Estimating ground temperature and humidity using geographically weighted regression

Nguyen Pham Viet Nam Dao Tuan Linh Bui Duc Anh Dao Xuan Loc

Abstract *Ground level measurement is necessarily scarce due to constraints in space and cost, while satellite data can cover a large space with high resolution. This report aims to illustrate the usage of geographically weighted regression (GWR) in estimating temperature and humidity from satellite-derived products. The GWR model developed showed good agreement between predicted values and ground measurements at 102 weather stations in Vietnam for both temperature and humidity.*

1 Introduction

Temperature and humidity data are very important for monitoring the environment: they are useful for weather forecast, climate change alerts, plant development and other agriculture related predictions. The most reliable way of obtaining these data is to measure them directly at the point of interest, however, it is not always easy or cheap to set up and maintain these ground stations. Due to the cost constraints and the fact that some places can be unstable or otherwise unsuitable for weather station, it is of great interest to make use of the freely available satellite-derived products, which is also of high resolution with wide coverage. In this report, we make use of data derived from MODIS sensors on Aqua and Terra satellite overhead Vietnam to calculate temperature and humidity data at places where ground stations are not available.

Temperature and humidity data depends heavily on the season, the time measurements were taken, and the positions of the place as well - we would expect coastal area to have very different measurements from mountainous area, for example.

In this study, we make use of GWR method, in particular *spgwr* package in R [1], which takes into account this fact. The software calculates different sets of coefficients for each measured points, portraying a spatially varying relationship.

2 Related work

Multiple efforts were made to make use of sparse ground station measurements to infer temperature of places without such stations, for example interpolation based method such as inverse distance weighting [2, 3], or Cokriging method [4].

There are also artificial neural network based method, which acts as a non-linear interpolation tool. The problem with this method is usually discontinuity of predicted measurements [5, 6].

3 Methodology

3.1 Study area and datasets

We focus on Vietnam territory, using preprocessed MODIS level 2 products in January 2014, latitude ranging from 6.4 to 25.6, longitude 100.1 to 111.8 with 3km resolution. Independent variables were generated from MOD06, MYD06, MOD07, MYD07, VIIRS for temperature; MOD06, MOD07, MYD06, MYD07 for pressure, and MOD05, MOD07, MYD05, MYD07 for water vapor. Temperature and humidity data from 102 weather stations in Vietnam collected daily at 6AM during the same time period were used as groundtruth. Figure 1 shows the area of study with ground stations marked with black dots.

3.2 Retrieval method

The preprocessed MODIS images were collected at different time throughout the day, and they do not always cover the whole study area. We generated independent variable data by averaging images of the same product from the same day, pixel by pixel (ignoring missing data) - see figure 2 for a sample. Pixel values corresponding to ground station positions are then extracted from these averaged images and written to a file for later processing.

