Inference in Meteorological Data Taken from January 8, 2011 at Three Different Locations

Lorentz JÄNTSCHI\textsuperscript{1,2}

\textsuperscript{1}Chair of Chemistry, Faculty of Engineering of Materials and Environment, Technical University of Cluj-Napoca, 400641 Cluj
\textsuperscript{2}Chair of Horticulture and Landscape architecture, University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 400372 Cluj
email: lorii@academicdirect.org

Manuscript received February 05, 2011; revised March 15, 2011.

Abstract: In thee meteorological observation points were installed identical weather stations and a network infrastructure were created to receive and record the environmental parameters at every minute. A series of six out of forty-four parameters relating air, and solar radiation were taken into the analysis. The inference between observed data in three different locations was analyzed in order to find similarities between locations and/or observables. The analysis revealed important associations between environmental parameters and geographical location.

Keywords: environmental monitoring; environmental parameters; statistical inference; cluster analysis; multiple linear regression

Introduction

The analysis of the inference in meteorological data gives more and more interest due to the development of new forecasting models. Several approaches keep the front line. Thus, neural networks [1] are often used to the cases of incomplete data inference. When series of measurements of different nature are available, the inference in data is often searched using multiple linear regression and principal component analysis [2]. Humidity and temperature analysis provides insight into the potential health impacts of climate change [3]. Dew point analysis has important civil and thermal engineering applications, such as in design of dehumidification equipment [4].

Starting with an home-made equipment for acquisition and measurement of indirect and total solar radiation in 2007 [5] which provided important information concerning the recovering of local solar energy using thermal collectors [6], the environmental parameters acquisition system [7] were later extended with three commercial weather stations, which were able to provide data for analysis of: air movement [8], soil and leaf parameters [9], influence of environmental conditions on fruit growing [10], and apple scab attacks under conditions of excessive rainfalls [11].
In the present study, the inference between observed data in three different locations was analyzed in order to find similarities between locations and/or observables.

Material

Three Vantage Pro2 weather stations were placed and records data (44 observables) at every minute at three different locations.

The location of the observation points are given in Table 1.

<table>
<thead>
<tr>
<th>Place</th>
<th>Weather station</th>
<th>GPS</th>
<th>Elevation</th>
<th>Distance from ground</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reghin</td>
<td>st1</td>
<td>N 46° 46’ 12.41”; E 24° 41’ 27.99”</td>
<td>390m</td>
<td>1.5m</td>
</tr>
<tr>
<td>USAMV-CN</td>
<td>st2</td>
<td>N 46° 45’ 34.00”; E 23° 34’ 20.53”</td>
<td>381m</td>
<td>1.5m</td>
</tr>
<tr>
<td>UT-CN</td>
<td>st3</td>
<td>N 46° 47’ 45.40’”; E 23° 37’ 34.33’</td>
<td>326m</td>
<td>20m</td>
</tr>
</tbody>
</table>

Six observables were selected (recorded in every observation point in same time) for analysis, being given in Table 2.

Table 2: Environmental observables

<table>
<thead>
<tr>
<th>Observable</th>
<th>Meaning</th>
<th>Measurement unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>t_out</td>
<td>Outside temperature</td>
<td>°C</td>
</tr>
<tr>
<td>h_out</td>
<td>Outside relative humidity</td>
<td>%</td>
</tr>
<tr>
<td>t_dwp</td>
<td>Outside dew point temperature</td>
<td>°C</td>
</tr>
<tr>
<td>w_spe</td>
<td>Wind speed</td>
<td>ms⁻¹</td>
</tr>
<tr>
<td>p_bar</td>
<td>Barometric pressure</td>
<td>milibars</td>
</tr>
<tr>
<td>s_rad</td>
<td>Solar radiation</td>
<td>Wm⁻²</td>
</tr>
</tbody>
</table>

The data recorded from 10 to 10 minutes (144 records) were included into the analysis.

Methods

Cluster analysis [12] may be defined as a mathematical way of assignment of a set of observations into subsets (called clusters) so that observations belonging to same cluster are similar in some sense. The clustering problem has been addressed in many contexts and disciplines [13], having applications in meteorological control [14]. Cluster analysis was used in the present study to infer the environmental data from a contingency of six observables and three locations.

