LINEAR MODELS OF DIFFERENT SCALES

Xian He[1,*], Xudong Zhu[2], Hao Zhang[1] & Qianlai Zhuang[2]
[1]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
[2]Department of Earth and Atmospheric and Planetary Sciences, Purdue University, West Lafayette, IN 47907, USA

ABSTRACT

Many statistical models for environmental studies can be run at different scales, e.g., for daily, weekly or monthly data. It is important to know how these models differ in terms of prediction. We provide some theoretical results to compare these models. We show that when there is no high order terms of explanatory variables, the small scale model yields more efficient estimators for model parameters and also produces better prediction. However, when there are high order terms of explanatory variables, a larger scale model can be run in two different ways: product-of-sum or sum-of-product. Current practice made it hard to compare directly the larger scale model with the small scale model and no explicit conclusions are drawn in general. We provide a case study on gross primary production of terrestrial ecosystems in the conterminous United States to demonstrate our results.

Key-words: *Gauss-Markov theorem, gross primary production, multiple linear regression, prediction, best linear unbiased estimator.*

1. INTRODUCTION

Many statistical models for environmental studies can be run at different scales, e.g., for daily, weekly or monthly data [1-3]. It is important to know when and how these models of different scales differ. Although there are some empirical studies on models of different scales [4-6], there is a lack of theoretical discussion and explicit conclusions on the scaling problem.

In many environmental studies, choosing a suitable temporal scale (e.g, hourly, daily, weekly or monthly) is one of the most important steps. With the improvement of remote sensing technology, it is feasible to acquire data at various spatial and temporal resolutions. We can therefore run a model at a larger scale or run it at a finer scale and then upscale the results. How would the results differ? This work is an attempt to answer the question.

We provide some theoretical results to show that when there is no high order terms of explanatory variables in a regression model, the small scale model yields more efficient estimators for model parameters and also produces better prediction. However, when there are high order terms of explanatory variables, a larger scale model can be run in two different ways. Current practice made it hard to compare directly the larger scale model with the small scale model and no explicit conclusions are drawn in general. We demonstrate our results through a case study on the gross primary production.

This paper is organized as follows. In Section 2, we theoretically compared the different temporal scale models in terms of prediction accuracy and efficiency. In Section 3, we present a case study on the gross primary production (GPP) [1,7] where we run and compare the models at three scales.

2. MODELS FOR DIFFERENT SCALES

2.1 The Scaling Issues

Suppose Y is the response variable to be regressed on $p-1$ explanatory variables $x_1,\ldots,x_{p-1}$. Each of the variables is observed at time points $t=1,\ldots,n$, say daily. The linear regression model becomes

$$y_t = 1+\sum_{i=1}^{p-1} x_{t,i}\beta_i + \varepsilon_t, t=1,\ldots,n, \tag{1}$$

where the error terms ε_t are assumed to be i.i.d. $N(0,\sigma^2)$.

However, there are situations when the model is applied at a larger scale, say, weekly. The aggregated variables $y_t^{(w)} = \sum_{i=1}^{s} y_{s(t-1)+i}, x_{t,k}^{(w)} = \sum_{i=1}^{s} x_{s(t-1)+i,k}, k=1,\ldots,p-1$ are used in the regression, where s denotes the time units the variables are aggregated upon (e.g., $s=7$ for the weekly scale). The model becomes

$$y_t^{(w)} = s+\sum_{i=1}^{p-1} x_{t,i}^{(w)}\beta_i + \varepsilon_t^{(w)}, t=1,\ldots,m, \tag{2}$$

The two models share the same linear parameters $\beta=[\beta_0,\beta_1,...,\beta_{p-1}]'$, but the error terms in (2) has a larger variance than (1). In addition, there are fewer observations for the larger scale model (2). Hereafter, we assume that $n=ms$.

The two central questions this work is concerned of are the following. First, how do the two scales affect the estimation of the parameters β_i and the variance σ^2? Second, how do the scales affect the prediction? More specifically, suppose we like to predict $y_{m+1}^{(w)}$. We can obtain this prediction from both models. How different would these two predictions be?