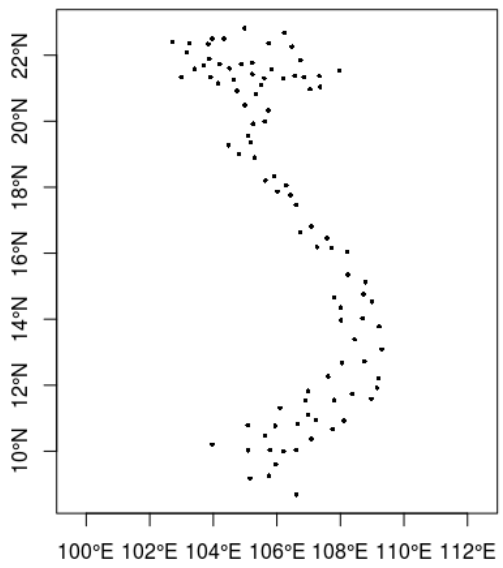

Fig. 1: Ground stations distribution

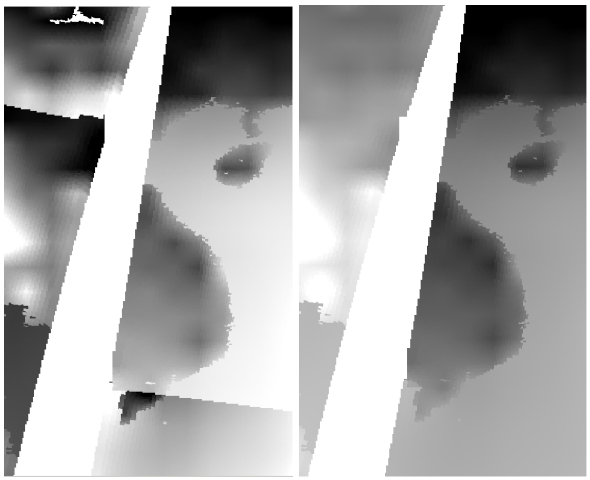

Fig. 2: MOD06 data (temperature). Left: original 4 images from the same day stacked together, right: averaged image

3.3 Geographically weighted regression

GWR, unlike normal regression, does not assume stationary process, or parameters being constant over space. Any spatial variations in the processes can only be measured by the error term. This problem leads to the central idea behind GWR: by allowing the relationship we are measuring to vary over space, we can capture nonstationarity directly. The equation is

$$y(g) = \beta_0(g) + \beta_1(g)x_1 + \beta_2(g)x_2 + ... + \beta_n(g)x_n + \varepsilon$$

where g refers to the location at which the estimates of the parameters are obtained, and the estimator is

$$\hat{\beta} = (X^T W(g) X)^{-1} X^T W(g) Y$$

where $W(g)$ is a matrix of weights specific to location g such that observations nearer to g are given more weights than observations further away.

There are multiple choices of weighting function. The one we used later for this study is simple Gaussian weighting scheme, bisquare and tricube functions.

The default *gwr.gauss* function is

$$w(g) = e^{-(d/h)^2}$$

where h is the bandwidth.

Bisquare function is

$$w_{ij}(g) = (1 - d_{ij}^2/d^2)^2$$

if $d_{ij} <= d$ else $w_{ij}(g) = 0$, where d is the bandwidth.

And tricube function looks similar to bisquare function

$$w_{ij}(g) = (1 - (d_{ij}/d)^3)^3$$

if $d_{ij} <= d$ else $w_{ij}(g) = 0$, where d is the bandwidth.

The bandwidth is an important parameter and can be chosen using function *gwr.sel*, which performs drop-1 cross validation to minimize the RMS prediction error. The bandwidth can be fixed or adaptive, with the fixed version being what the name suggests and adaptive version means using some percentage of number of closest neighbors to perform prediction.

3.4 Model evaluation

To estimate temperature, data from cloud product and atmospheric profiles (Aqua and Terra), as well as from VIIRS were averaged (ignoring missing values) to produce the only independent variable, which is surface temperature as observed with remote sensor.

To estimate humidity, data from cloud product and atmospheric profiles (Aqua and Terra) were also averaged to generate the first independent variable, average

observed pressure. Data from cloud product and total precipitable water were averaged to generate the second independent variable, average observed water vapor. In addition to evaluating different weighting functions and choice of bandwidth, we also test different formulas for the regression equation: humidity depending on water vapor alone, pressure alone, and both of them at once.

All bandwidths (fixed or otherwise) were chosen by *gwr.sel* function in *spgwr* package, which performs leave one out cross validation to choose the optimal value.

We performed 5-fold cross validation for the collected data for each day, and then calculated RMSE, R2 and RE. The folds were created by stratified sampling over latitude (since the study area varies more significantly over latitude than longitude) to ensure a somewhat representative set of test data for each fold. Another reason, perhaps more important, is that GWR would not be able to handle regions in which it has no data to learn the coefficients from.

The mean RMSE, median R2 and mean RE (%) across days in the dataset is reported in table 1 for temperature and table 2 for humidity.