Dew point is an estimated variable according to Davis Inc [15] with an approximating formula recommended by WMO [16]. Using same notations as in Table 2, the estimated outside dew point temperature (tedwp) is - according to [15] - given by:

\[
\text{tedwp} = \frac{243.12 \cdot \ln(\text{vpv}) - 440.1}{19.43 - \ln(\text{vpv})}
\]

By using our range of observables values (t_out were from -2.1°C to 4.1°C; h_out were from 83% to 93%) a MathCad representation (Figure 1) shown that the tedwp have
a monotone approximately plane dependency, which allow us using of multiple linear regression to obtain simpler regression equations estimating $t_{dwp}$.

Figure 1: $tedwp(h_{out}, t_{out})$ variable plot ($h_{out}$: 80% to 100%; $t_{out}$: -3°C to 5°C)

Correlation analysis, initially introduced to measure the strength of linear dependence [17], was later extended to a more general case to infer monotone dependences [18]. Correlation analysis was used in the present study to compare the Dew point estimation method reported in [15] with recorded from weather stations values of the dew point.

Multiple linear regression uses various strategies [19] to minimize the disagreement between a set of observables [20] under assumption of linear dependence [17], having many environmental analysis applications [21]. Multiple linear regression were used in the present study to construct multiple linear relationship (MLR) models for dew point.

Results

The analysis of correlation between observed dew points and calculated ones according to [15] are given in Table 3. Table 3 contains the correlation analysis of all data (from all three weather stations, the data being taken pair by pair (432 pairs).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>0.9997</td>
<td>0.9990</td>
<td>0.9666</td>
<td>0.9861</td>
<td>0.9644</td>
<td>0.9982</td>
</tr>
<tr>
<td>Wrong model probability*</td>
<td>0.0e-1</td>
<td>0.0e-1</td>
<td>9.6e-1088</td>
<td>1.2e-1077</td>
<td>1.8e-1077</td>
<td>1.3e-1111</td>
</tr>
</tbody>
</table>

*from 'Student t' test

Correlation coefficients from Spearman to Gamma express different measures of monotone association (their definitions uses different manners of ties treatment).

A tree diagram for 6 (observables) times 3 (locations) using 144 observations were obtained when single linkage of Euclidian distances were calculated (Figures 2 to 4).
Figure 2: Linkage distance between observables \times locations

Figure 3: Zoom in 0 to 1000 range of Figure 1
Table 4 gives four multiple linear regression analyses between \( t\_dwp \) (as dependent variable) and \( t\_out \) and \( h\_out \) (as independent variables).

