Correlation coefficients, temperature and distance separation. A dissection.

by Geoffrey Sherrington.

Lagged data strings.

An early step in the application of geostatistical mathematics to a data string is commonly the construction of a semivariogram. From Wikipedia, this and following material are descriptive.

http://en.wikipedia.org/wiki/Semivariogram

Quote: In spatial statistics the theoretical variogram $2\gamma(x,y)$ is a function describing the degree of spatial dependence of a spatial random field or stochastic process $Z(x)$. It is defined as the variance of the difference between field values at two locations across realizations of the field (Cressie 1993):

$$2\gamma(x,y) = \operatorname{var}(Z(x) - Z(y)) = E\left(|(Z(x) - \mu(x)) - (Z(y) - \mu(y))|^2\right).$$

If the spatial random field has constant mean μ, this is equivalent to the expectation for the squared increment of the values between locations x and y (Wackernagel 2003):

$$2\gamma(x,y) = E\left(|Z(x) - Z(y)|^2\right),$$

where $\gamma(x,y)$ itself is called the semivariogram. In case of a stationary process the variogram and semivariogram can be represented as a function $\gamma_s(h) = \gamma(0,0 + h)$ of the difference $h = y - x$ between locations only, by the following relation (Cressie 1993):

$\gamma(x,y) = \gamma_s(y - x)$. (End quote)

The semivariogram introduces the method of comparing data strings by looking at differences when the strings are shifted 1, 2, 3, … n points apart. In a familiar example, a set of chemical analyses of 1000 mm cuts, from adjacent intervals from a drill hole, would be expected to correlate best when adjacent cuts are chosen. If the correlation is made between intervals 10 cuts apart, a worse correlation is expected intuitively. The method is an approach to determining the separation at which a value at one place has some predictive power in estimating another value distant from it.

In the following essay, the formal mineral semivariogram mathematics are not used. They involve other properties such as stationarity, nugget effects, variations in host lithologies etc., to be considered. However, related methods of lagging data strings by 1, 2, 3, … n intervals can be used in conjunction with classical correlation coefficients, here calculated by the standard Excel command "CORREL".

- The equation for the correlation coefficient is:

$$Correl(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

where x and y are the sample means AVERAGE (array1) and AVERAGE(array2).

The problem of R in climate temperature strings.

With the unpredictable nature of weather, how do we arrive at correlation coefficients higher than 0.9? That is the fundamental question for this essay.

We shall tease apart data starting from annual temperatures, then to monthly, then to daily. We shall use one weather station mainly, because there is zero separation of distance and so we start with a best case. The record for Melbourne Australia suits. It extends at one site from 1854. We are not concerned with effects like UHI at this site for the methods used. Latitude and longitude are 37.8075S, 144.9700E, Australian Bureau of Meteorology Number is 106071, World Meteorological Number is 94868, altitude is 31m.

The variation of correlation coefficient with distance separation is integral to the study released in preliminary form by Rohde et al as part of the Berkeley University BEST project to examine UHI on a global scale. One can see that R>0.9 at some close separations, but in extreme cases R>0.8 at separations of 1000 to 1500 km. How can this be? It is atypical of much Earth Science data. The following figure with caption is from the BEST report of October 2011.