RMSE is calculated by

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

where $e_t = y_t - \hat{y}_t$

R2 is calculated by

$$R2 = 1 - \frac{\sum_{t=1}^{n} e_t^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

where $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ and RE is

$$RE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right|$$

4 Experimental results

From cross validation result (table 1), we can see that for temperature model, the choice of weighting function and whether bandwidth is adaptive do not significantly affect the prediction result. Both the mean and variance of these statistics do not shift by any noticeable amount when we change variants of GWR model. They also do tend to perform similarly on the same day (compare figure 3 and 4 for an example). Since adaptive bandwidth with Gauss weighting

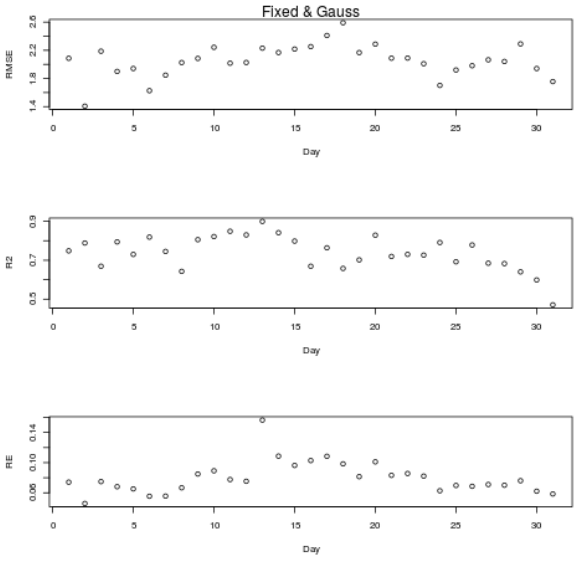

Fig. 3: Error for 5-fold cross validation in temperature model fixed bandwidth, weighting function Gauss

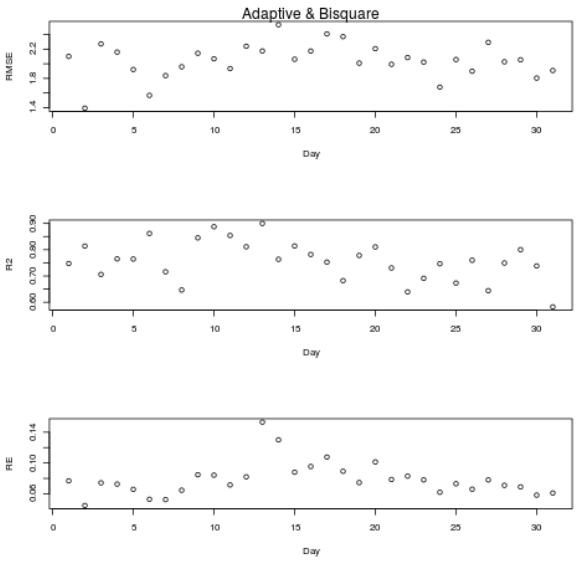

Fig. 4: Error for 5-fold cross validation in temperature model adaptive bandwidth, weighting function Bisquare

scheme works slightly better than the rest, it is the variant we chose to report and do further experiment on. Table 3 shows average test error statistics across 5 folds for each day, and table 4 reports error statistics for all data of each day.

Similarly, for humidity model, the choice of weighting function, whether bandwidth is adaptive and

Adapt	Weight	RMSE	R2	RE
Adaptive	Gauss	2.035	0.767	7.861
Fixed	Gauss	2.077	0.763	8.069
Adaptive	Bisquare	2.048	0.772	7.963
Fixed	Bisquare	2.069	0.764	8.030
Adaptive	Tricube	2.063	0.758	7.989
Fixed	Tricube	2.132	0.742	8.133

Table 1: Average dev error metrics across days for temperature model evaluation (5-fold cross validation)

even regression formula do not affect the result very significantly. We chose to proceed with regression formula of only one dependent variable average pressure, and Gauss weighting function with adaptive bandwidth because it performs better than the rest and has low variance in RMSE across days, suggesting that its performance is quite stable. Table 5 reports average test error statistics across 5 folds each day for this variant, and table 6 reports error of the model fitted on all data of each day.