**Table 4:** Multiple linear regression analysis results for \( t\_dwp = \text{MLR}(t\_out, h\_out) \)

<table>
<thead>
<tr>
<th>Observations*</th>
<th>( t_dwp )</th>
<th>( t_2dwp )</th>
<th>( t_3dwp )</th>
<th>( t_dwp )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data: ( Y = t_dwp, X_1 = t_out, X_2 = h_out )</td>
<td>144</td>
<td>144</td>
<td>144</td>
<td>432</td>
</tr>
<tr>
<td>Model: ( Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 ); Assumption: ( Y ) is a normal estimates for ( Y ) (see [22] for details)</td>
<td>Intercept (-15.24 \notin [-15.61, -15.37] = CI(\text{Intercept})_{t_dwp} ); idem for ( t_3dwp ); ( t_2dwp )</td>
<td>Coef(( t_out )) ( t_1dwp = 0.9823 \notin [0.9769, 0.9854] = CI(\text{Coef}(( t_out )))_{t_dwp} ); idem for ( t_2dwp ) and ( t_3dwp ); ( h_out )</td>
<td>Coef(( t_out )) ( h_1dwp = 0.1534 \notin [0.1560, 0.1589] = CI(\text{Coef}(( t_out )))_{t_dwp} ); idem for ( h_3dwp ); ( h_2dwp )</td>
<td>Linear association measure: ( r = \text{r}(Y, \hat{Y}) ) (see [17] for details)</td>
</tr>
<tr>
<td>Pearson ( r )</td>
<td>0.9994</td>
<td>0.9997</td>
<td>0.9996</td>
<td>0.9997</td>
</tr>
<tr>
<td>Common sense statistics: confidence intervals for coefficients (see [22] for details)</td>
<td>CI(\text{Intercept})<em>{t_dwp} (-15.60, -14.88) \notin [-15.61, -15.37] = CI(\text{Intercept})</em>{t_dwp}; ( t_3dwp ); idem for ( t_dwp ); CI(\text{Coef}(( t_out )))<em>{t_dwp} ( t_1dwp = [0.9760, 0.9885] \notin [0.9707, 0.9830] = CI(\text{Coef}(( t_out )))</em>{t_dwp}; ( t_3dwp ); idem for ( t_dwp ); CI(\text{Coef}(( t_out )))<em>{t_dwp} ( h_1dwp = [0.1495, 0.1572] \notin [0.1548, 0.1575] = CI(\text{Coef}(( t_out )))</em>{t_dwp}; ( t_3dwp ); idem for ( h_dwp );</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* ( t_dwp ): from st1; ( t_2dwp ): from st2; ( t_3dwp ): from st3; ( t_dwp ): all together; CI: at 95% probability coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Followings can be observed in Table 4:

- \( \text{Intercept}_{t\_dwp} = -15.24 \notin [-15.61, -15.37] = \text{CI}(\text{Intercept})_{t\_dwp}; \) idem for \( t\_3dwp \); 
- \( \text{Coef}(t\_out)_{h\_1dwp} = 0.9823 \notin [0.9829, 0.9878] = \text{CI}(t\_out)_{h\_dwp}; \) idem for \( t\_2dwp \) and \( t\_3dwp \); 
- \( \text{Coef}(t\_out)_{h\_3dwp} = 0.1534 \notin [0.1548, 0.1575] = \text{CI}(t\_out)_{h\_dwp}; \) idem for \( h\_3dwp \); 
- \( \text{Coef}(t\_out)_{h\_2dwp} = 0.1534 \notin [0.1548, 0.1575] = \text{CI}(t\_out)_{h\_dwp}; \) idem for \( h\_3dwp \);
Discussions

Correlation analysis from Table 3 reveals that the reported (in [15]) formula of calculation doesn't "fit exactly" on the weather station given data. The main result of the analysis given in Table 3 is that a 'monotone association' is less likely than a particular case of it, 'linear association'. A better agreement was obtained when the data are correlated (0.9997) than their ranks are correlated (all others below 0.9990); this is not the expected result when a formula of calculation are applied, but on the contrary, when experimental error occurs [23]; thus, the only conclusion which can be drawn from here is that the reported formula (in [15]) and their implementation in the weather station device (or software) has some (minor) leaks.

The tree diagram from Figures 2-4 reveals the degree of the similarity between observables. The observed groups of data had shown that:

- Wind speed (due to it's stationery almost all the time behavior - no wind activity in over 75% of the cases) are one of the best group of relatives (clearly shown on Figure 4); inside this group the association between observations from Reghin (st1) and USAMV-CN (st2) is attributed to the relative altitude from ground of the observation points: both weather stations are at near to the ground level while UT-CN (st3) observation point is at about 20m from the ground, and thus is expected that wind activity to be somehow different - is almost twice much more wind activity (62 vs. 36 snapshots out of 144 with wind activity) and is over twice more intense (54.7 ms\(^{-1}\) vs. 20.2 ms\(^{-1}\) sums of wind speeds from 144 moments) at observation point st3 (Cluj-Napoca, 20m from ground) vs. observation point st2 (Cluj-Napoca, near to the ground); due to this fact we can draw the conclusion that wind speed observations array is a better location indicator in terms of the height from the ground position;

- Dew points from st2 and st3 observation points are relatives due to similar environmental conditions (temperature, relative humidity, barometric pressure) but are much relatives than any of them (Figures 3 and 4); this similarity should be assigned to the neighborhood of the observation points (placed at about 5.8 Km one to the other); due to this fact, we can draw the conclusion that dew point observations array is a better location indicator than others in terms of horizontal geographical position;