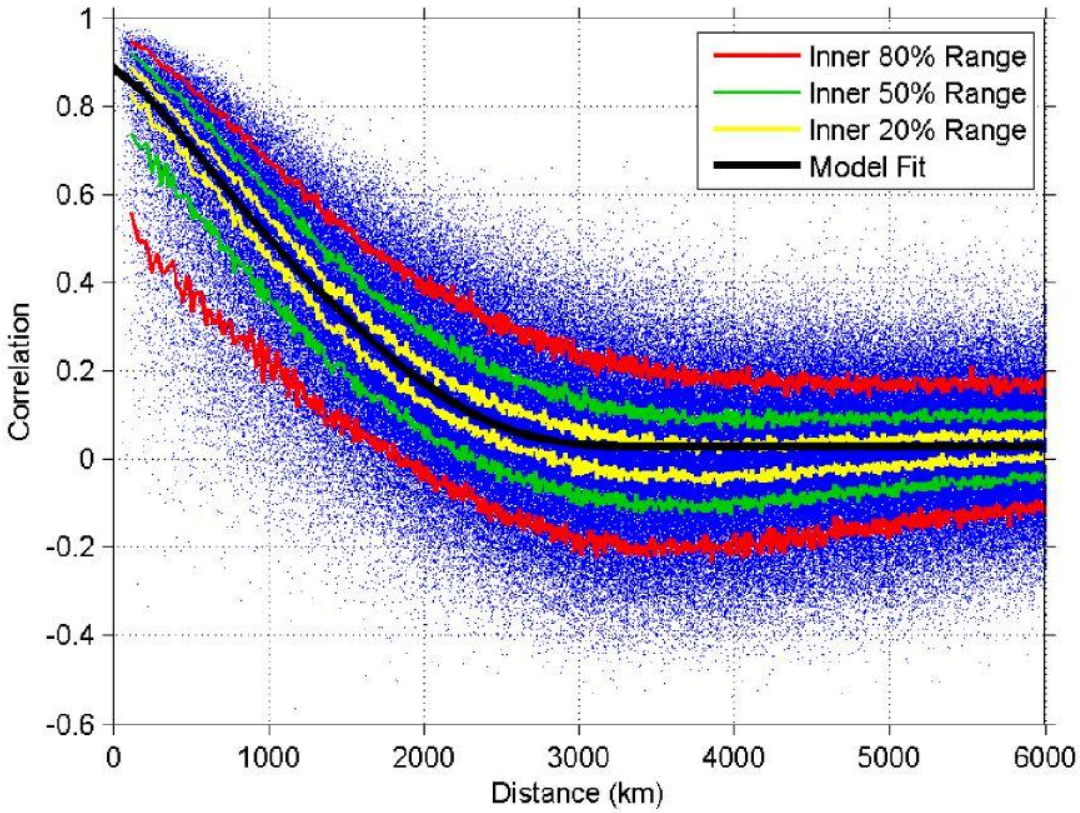
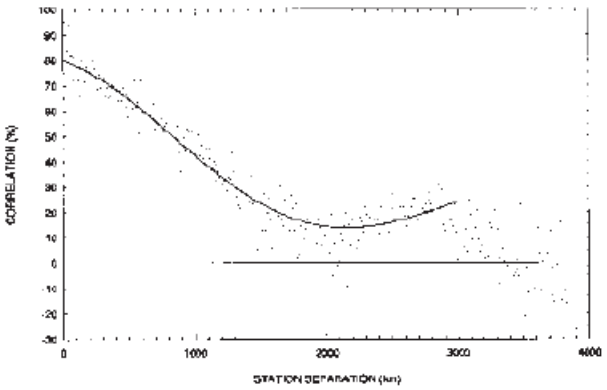


Figure 2. Mean correlation versus distance curve constructed from 500,000 pair-wise comparisons of station temperature records. Each station pair was selected at random, and the measured correlation was calculated after removing seasonality and with the requirement that they have at least 10 years of overlapping data. Red, green, and yellow curves show a moving range corresponding to the inner 80, 50, and 20% of data respectively. The black curve corresponds to the modeled correlation vs. distance reported in the text. This correlation versus distance model is used as the foundation of the Kriging process used in the Berkeley Average.

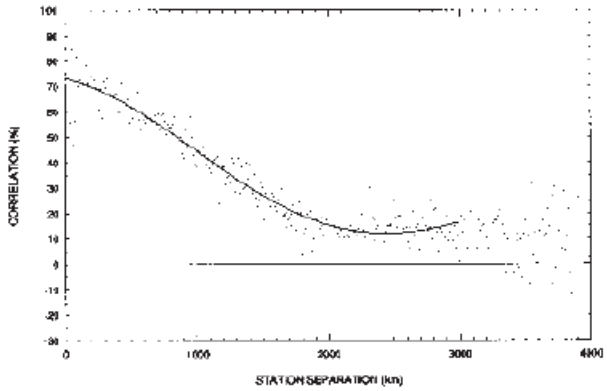
A related figure is in Della-Marta,Paul, Collins, Dean and Braganza, Karl. Aust. Met. Mag. 53 (2004) 75-93.

Fig. 4 **1961-1990 inter-station correlations (expressed as %) for annual mean (a) maximum and (b) minimum temperatures as a function of separation distance. Exponentially damped cosine functions are fitted where the station separation is less than 3000 km.**

(a)



(b)