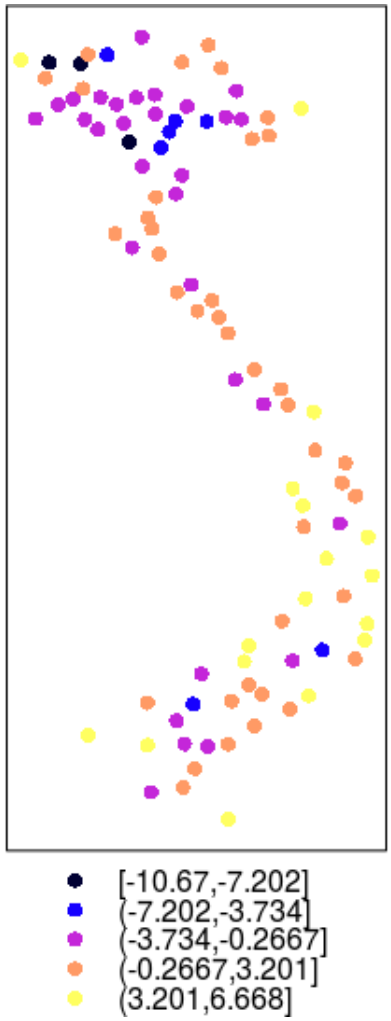

Fig. 5: Prediction error of one randomly sampled day for temperature, normal regression model (best viewed in color)

5 Discussions

We compared ordinary regression model with GWR method on a randomly chosen day and observed an increase in R2 score (temperature: 0.682 to 0.865, humidity: -0.00083 to 0.822), which is expected because GWR, being a local regression method (in geographic space instead of parameter space like normal local regression), provides more flexibility to fit data around training points.

Figure 8 shows the residuals at each station on January 14th. There is no obvious spatial pattern to the error, unlike the normal regression model shown in figure 7, where the regression model systematically underestimates humidity for the north, and overestimates it for the middle part of Vietnam. Looking at figure 10, we see that there does not seem to be any obvious trend in residual vs fitted plot, even though QQ plot seems to deviate from the ideal straight line shape. Its pattern suggests that the central quantiles are more closely spaced for residuals than for predicted, and that they might not have come from the same distribution.

Comparing figure 5 and 6, we observed the same pattern as previously mentioned in humidity model: the normal regression model underestimates the northern part's temperature more and overestimates the south's more, but in GWR model this pattern disappears. For temperature model, even though it has better performance than humidity model (comparing only RE and R2, not RMSE), a closer look at error plots (figure 9) revealed some non-linearities in the relationship between residual and fitted values: QQ plot is closer to an S-shape than the ideal straight line, and residual vs fitted plot is not randomly distributed around 0. There seems to be a lack of homoscedasticity: higher fitted values seem to have lower residuals, but this might be an artifact of the low number of data points. One potential source of complications for our models could be the time difference between ground measurement and satellite data collection time, since humidity and tem-

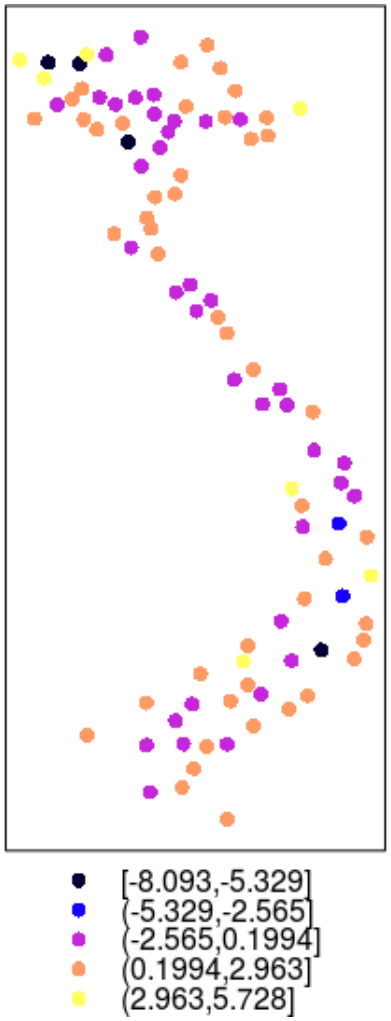

Fig. 6: Prediction error of the same day for temperature, GWR model (best viewed in color)

perature are highly variable throughout the day.