2.2 Theoretical Results

In this section, we provide some theoretical results that allow us to draw some explicit conclusions. Denote by $\hat{\beta}$ and $\hat{\sigma}^2$ the least squares estimators of β and σ^2, respectively, which are obtained by fitting model (1), and by $\hat{\beta}^{(w)}$ and $\hat{\sigma}^{2(w)}$ the least squares estimators according to model (2). If we denote by X the design matrix in model (1) and by Y the vector of response variable, then

$$\hat{\beta}=(X'X)^{(-1)}X'Y, \hat{\sigma}^2=\|Y-X\hat{\beta}\|^2/(n-p).$$

The design matrix $X^{(w)}$ and the vector aggregated response variable $Y^{(w)}$ are related to X and Y in the following way

$$X^{(w)}=JX,\ Y^{(w)}=JY,$$

where $J=I_m\otimes 1_s$ is an $m\times n$ matrix where I_m is an $m\times m$ identity matrix and 1_s is an s-dimension vector of all 1s. The estimates from model (2) can be written

$$\hat{\beta}^{(w)}=(X^{(w)'}X^{(w)})^{(-1)}X^{(w)'}Y^{(w)}, \hat{\sigma}^{2(w)}=s\|Y^{(w)}-X^{(w)}\hat{\beta}^{(w)}\|^2/(m-p)$$

where s is the period of time units the large scale is aggregated upon.

The following proposition says that the smaller scale model yields more efficient estimators than the larger scale model.

Proposition 2.1 *Observing $y_1,\ldots,y_n$ with $n=ms$, the estimators given through the vector of response variable of two models (1) and (2) have the following properties.*

- Both $\hat{\beta}^{(w)}$ and $\hat{\beta}$ are unbiased estimators of β but the former is more efficient, i.e.,

$$E(\hat{\beta}^{(w)})=E(\hat{\beta})=\beta,$$

and $Var(\hat{\beta}^{(w)})-Var(\hat{\beta})$ is positive semi-definite.

- Both $\hat{\sigma}^2$ and $\hat{\sigma}^{2(w)}$ are unbiased estimators of σ^2. In addition,

$$Var(\hat{\sigma}^2)=\frac{2\sigma^4}{n-p}, Var(\hat{\sigma}^{2(w)})=\frac{2\sigma^4}{m-p}.$$

Hence $\hat{\sigma}^2$ is more efficient.

The Proposition readily follows the Gauss-Markov theorem [?]. We only sketch the proof here. It is obvious that both $\hat{\beta}^{(w)}$ and $\hat{\beta}$ are unbiased. Since $\hat{\beta}^{(w)}$ is a linear unbiased estimator, the Gauss-Markov theorem implies that $\hat{\beta}$ is more efficient than $\hat{\beta}^{(w)}$. It is well-known that $\|Y-X\hat{\beta}\|^2/\sigma^2$ has a χ^2-distribution with $(n-p)$ degrees of freedom. It follows that

$$Var(\hat{\sigma}^2)=\frac{2\sigma^4}{n-p}.$$

Indeed, the above can be found in classical textbooks on regression. Similarly, because

$$\frac{\| Y^{(w)} - X^{(w)} \hat{\beta}^{(w)} \|^2}{\sigma^4 / s}$$

has a χ^2-distribution with $(m-p)$ degrees of freedom with a variance $2(m-p)$, it follows that

$$Var(\hat{\sigma}^{2(w)}) = \frac{2\sigma^4}{m-p}.$$

Next, we consider the effects of scales on prediction. If we observe the explanatory variables at s consecutive time points, $n+1,\ldots,n+s$, and want to make a prediction of the aggregated response variable $y^{(w)}$, we could obtain the prediction in two ways, using the two models (1) and (2). The explanatory variables for the larger scale model is $x_{m+1}^{(w)} = \sum_{t=n+1}^{n+s} x_i$, where x_i is the vector of explanatory variables at time i. We could get the prediction of y from the two different temporal scale models as follows:

$$\hat{Y} = \sum_{i=n+1}^{n+s} x_{i'} \hat{\beta} = x^{(w)'} \hat{\beta}. \tag{3}$$

$$\hat{Y}^{(w)} = x^{(w)'} \hat{\beta}^{(w)}. \tag{4}$$

The comparison of the two predictions is given in the following proposition.