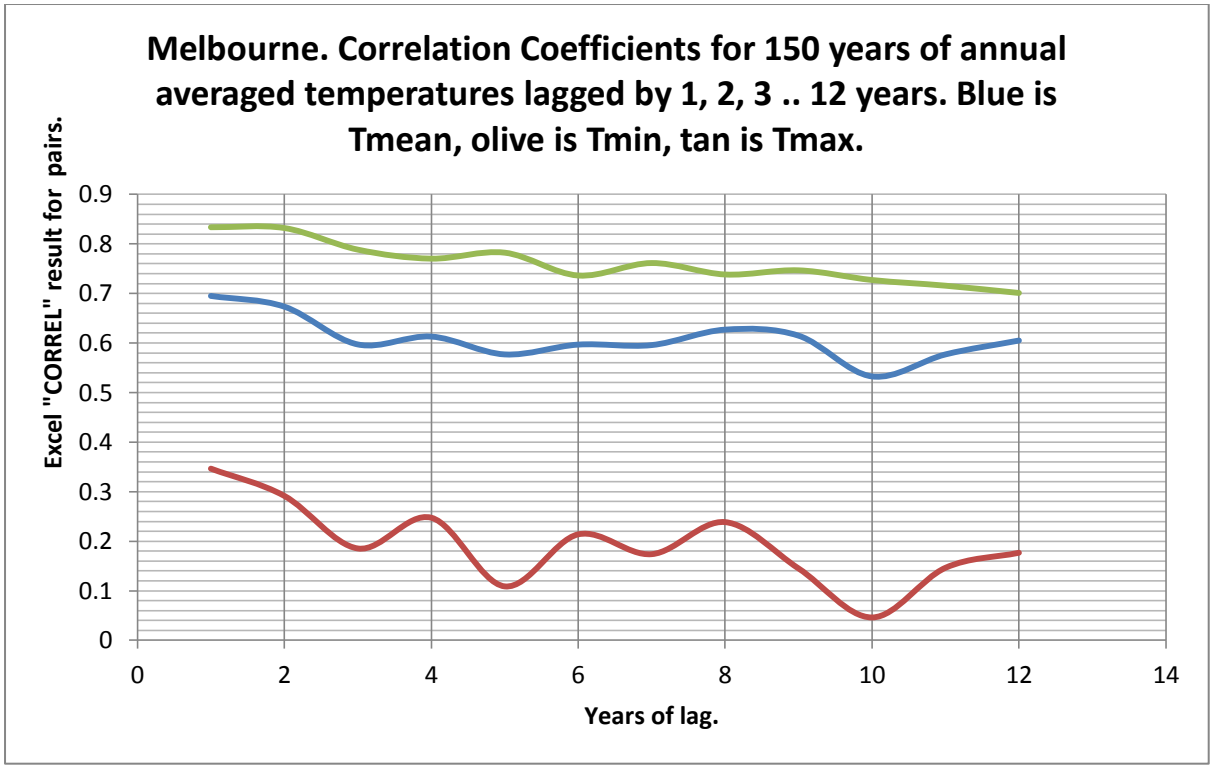


Methodology.

The temperature record for Melbourne has occasional missing values. These were infilled unscientifically by an approximate mean of those around them. There are too few corrections to affect the conclusions here. T mean was calculated as half of (Tmax+Tmin).

Stage 1. Melbourne annually for 150 years.

The approximately 150 years of annual averaged temperature data were analysed on Excel by lagging. That is, the correlation coefficient between the 150 values, and the same values moved down one year, was calculated. This was repeated with a 2 year lag, and so on up to 12 years. A year of data drops off the end at each shift so that finally about 137 pairs were correlated. A graph summarising the results is shown.