We used the whole image created as per section 3.2 and the best model as discussed in the previous section to extrapolate temperature data (figure 11) and humidity data (figure 12) for the whole region. We expected that the further away from original fit points the pixels are, the larger the error, but from this simple interpolation we can already observe some nice, intuitive trends: the fact that the north is on the whole colder than the south and drier regions are further away from the sea.

6 Conclusions

In this report, we have explored GWR method and applied it to the calculation of temperature and humidity in places without ground stations using satellite-derived products, which can be very cost-effective. These are important basic measures, useful for weather forecasting and various other research efforts. Even though our models might need further refinement, GWR is a promising method providing adequate explanation to spatially varying relationship. Future work could include more sophisticated data preprocessing methods and other spatial weighting schemes.

Var	Adapt	Weight	RMSE	R2	RE
Vapor	Adaptive	Gauss	8.475	0.463	10.837
Vapor	Fixed	Gauss	8.465	0.474	10.735
Vapor	Adaptive	Bisquare	8.476	0.473	10.796
Vapor	Fixed	Bisquare	8.699	0.446	11.040
Vapor	Adaptive	Tricube	8.551	0.471	10.845
Vapor	Fixed	Tricube	8.571	0.474	10.959
Press.	Adaptive	Gauss	8.146	0.506	10.491
Press.	Fixed	Gauss	8.580	0.484	10.706
Press.	Adaptive	Bisquare	8.317	0.490	10.523
Press.	Fixed	Bisquare	9.522	0.434	11.567
Press.	Adaptive	Tricube	8.469	0.479	10.734
Press.	Fixed	Tricube	9.455	0.427	11.494
Both	Adaptive	Gauss	8.379	0.472	10.800
Both	Fixed	Gauss	8.779	0.478	11.084
Both	Adaptive	Bisquare	8.471	0.477	10.810
Both	Fixed	Bisquare	9.080	0.455	11.296
Both	Adaptive	Tricube	8.459	0.479	10.823
Both	Fixed	Tricube	9.966	0.444	11.759

Table 2: Average dev error metrics across days for humidity model evaluation (5-fold cross validation)

References

[1] Bivand, R., 2017. *Package spgwr*. PDF file https://cran.r-project.org/web/packages/spgwr/spgwr.pdf.

[2] Willmott, C., Robeson, S., and Philpot, W., 1985. "Small-scale climate maps: A sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring". *Am. Cartographer*.

Day	RMSE	R2	RE
1	2.074	0.7208	7.629
2	1.395	0.8191	4.612
3	2.096	0.7208	6.957
4	2.046	0.7508	6.843
5	1.968	0.7414	6.707
6	1.61	0.8415	5.591
7	1.81	0.7605	5.444
8	1.912	0.6865	6.334
9	1.996	0.8784	7.793
10	2.125	0.8481	8.579
11	2.051	0.8761	7.926
12	1.994	0.8654	7.321
13	2.35	0.8645	15.43
14	2.251	0.8625	11.4
15	2.093	0.8159	8.871
16	2.356	0.7499	9.944
17	2.607	0.701	11.4
18	2.113	0.8044	8.331
19	1.985	0.8179	7.066
20	2.42	0.7853	11.01
21	2.185	0.7873	8.761
22	2.073	0.7331	8.254
23	1.969	0.6919	7.824
24	1.706	0.7759	6.251
25	1.871	0.8383	6.818
26	1.897	0.804	6.875
27	2.105	0.6318	7.26
28	2.047	0.7093	6.975
29	2.244	0.6754	7.353
30	1.976	0.5657	6.379
31	1.766	0.6491	5.773
Mean	2.035	0.767	7.861

Table 3: 5 fold cross validation for temperature GWR model with Gauss weighting scheme and adaptive bandwidth