Proposition 2.2 *Under the formulation of models (1) and (2), the two predictors (3) and (4) have the following properties:*

- $E(\hat{Y}) = E(\hat{Y}^{(w)}) = x^{(w)'} \beta.$
- $Var(\hat{Y}^{(w)}) \geq Var(\hat{Y}).$

This proposition follows from the unbiasedness of $\hat{\beta}$ and $\hat{\beta}^{(w)}$, and the fact that $\hat{\beta}$ is the best unbiased linear estimator of β. Indeed, the Gauss-Markov theorem implies that for any vector x,

$$Var(x' \hat{\beta}^{(w)}) \geq Var(x' \hat{\beta}).$$

2.3 Scaling Issues with Polynomial Regression

In this section, we consider the scaling issue in the polynomial regression. What complicates in this case is that there are two possible ways to run the model at the larger scale. Suppose the regression model at the smaller scale is

$$y_t = \beta_0 + \sum_{i=1}^{p-1} x_{t,i}\beta_i + \sum_{(i,j)\in\Delta} x_{t,i}x_{t,j}\beta_{ij} + \varepsilon_t, t = 1,\ldots,n, \tag{5}$$

where Δ is an index set for the high order term. For example, $\Delta = \{(i,j), i,j = 1,\ldots,p-1, i \neq j\}$ if all second order terms are included in the model.

One way to formulate the larger scale model is to aggregate all variables as in model (2)

$$y_t^{(w)} = s\beta_0 + \sum_{i=1}^{p-1} x_{t,i}^{(w)}\beta_i + \sum_{(i,j)\in\Delta} x_{t,ij}^{(w)}\beta_{ij} + \varepsilon_t^{(w)}, t = 1,\ldots,m, \tag{6}$$

where $y_t^{(w)}$ and $x_{t,i}^{(w)}$ are defined the same as in (2), $x_{t,ij}^{(w)} = \sum_{k=1}^{s} x_{s(t-1)+k,i}\, x_{s(t-1)+k,j}$ is the aggregated cross product $x_{t,i}x_{t,j}$. Comparison between models (5) and (6) follows the discussion in the previous section. We can say that model (5) at the smaller scale results in more efficient estimation and better prediction.

In practice, however, the larger scale model is often run as follows.

$$y_t^{(w)} = \beta_0^{(w)} + \sum_{i=1}^{p-1} x_{t,i}^{(w)}\beta_i^{(w)} + \sum_{(i,j)\in\Delta} x_{t,i}^{(w)}x_{t,j}^{(w)}\beta_{ij}^{(w)} + \varepsilon_t^{(w)}, t = 1,\ldots,m, \tag{7}$$

where $x_{t,i}^{(w)}$ is same as defined previously. The high order terms are now aggregated differently. The larger scale model (7) and the small scale model (5) have different sets of parameters. Therefore, unlike in the previous section, a direct comparison between the two models is difficult if not impossible. For example, it does not make sense to compare the efficiency of estimators because the parameters in the two models are different. Similarly, for prediction, the two models assumed different expected value to start with. Therefore, the two models may yield different prediction results.

The example in the next section reveals that the predicted value given by the larger scale model may be either smaller or larger than that given by the smaller scale model.

3. AN EXAMPLE

In this section, we consider an example of real data set, which motivated this work and also helps to show the difference the scales can make to statistical inferences. The response variable in this example is the gross primary production (GPP), which is the total amount of energy primarily produced by plants through photosynthesis. The GPP can be calculated from the observations at the eddy flux towers. However, for over a region such as a country or continent, the GPP has to be estimated by employing either statistical models or ecosystem models, which may range in complexity from empirical models (e.g., [?, ?]) to biogeochemical models (e.g.,[?, ?, ?]). Linear regression models have been employed to estimate the regional GPP. For example, Zhang, et al. [?] used an empirical piecewise regression model to map GPP for the Northern Great Plains grasslands from flux tower measurements. Xiao, et al. [?] developed an upscaling model based on the regression tree method to extrapolate eddy flux GPP data to the continental scale and producing continuous GPP estimates across multiple biomes. Mueller, et al. [?] studied the variability of carbon flux measurement across different temporal scales. We will examine estimations of regional GPP given by models of different time scales.