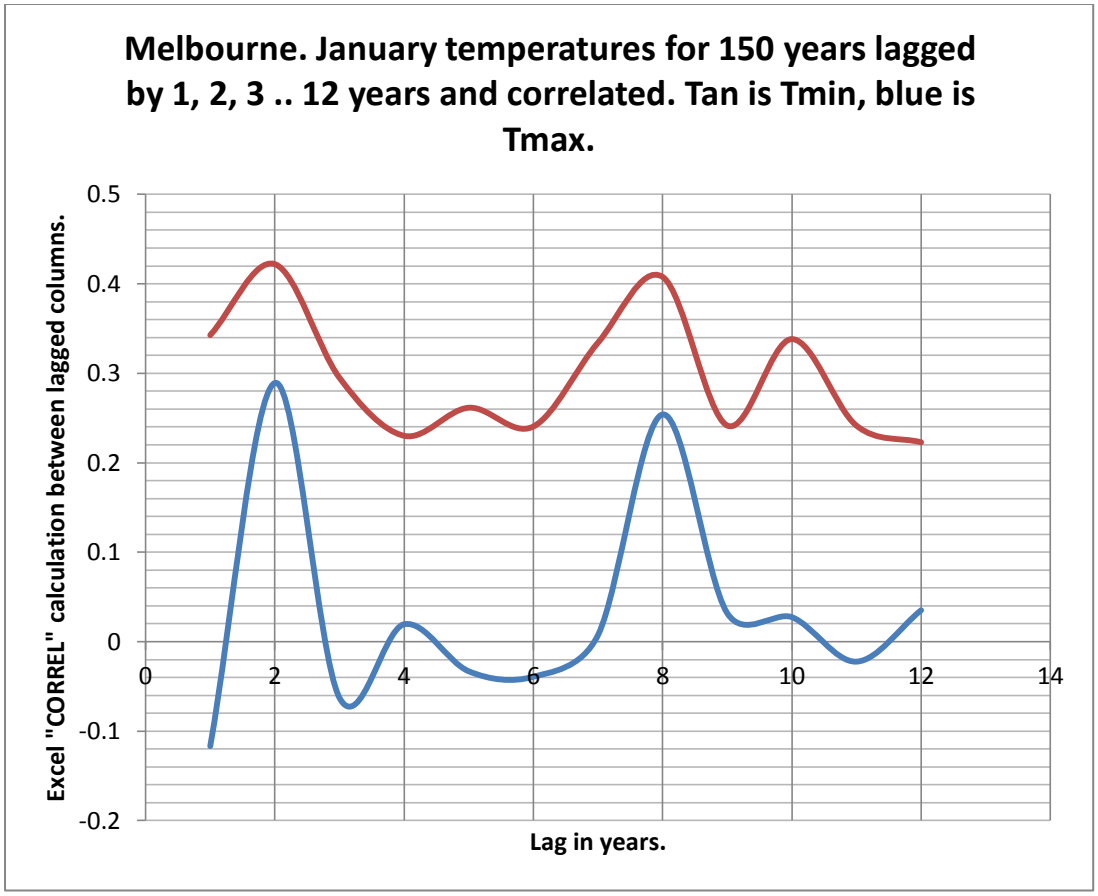


The interpretation is simplified by the choice of one site. The lag 1 results show how temperatures correlate with a time separation of 1 year. There are three obvious features. First, the best value of R is about 0.85, well below many of the correlation shown on the figures above for separated sites. Second, the Tmax correlations are greatly worse than the Tmin for reasons unknown. Therefore, a derivation of Tmean from them has a component of unknown origin. It is Tmean that is plotted on the BEST separation graph referenced. Third, as expected, the correlations worsen as the lags increase, showing that values well separated in time have reduced ability to predict earlier values.

These annual data were calculated from the arithmetic averages of daily data.

Stage 2. Melbourne, month of January for 150 years.

January, as a month, was selected for no particular reason for the next exercise, which can be repeated for any month. The results will be essentially similar.



Interpretation. The monthly result (here for January) is somewhat similar to annual result except that correlations are lower. There is little ability to estimate temperature in year Y+1 from year Y, by using monthly data. The peak at lag 8 years arises from unknown causes. It mildly infers an 8 year cycle in this month in this place, but it is contrary that there is no sign of an 11 year cycle that some might expect to see.

A second part of a monthly analysis is to simply compute correlation coefficients between adjacent months. For the full 150 years, here are the correlation coefficients for Jan-Feb, Feb-Mar, Mar-Apr etc in turn, first Tmax then Tmin.

JF	FM	MA	AM	MJ	JJ	JA	AS	SO	ON	ND	1856-2005.
0.221	0.260	0.020	0.289	0.328	0.418	0.187	0.293	0.202	0.252	0.220	
0.561	0.587	0.501	0.423	0.313	0.367	0.353	0.537	0.503	0.527	0.572	

Thus, temperatures in one month do not correlate highly with the adjacent month. Tmax again gives systematically poorer correlations than Tmin. All are well below “significant” correlation in a loose use of that word. These monthly data were calculated from the averages of daily data.

Stage 3. Melbourne, Daily Data for 150 days.

To reduce file sizes, only a few daily pairs were calculated. That is, I examined the correlation coefficient between pairs of days that followed each other (lag 1 day only). Because we have been using a series about 150 values in length, I selected a few strings about 150 days long. I repeated this procedure at several dates so we can see the variations. The chosen 150 day terms are all in the 1960-80 period because instrumental changes away from glass thermometers commenced in the late 1980s in Australia; and because the Australian temperature change was not so severe in the 1960-80 period. (I am trying to minimise spurious effects and to maximise the chance of hitting a high R).

Four of these exercises gave

Tmax (R)	Tmin (R)	
0.767	0.706	Started 10 Feb 1962
0.545	0.398	Started 5 Nov 1964
0.653	0.604	Started 5 Aug 1970
0.411	0.430	Started 1 Dec 1976

There are too few of these exercises here to find R>0.85, so the conclusion is that even adjacent days, taken over a 150 day term, correlate poorly but probably positively near to 0.5 on average.