Day	RMSE	R2	RE
1	1.692	0.8516	5.81
2	1.091	0.8843	3.525
3	1.825	0.8104	5.85
4	1.804	0.8081	6.234
5	1.733	0.7873	5.621
6	1.211	0.9137	4.015
7	1.232	0.874	3.568
8	1.148	0.9014	3.464
9	1.622	0.9158	6.367
10	1.67	0.9141	6.267
11	1.408	0.9305	5.094
12	1.687	0.9002	6.249
13	2.036	0.9127	13.66
14	1.784	0.8953	8.411
15	1.613	0.8915	6.656
16	1.611	0.875	6.491
17	2.232	0.7924	9.843
18	1.336	0.9046	4.474
19	1.461	0.8878	4.962
20	1.905	0.8647	8.186
21	1.557	0.866	5.81
22	1.589	0.8409	5.679
23	1.661	0.8147	6.125
24	1.357	0.8499	4.779
25	1.278	0.8875	4.394
26	1.548	0.8576	5.071
27	1.753	0.7954	5.742
28	1.664	0.8185	5.38
29	1.763	0.8208	5.589
30	1.705	0.6898	5.177
31	1.533	0.696	4.926
Mean	1.597	0.853	5.917

Table 4: Temperature model error on each day's dataset. Gauss weighting scheme, adaptive bandwidth

Day	RMSE	R2	RE
1	6.825	0.5178	9.22
2	5.735	0.4411	7.782
3	9.529	0.263	11.25
4	7.424	0.5609	9.096
5	8.339	0.172	10.24
6	6.346	0.6684	7.468
7	8.583	0.4281	11.22
8	9.143	0.4341	12.33
9	9.338	0.4098	10.92
10	7.418	0.6913	8.189
11	8.173	0.7004	9.327
12	10.02	0.4648	11.36
13	8.981	0.5515	10.12
14	8.521	0.6283	11.03
15	7.483	0.5159	10.78
16	7.346	0.4975	8.956
17	9.69	0.3363	10.84
18	7.551	0.6834	10.66
19	7.938	0.5984	10.49
20	9.36	0.6385	11.78
21	6.993	0.7555	10.81
22	6.693	0.702	10.76
23	8.266	0.4966	12.92
24	7.344	0.3248	10.86
25	7.374	0.5848	9.511
26	8.009	0.6101	9.527
27	8.858	0.3744	11.48
28	8.027	0.4773	9.97
29	10.1	0.532	12.75
30	8.107	0.3152	10.92
31	9.031	0.2994	12.65
Mean	8.146	0.506	10.491

Table 5: 5 fold cross validation for humidity GWR model with Gauss weighting scheme and adaptive bandwidth

Day	RMSE	R2	RE
1	5.008	0.7389	6.694
2	4.182	0.6849	5.78
3	8.171	0.4307	8.731
4	5.426	0.7925	6.662
5	5.818	0.636	6.42
6	4.155	0.8844	4.832
7	6.666	0.7258	8.122
8	9.027	0.4949	12.03
9	7.23	0.6855	8.27
10	5.687	0.7853	5.988
11	5.96	0.8276	6.62
12	5.43	0.8125	6.083
13	6.841	0.7722	7.682
14	6.473	0.8015	8.206
15	5.671	0.7319	7.907
16	5.391	0.775	6.384
17	7.093	0.6997	7.798
18	5.769	0.8363	7.85
19	5.563	0.8044	7.027
20	6.862	0.7978	8.211
21	5.731	0.8256	8.625
22	5.055	0.8218	8.143
23	6.396	0.7174	9.832
24	5.826	0.6326	8.423
25	5.563	0.7764	6.902
26	4.884	0.8566	5.436
27	6.463	0.6892	8.415
28	4.421	0.8323	5.196
29	8.986	0.5633	11.51
30	4.401	0.8164	5.762
31	6.31	0.6832	8.245
Mean	6.01	0.74	7.54