- The next cluster of interest groups wind speeds, outside temperatures and dew points (Figure 4); this is a surprising association, being known [16] that dew point is in relationship with temperature (t\(_{out}\)) and relative humidity (h\(_{out}\));

- Three other clusters groups the observables by their locations; by degree of association these groups are (Figure 3): Outside relative humidity (at linkage distance of about 26), Solar radiation (at linkage distance of about ten times higher), and Barometric pressure (at linkage distance of about eleven times higher);

The multiple linear regression analysis from Table 4 reveals that even if the models are high statistically significant, this is not a enough condition to be accepted as to be true - under assumption that the equation \(t_{dwp} = \text{MLR}(t_{out}, h_{out})\) exists, then must
be true when the model are feed with all possible data and should converge to it's true value when number of observations increases; more than that, outside of the 95% probability confidence interval should be only 5% of the possible cases, while analysis given in Table 4 reveals that in 7 out of 9 cases the coefficients are outside of the 95% confidence interval. This result should be correlated with the fact that the data are split according to their location of observation. Due to this fact we can draw the conclusion that dew point temperatures depends (in small amount) by at least one parameter more than temperature t_out and relative humidity h_out (such as wind speed, as Figure 4 strongly suggest).

Somebody may say: Why the linear regression was considered? What was the reason (only the simplicity)? - and we should raise an answer to these questions too. Indeed, the simplicity characterizes the linear regression. But more, as we depicted in Figure 3, for small ranges (as we have) of our observables (tedwp, h_out, t_out) values at least to a monotonic tedwp(h_out, t_out) function we should expect. More than that, correlation analysis from Table 3 reveals that 'monotone association' is less likely than a particular case of it, 'linear association'. Note that this it not means by necessity that the model is linear, it means only that the linear model is a 'approximating enough' model for the association in small ranges of the values of the observables and is according to our common manner of sin(Θ) ~ Θ approximations when Θ are small enough. Going further with this reasoning, a polynomial Ŷ=f(X,Z) of a given range is expected to increase the accuracy of the estimates Ŷ, but not of the confidence of the coefficients of the model too.

Another point of view expecting an answer may be: May down-sampling from one minute to ten minutes to produce loosing of information? - Always down-sampling produces the loose of the information. Thus, the question should be reformulated: the loose of the information may affect the interpretation? - And here the answer is - definitely no. When a model cannot be rejected (as the proof given in Table 4) the increasing of the number of observations (let's say ten times more observations) the only expectation which may have is to obtain even smaller confidence intervals for same probability coverage (in an exact ratio of square root of ten smaller, given by the expression of standard error of the sampling as function of population standard deviation and the sample size). Thus, the increasing of the sample size is expected not to reject the observations drawn from Table 4 (Interceptt1dwp ∉ Cl95%(Intercept)t_dwp and the following ones), but on contrary, to even strongly proof it.

The present study focused on the inference of a small group of six out of forty-four weather stations provided parameters (more exactly six out of twenty-two weather related parameters). The selection of the five out of six parameters (outside dew point temperature entering by default into the analysis) was based on physical reasoning - the known relationship giving an approximation ('approximating enough model') of dew point temperature as function of temperature and relative humidity as well as other three parameters characterizing three physical phenomena which may infer the occurrence of the dew point - solar radiation, air pressure and air movement. The study revealed (see Figure 4) that the most dew point estimates affecting parameter from all three is wind speed. Thus, the study reveals that a better estimating accuracy for dew point should be obtained when the model of the estimate it include together with temperature and relative humidity the wind speed too.
Conclusions

Cluster analysis of the data recorded on January 8, 2011 in three different locations revealed that:

- Wind speed observations array is a very good location indicator in terms of terms of the height from the ground position.
- Dew point observations array is a very good location indicator in terms of horizontal geographical position.
- Dew point temperature model depending on air temperature and relative humidity proved to have not enough accuracy when location are changed, and at least one more parameter (such as wind speed, as present study shown) are necessary in the equation to correct it's accuracy.

References