Nothing done here so far points to R>0.85, the goal. Remember, all treatments so far are from the same location, Melbourne. The effect of distance separation cannot be expected to improve this situation.

The next logical steps are to examine (a. shorter data strings, below 150 values; and (b. sites other than Melbourne.

Stage 4. Shorter data strings.

The caption from the BEST diagram above mentions that their comparisons were done on years, where each string was more than 10 years long. Therefore, I took the Mebourne data later than year 1900 (to try for better quality, avoid Stevenson screen date complications, etc.) and calculated 10 strings each of 10 years, using the one year lag method only. That is, for 10 periods of a decade each, I looked at the correlation coefficients between one year and the next - using averaged annual data, not daily sampling.

Decade starting	Tmax R	Tmin R
1907	0.561497	-0.10998
1917	0.099624	0.034864
1927	-0.39736	0.30504
1937	0.065989	0.03306
1947	-0.22819	0.19784
1957	-0.07647	0.268644
1967	-0.00553	0.435924
1977	0.432846	-0.14064
1987	0.050154	0.032175
1997	-0.11212	0.117196

The conclusion is that in a single location (Melbourne), the correlation coefficients between temperature data from adjacent years, within a decade, have no systematic pattern.

I have shown at the beginning of this essay that the same data taken over a full 150 years does have some systematics. The question therefore arises as to the valid length of data required to produce credible correlations in adjacent years. Therefore, 144 consecutive years of data ending in 2005 were split into halves, then quarters, with this result.

TERM	144 yrs	1st 72 yrs	2nd 72 yrs	1st 36 yrs	2nd 36 yrs	3rd 36 yrs	4th 36 yrs
Tmax,R	0.4470	0.5771	0.3276	0.5822	0.6148	-0.0267	0.5166
Tmin, R	0.8371	0.4707	**0.8545**	0.3612	0.1630	0.4994	0.4459

The elusive R>0.85 appears in the second half of the Melbourne data in Tmin.

However, one asks how this happens, because the other intervals selected in that year appear to show no pattern. While R for Tmin is highest in this table in the period 1932 to 2005, this is the lowest (bar 1) period for Tmax. Systematic patterns are hard to find in all of the essay to date.

The question arises whether a correlation coefficient of this type returns the same answer from annually averaged inputs, as from daily inputs over the same term. Of course, it would not be the same answer because one set is lagged by a year and the other by a day.

Using the table above and selecting 4th 36 years (1968-2005),

Tmax using yearly averaged and annual lag 0.5166

Tmin using yearly averaged and annual lag 0.4459

T max using daily data and lagged by one year 0.5121

T max using daily data and lagged by one year 0.5464

T max using daily data and lagged by one day 0.7310

T min using daily data and lagged by one day 0.7706

……..

It is an obvious conclusion that the choice of sampling interval, be it daily, weekly, monthly or annual, has a dramatic effect of the calculations. Probably, it also follows logically that smoothing will have an effect; and that other mathematical manipulations such as detrending of cyclic data and infilling of missing data should be treated with tests of the types shown above.

Note also that in most instances, the correlation coefficient on Tmin is better than that on Tmax. There are probably physical reasons for this, but it would be idle to speculate when experiments can be performed. Another cautionary note is that the method of calculation of Tmean has changed over the years. In liquid-in-glass thermometer days, it was commonly Tmean = (Tmax+Tmin)/2. Since the early 1990s, many systems have been samples many times a day, allowing spike rejection and smoothing of select Tmax and Tmin from a curve. Given that spikes are probably more prevalent in day time than night time, spike rejection could be one reason for the discrepancy between Tmax and Tmin correlations.

Data strings that have been subjected to homogenisation, TOBS, Filnet and other manipulations should be used with extreme care.

There is a strong case to argue that the cluster of extreme values of high correlation with short distance separation, as on the graphs shown above, is simply an artifice. That is, when a correlation coefficient is calculated, some mathematical result is obtained and there are occasional chance times when a high result is obtained, even though it has little to nothing to do with the systematics of climate.

The results above also call for caution when investigating distance correlations at different latitudes. Since Tmax and Tmin seldom agree well in these examples above, at latitude 38 deg S, one wonders at the effect of polar days and nights that last half a year. The could be expected to be different to places near the Equator.

It is hard to argue that sampling at one location and comparing data as short as a day apart, should logically give lower correlation coefficients than two places separated by 1,000 km.

Yet, this result has been obtained here. Until it is explained, the data should not be used, especially for corrective or predictive purposes.

Recommendation.

The simple calculations shown above should be replicated many times so that an expected distribution of results for each small variation of technique is obtained. Remember that this example has dealt only with the city of Melbourne, which might have some atypical characteristics.