Table 6: Humidity model error on each day's dataset. Independent variable: pressure, weighting scheme: Gauss, adaptive bandwidth

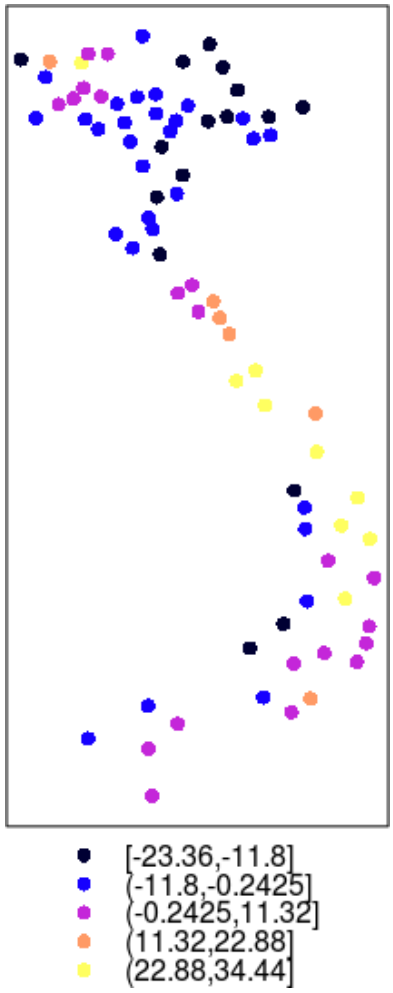

Fig. 7: Prediction error of one randomly sampled day for humidity, normal regression model (best viewed in color)

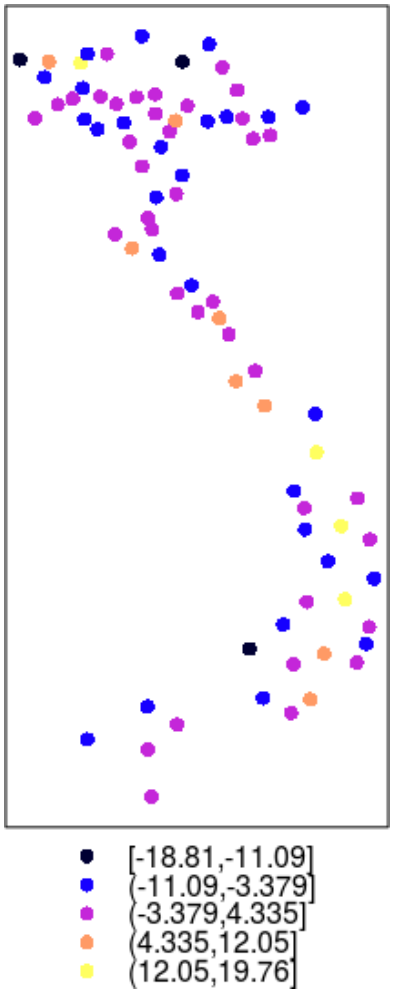

Fig. 8: Prediction error of the same day for humidity, GWR model (best viewed in color)

[3] J. S. Yang, Y. Q. W., and August, P. V., 2004. "Estimation of land surface temperature using spatial interpolation and satellite-derived surface emissivity". *Journal of Environmental Informatics*.

[4] Ishida, T., and Kawashima, S., 1992. "Use of cokriging to estimate surface air temperature from elevation". *Theor. Appl. Climatol.*

[5] Dolling, O., and Varas, E., 2002. "Artificial neural networks for streamflow prediction". *Journal of Hydraulic Research*.

[6] Flores, P. F., and Lillo, S. M., 2010. "Simple air temperature estimation method from modis satellite images on a regional scale". *Chilean Journal of Agricultural Research*.