3.1 Data and Model

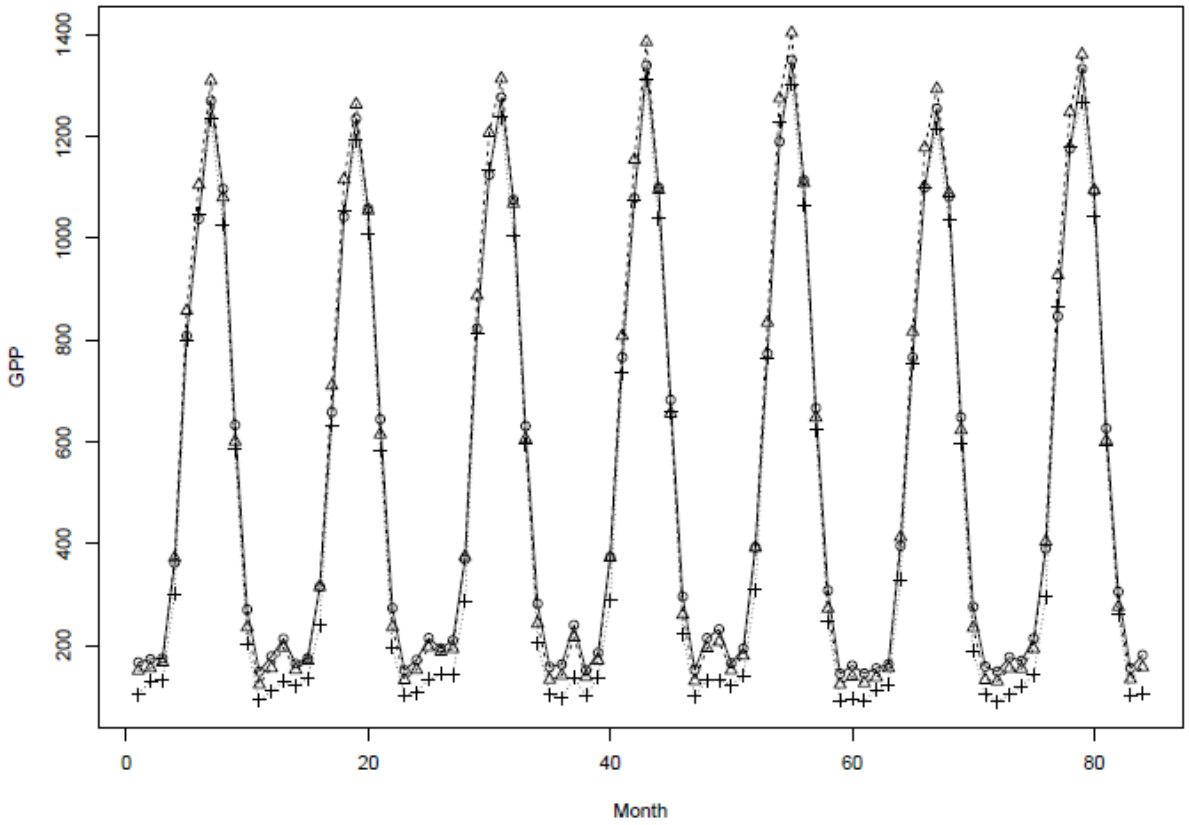
We used the data collected at the AmeriFlux towers at 70 sites (http://ameriflux.ornl.gov/). We obtained the level 4 data from http://cdiac.ornl.gov/ftp/ameriflux/data/Level4/. The data consist of observations collected every half hour ranging from 2000 to 2007 at each site. The response variable is GPP and six explanatory variables are air temperature, global radiation, precipitation, vapor pressure deficit, land-cover type and enhanced vegetation index (EVI). These six variables were chosen based on previous studies. The first five variables were observed at the AmeriFlux sites and EVI was calculated from the Moderate Resolution Imaging Spectroradiometer (MODIS) every 8 days, which is the reason we choose the 8-day scale instead of the weekly scale. The land-cover type is a qualitative variable with 6 levels representing 6 land-cover categories. Based on these data, we fitted a polynomial regression of order 2 from (7) at three different scales: daily, 8-day, and monthly. We therefore have three fitted regression models.

To predict GPP at a site that is not part of AmeriFlux net, we use data from the North American Regional Reanalysis (NARR) (http://www.emc.ncep.noaa.gov/mmb/rreanl/). This data set has a spatial resolution of 0.5×0.5 degrees over the conterminous US, and the time range is 2001-2007. In total, the whole US has 3252 pixels. We predict the GPP at each of the pixel using the three fitted models and calculated the total GPP over the US by adding the pixel-level GPP.

3.2 Results

The first conclusion we can draw is that a large scale model can result in larger or smaller prediction. This can be seen in Table 1 which summarized the total GPP over the US for each year. We see that the 8-day model yields higher total GPP than the daily model in each of the seven years while the monthly model yields lower total GPP than the daily model. Figure 1 shows three predicted monthly total GPP over the US for each month between 2001 and 2007 in the whole US, from which we can see that the predicted monthly GPP from the three different temporal scale models are different. The 8-day model consistently provides higher predicted total GPP, which is consistent to what we observed from Table 1.

Figure 1: Predicted monthly GPP (TgC) across 2001-2007 given by the daily model ($\circ$), the 8-day model (Δ), and the monthly model ($+$).



Year	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
Daily	6328	6109	6528	6587	6697	6299	6673
8-Day	6338	6133	6572	6604	6753	6348	6728
Monthly	5770	5509	5903	5941	6128	5744	6084

Table 1: The predicted annual GPP(Units: $TgCyr^{-1}$) over the US by year.

In Figure (2), we plot the predicted annual GPP for the year 2007 at each pixel. The three different models reveal about the same spatial trend, but a careful examination also reveals some differences of the predicted GPPs in some areas.

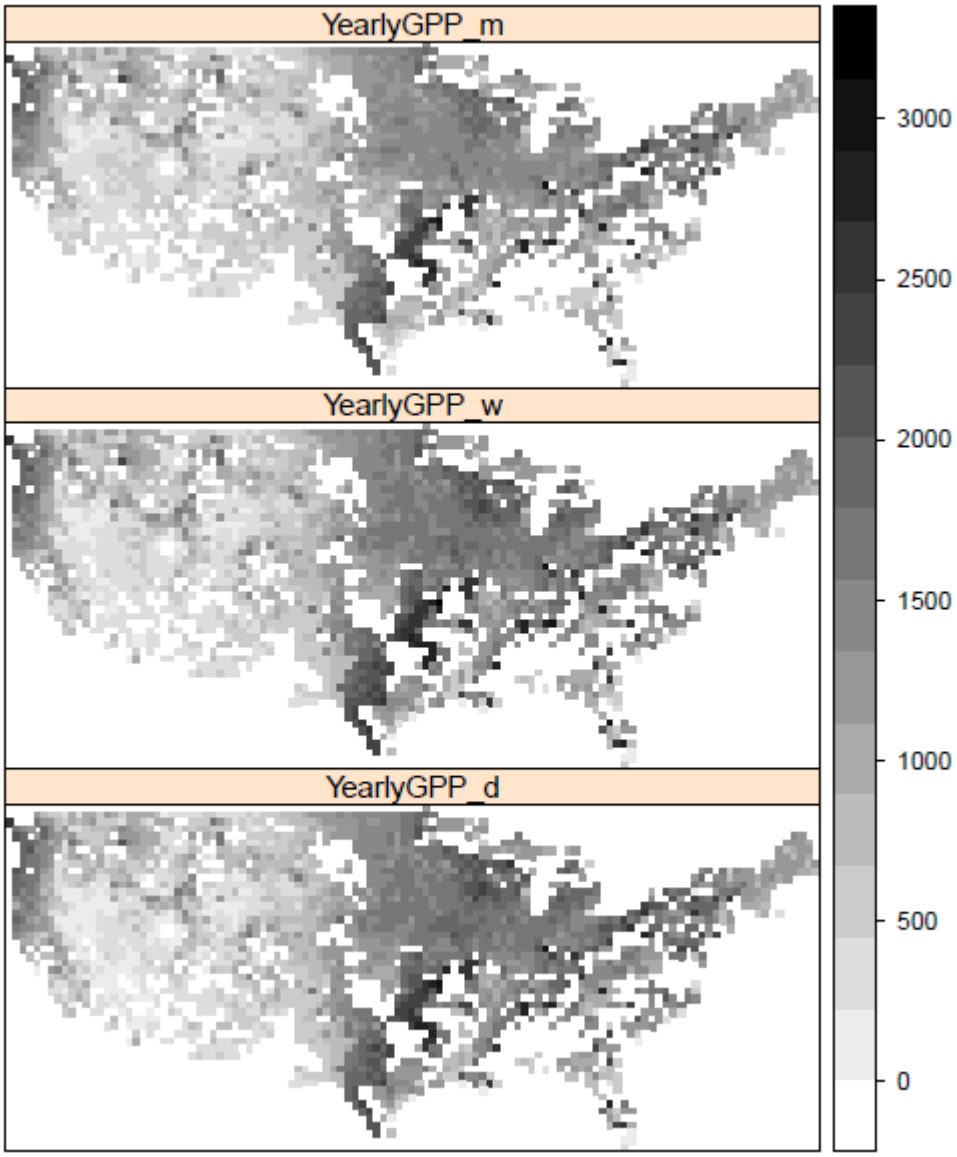


Figure 2: Annual GPP(Units: $gCm^{-2}yr^{-1}$) predicted by three models for year 2007: monthly model (top), 8-day model (middle) and daily model (bottom).

Next we compare the prediction variances given by the three models at each pixel. Figure 3 plots the standard errors given by the three different temporal scale models at each pixel for year 2007. It is evident that the prediction error is smaller for finer resolutions, although we cannot justify this theoretically in this case.

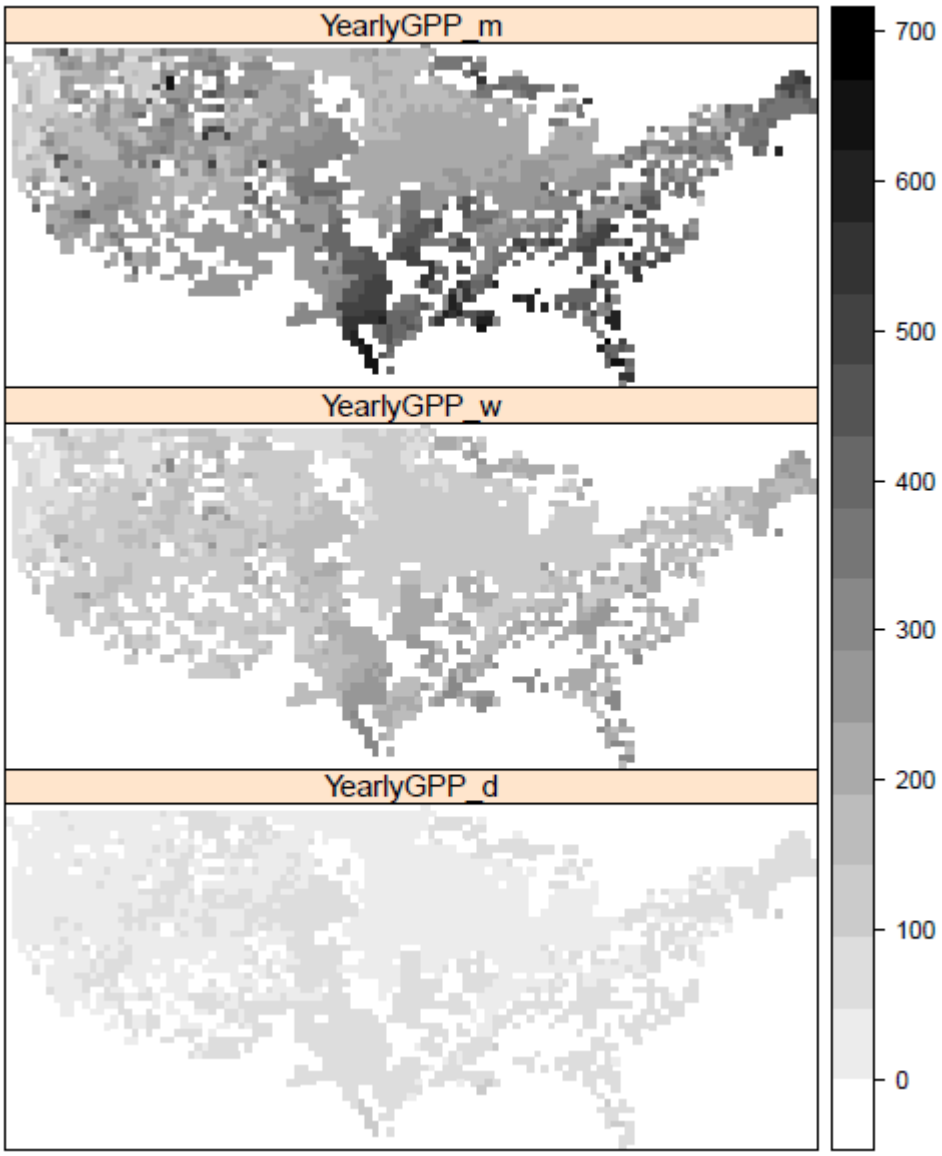


Figure 3: Standard error of of GPP(Units: $gCm^{-2}yr^{-1}$) at each pixel for year 2007: monthly model (top), 8-day model (middle) and daily model (bottom).

4. DISCUSSION

In this paper, we provided some theoretical discussions on the scale issue in linear models. When there is no high order terms in the model, the smaller scale model is preferred whenever possible. However, if the model includes high order terms of the explanatory variables, direct comparisons are difficult and no explicit conclusions are given in this paper. The example revealed that a larger scale model can yield either larger or smaller predictions. For the polynomial regression, it would be an interesting problem to provide some conditions under which the larger scale model yields larger predictions, or conditions under which the larger scale model yields smaller predictions. It would be also interesting to investigate how the scales affect the prediction variance.

ACKNOWLEDGMENT

This research is partially supported by grants from the National Science Foundation of the US (IIS-1028291, NSF-1028291 and NSF-0919331, NSF-0630319), NASA (NNX09AI26G), and by Department of Energy (DE-FG02-08ER64599).

REFERENCE

[1]. S. D. Prince and S. N. Goward, "Global primary production: A remote sensing approach," *Journal of Biogeography*, vol. 22, no. 4/5, pp. 815–835, Jul. 1995.

[2]. V. Yadav, K. L. Mueller, D. Dragoni, and A. M. Michalak, "A geostatistical synthesis study of factors affecting gross primary productivity in various ecosystems of north america," *Biogeosciences*, vol. 7, no. 9, pp. 2655–2671, Sep. 2010.

[3]. B. E. Law, D. Turner, J. Campbell, M. Lefsky, M. Guzy, O. Sun, S. V. Tuyl, and W. Cohen, "Carbon fluxes across regions: Observational constraints at multiple scales," in *SCALING AND UNCERTAINTY ANALYSIS IN ECOLOGY*. Springer Netherlands, Jan. 2006.

[4]. V. J. Berrocal, P. F. Craigmile, and P. Guttorp, "Regional climate model assessment using statistical upscaling and downscaling techniques," *Environmetrics*, vol. 23, no. 5, p. 482–492, 2012.

[5]. K. L. Mueller, V. Yadav, P. S. Curtis, C. Vogel, and A. M. Michalak, "Attributing the variability of eddy-covariance co2 flux measurements across temporal scales using geostatistical regression for a mixed northern hardwood forest," *Global Biogeochemical Cycles*, vol. 24, no. 3, 2010.

[6]. A. Patil and Z. Deng, "Temporal scale effect of loading data on instream nitrate-nitrogen load computation," *Water science and technology: a journal of the International Association on Water Pollution Research*, vol. 66, no. 1, pp. 36–44, 2012.

[7]. J. Xiao and Q. Zhuang, "A continuous measure of gross primary productivity for the conterminous u.s. derived from modis and ameriflux data," *Remote sensing of environment*, vol. 114, pp. 576–591, 2010.

[8]. J. H. Stapleton, *Linear Statistical Models*. John Wiley & