Interpolation Sometimes Enhances and Sometimes Impedes Spatial Correlation: Simple Pedagogical Examples

Olga Kosheleva[1] and Vladik Kreinovich[2]
[1]Department of Teacher Education
[2]Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968, USA
olgak@utep.edu, vladik@utep.edu

Abstract

A natural way to check whether there is a dependence between two quantities is to estimate their correlation. For spatial quantities, such an estimation is complicated by the fact that, in general, we measure the values of the two quantities of interest in somewhat different locations. In this case, one possibility is to correlate each value of the first quantity with the value of the second quantity measured at a nearby point. An alternative idea is to first apply an appropriate interpolation to each of the quantities, and then look for the correlation between the resulting spatial maps. Empirical results show that sometimes one of these techniques leads to a larger correlation, and sometimes the other one. In this paper, we provide simple pedagogical examples explaining why sometimes interpolation enhances spatial correlation and sometimes interpolation impedes correlation.

1 Two Approaches to Spatial Correlation: Empirical Evidence and Need for Simple Pedagogical Examples

Formulation of the problem. In many practical situations, we want to find the correlation between spatially distributed quantities $a(x, y)$ and $a'(x, y)$.

For example:

- we have pollution measurements at different spatial points,
- we have frequencies of allergies in different geographic locations, and
- we want to see if some of these allergy cases are caused by pollution.

Another example:

- we use different techniques to measure the qualify of the road pavement, and
- we want to make sure that the corresponding measurements indeed measure the same quantity.

In the *ideal* case,when both quantities a and a' are measured at the exact *same* geographic locations (x_1, y_1), ..., (x_n, y_n). In this case, we have n pairs of values $a_i \stackrel{\rm def}{=} a(x_i, y_i)$ and $a'_i \stackrel{\rm def}{=} a'(x_i, y_i)$, and we can use the standard statistical formula to find the correlation between the corresponding a- and a'-values; see, e.g., [1].

Often, however, the measurements of the two quantities is performed at *different* spatial locations:

- the quantity a was measured at the locations (x_1, y_1), ..., (x_n, y_n), while
- the quantity a' was measured at spatial locations (x'_1, y'_1), ..., $(x'_{n'}, y'_{n'})$.

In this case, we have two main options:

- we can simply try to correlate each value $a_i = a(x_i, y_i)$ with the values $a'_{j(i)} = a'(x'_{j(i)}, y'_{j(i)})$ measured at a location $(x'_{j(i)}, y'_{j(i)})$ which is the closest to (x_i, y_i), or
- we can interpolate both data to the whole region – i.e., build two maps $A(x, y)$ and $A'(x, y)$, and then look for the correlation between the corresponding interpolated values $A(x, y)$ and $A'(x, y)$.

Comment. In addition to these two approaches, we can also *combine* these two techniques. For example, we can interpolate the values a' into a map, and look for correlation between:

- the measured values of the quantity a, i.e., the values $a_i = a(x_i, y_i)$ and
- the values $A'(x_i, y_i)$ obtained by interpolation from the results $a'(x'_j, y'_j)$ of measuring the quantity a'.

Sometimes, we have statistically significant correlation only in one of these techniques. Sometimes, we can see statistically significant correlation between the original data values, but once we interpolate and look for the relation between the maps, the correlation disappears. This makes sense from the engineering viewpoint:

- interpolation, crudely speaking, is adding "guesses" to the measurement results, and
- guesses corresponding to quantities a and a' can go in different directions, thus decreasing the observed correlation.

From this viewpoint, one would expect that interpolation always impedes spatial correlation. However, interestingly, there are cases when the effect is the opposite: interpolation enhances the correlation (sometimes to the extent that the statistically significant correlation can only be detected after the interpolation).

Problem: we need simple examples explaining the corresponding phenomena. The above two types of examples occur with real data, when we have a large number of data. These examples come as a result of extensive data processing, and it is not intuitively clear what is going on and what is the reason for the corresponding enhancing or impeding.

To clarify the situation, it would be nice to have simple examples, traceable by hand, when the correlation is enhanced and when the correlation is impeded.

What we do in this paper. In this paper, we provide two simple pedagogical examples of such a phenomena.

2 Examples

Let us make our examples as simple as possible. To make the examples as simple as possible, let us make them 1-dimensional instead of 2-dimensional. In other words, in both examples, we have values $a(x)$ and $a'(x')$ measured at certain points $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_{n'}$.

Another thing that we do to make examples simple is to use a simple easy-to-trace-by-hand piece-wise linear interpolation instead of possible more sophisticated ones (which are usually not easy to trace by hand). Thus, if we, e.g., know the values $a_i = a(x_i)$ at points $x_1 < x_2 < \ldots < x_n$, then, for each location x between x_i and x_{i+1}, as an interpolated value $A(x)$, we take

$$A(x) = a_i + \frac{x - x_i}{x_{i+1} - x_i} \cdot (a_{i+1} - a_i).$$

A simple example when interpolation impedes correlation. Let us assume that the actual values of the quantity a are $a(x) = x$, and that the actual values of the quantity a' are $a'(x) = x^2$. Suppose that:

- we have measured the value of the quantity a at two locations $x_1 = 0$ and $x_2 = 1$, the resulting values are $a_1 = a(x_1) = 0$ and $a_2 = a(x_2) = 1$, and
- we have measured the value of the quantity a' at three locations $x'_1 = 0$, $x'_2 = 0.5$, and $x'_3 = 1$; the resulting values are $a'_1 = 0$, $a'_1 = 0.25$, and $a'_2 = 1$.

In this case, if we look for correlation without interpolation, then:

- we compare the value $a_1 = a(x_1) = 0$ measured at the location $x_1 = 0$ with the value $a'_1 = a'(x'_1) = 0$ measured at the closest (here, identical) location $x'_1 = 0$;

- similarly, we compare the the value $a_2 = a(x_2) = 1$ measured at the location $x_2 = 1$ with the value $a'_3 = a'(x'_3) = 1$ measured at the closest (here, identical) location $x'_3 = 1$.

The resulting two tuples $a_1 = 0$ and $a_1 = 1$ and $a'_{j(1)} = 0$ and $a'_{j(2)} = 1$ are identical, so the correlation between them is 1.

On the other hand, if we use linear interpolation to interpolate $a(x)$ to the midpoint $x_2 = 0.5$, we get $A(x_1) = 0.5$. Now, we have a correlation between values $a'_i = (0, 0.25, 1)$ and the corresponding values $A_i = (0, 0.5, 1)$. These two tuples cannot be obtained from each other by a linear transformation and thus, the correlation is smaller than 1.

In other words, in this simple example, interpolation impedes correlation.

A simple example when interpolation enhances correlation. Let us consider the simple case when $a(x) = a'(x) = x$. Let us assume that we measure the quantity a at some points x_i and we measure a' at a different set of points x'_j.

Both functions $a(x)$ and $a'(x)$ are linear. For these functions, piece-wise linear interpolation reconstruct the exact values of $a(x)$ and $a'(x)$ for all intermediate values x. Thus, after interpolation, we get $a(x) = a'(x)$ for all x. These values are identical, so the correlation between them is clearly 1.

However, since the locations x'_j at which we measure the quantity a' are, in general, different from the values x_i at which we measure the quantity a, the correlation between:

- the measured values of a, i.e., values $a(x_i) = x_i$, and
- the corresponding values $a'_{j(i)}$ at points $x'_{j(i)}$ which are the closest to x_i

may get smaller than 1.

Let us consider a simple example. Suppose that:

- we have measured the quantity a at locations $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$, and
- we have measured the quantity a' at locations $x'_1 = 0.9$, $x'_2 = 2.1$, and $x'_3 = 3.0$.

In this simple example, for each i, the points x'_i is the closest to x_i, i.e., we have $j(i) = i$ for all i. Thus, when we compute the correlation between the original (non-interpolated) values, we must look for correlation between:

- the values $a_1 = a(x_1) = 1$, $a_2 = a(x_2) = 2$, and $a_3 = a(x_3) = 3$, and
- the values $a'(x'_{j(1)}) = a'(x'_1) = 0.9$, $a'(x'_{j(2)}) = a'(x'_2) = 2.1$, and $a'(x'_{j(3)}) = a'(x'_3) = 3.0$.

In this example, both sample means are equal to $\mu = \mu' = 2$, so:

- the differences $a_i - \mu$ are equal to -1, 0, and 1, and
- the differences $a'_i - \mu'$ are equal to -1.1, 0.1, and 1.

Here, the covariance is equal to

$$C(a,a') = \frac{(-1)\cdot(-1.1) + 0\cdot 0.1 + 1\cdot 1}{2} = 1.05,$$

$$V(a) = \frac{(-1)^2 + 0^2 + 1^2}{2} = 1,$$

and

$$V(a') = \frac{(-1.1)^2 + (0.1)^2 + 1^2}{2} = \frac{1.21 + 0.01 + 1}{2} = \frac{2.22}{2} = 1.11.$$

Thus, the correlation is equal to

$$\rho = \frac{(C(a,a')}{\sqrt{V(a)}\cdot\sqrt{V(a')}} = \frac{1.05}{\sqrt{1.11}}.$$

Here, $1.05^2 = 1.1025 < 1.11$, thus

$$\rho = \frac{1.05}{\sqrt{1.11}} = \sqrt{\frac{(1.05)^2}{1.11}} < 1.$$

So, in this case indeed, without interpolation, we get a smaller correlation value than with correlation – i.e., in this case, interpolation enhances correlation.

Comment. The above example describes a simplified case when the quantities a and a' are identical, and the only difference is between the spatial locations where they are measured. Can we get similar results when a and a' are strongly correlated but not identical? Absolutely.

We can emulate this by making small changes to the values $a(x_i)$ and $a'(x_j)$. Since all our formulas are continuous, when the changes are small, we will have the same inequality between the two correlation estimates – but already in a situation when the actual values $a(x)$ and $a'(x)$ are somewhat different.

Acknowledgments. This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

References

[1] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.