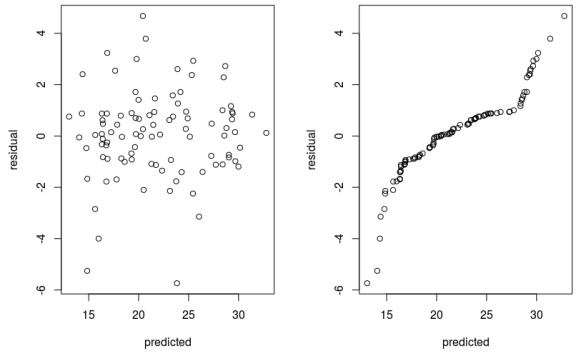

Fig. 9: Residual vs fitted (left) and QQ plot (right) of one randomly sampled day for temperature

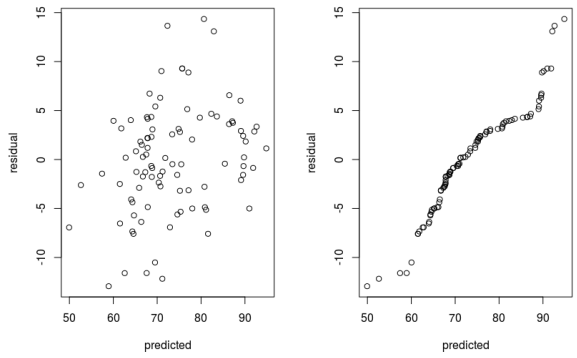

Fig. 10: Residual vs fitted (left) and QQ plot (right) of one randomly sampled day for humidity

Fig. 11: New temperature plot generated from the best model for January 12th, 2014 (best viewed in color), warmer color is for higher temperature. The trend for colder northern part and warmer southern part is clear

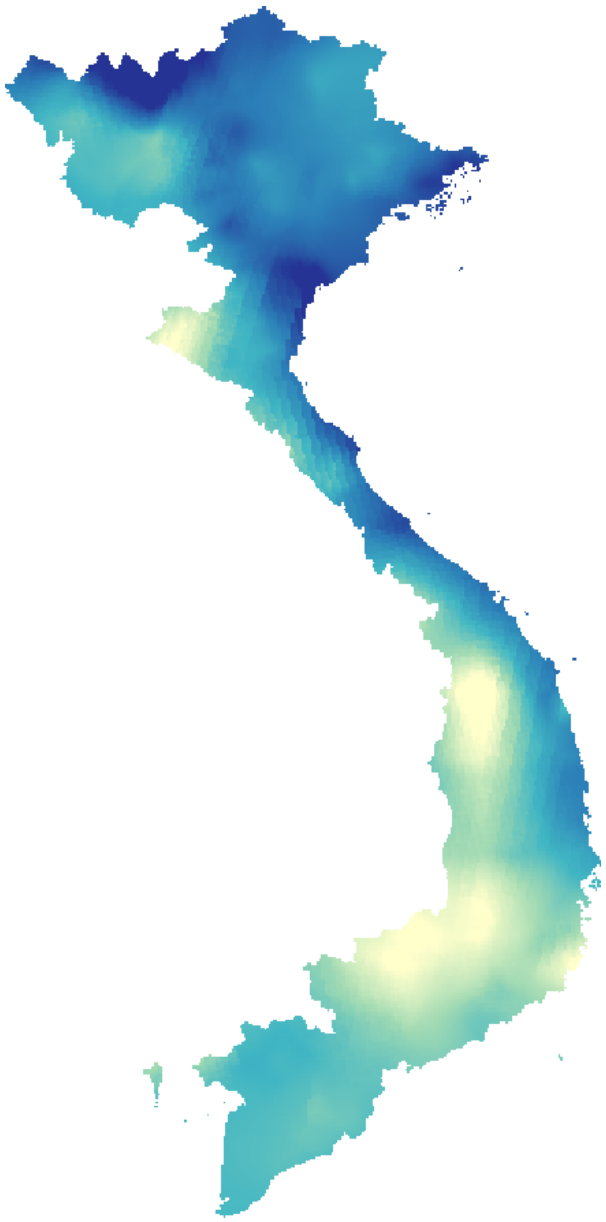

Fig. 12: New humidity plot generated from the best model for January 28th, 2014. Colder color is for more humid region (best viewed in color). Humid regions tend to cluster around the coast