation of air temperature according to atmospheric circulation patterns in southeast France. *International Journal of Climatology* **19**: 365–378.

De Berg M, Van Kreveld M, Overmars M, Schwarzkopf O. 2000. *Computational Geometry: Algorithms and Applications*, 2nd edn. Springer: Berlin, 367.

Fotheringham A, Brundson C, Charlton M. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons, Ltd.: Chichester; 288.

Goovaerts P. 1997. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology* **228**: 113–129.

Gunst RF. 1983. Regression analysis with multicollinear predictor variables: definition, detection, and effects. *Communications in Statistics: Theory And Methods* **12**: 2217–2260.

Helland IS. 1990. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics* **17**: 97–114.

Hocking RR. 1976. The analysis and selection of variables in linear regression. *Biometrics* **32**: 1–40.

Joly D. 2007. L'information géographique au service de la climatologie. In *Information Géographique et Aménagement du Territoire; Information géographique et climatologie*. Hermes: Paris; 23–72.

Joly D, Brossard T, Cardot H, Cavaillès J, Hilal M, Wavresky P. 2008. Interpolation par recherche d'information locale. *Climatologie* **5**: 27–48.

Joly D, Brossard T, Cardot H, Cavaillès J, Hilal M, Wavresky P. 2009. Interpolation par régressions locales: application aux précipitations en France. *L'Espace Géographique* **2**: 157–170.

Joly D, Brossard T, Cardot H, Cavaillès J, Hilal M, Wavresky P. 2010. Les types de climat en France, une construction spatiale. *Cybergéo* **501**. Available from http://cybergeo.revues.org/index23155.html [accessed 18 June 2010].

Joly D, Nilsen L, Fury R, Elvebakk A, Brossard T. 2003. Temperature interpolation at a large scale; test on a small area in Svalbard. *International Journal of Climatology* **23**: 1637–1654.

Loader C. 2004. Smoothing: local regression techniques. In *Handbook of Computational Techniques*, Gentle J, Härdle W, Moty Y (eds). Springler: New York; 539–564.

Matheron G. 1970. *Traité de Géostatistique Appliquée, Tome 1. Mémoires du bureau de recherches géologiques et minières, n°14*, Technip: Paris.

Mitas L, Mitasova H. 1999. Spatial interpolation. In *Geographical Information Systems: Principles and Technical Issues*, Vol. 1, Longley PA, Goodchild M, Maguire D, Rhind DW (eds). John Wiley & Sons, Inc.: New York; 481–492.

Plutowski M, Sakata S, White H. 1994. Cross-validation estimates IMSE. In *Advances in Neural Information Processing Systems 6*, Cowan JD, Tesauro G, Alspector J (eds). San Mateo, CA: Morgan Kaufman, 391–398.

Wackernagel H. 2003. *Multivariate Geostatistics: An Introduction with Applications*, Third completely revised edn. Springler-Verlag: Berlin.

Wand MP, Jones MC. 1995. *Kernel Smoothing*, *Monographs on Statistics and Applied Probability, 60*, Chapman & Hall: London, 2.

Wold S, Ruhe A, Wold H, Dunn WJ III. 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Scitific and Statistical Computations* **5**(3): 735–743.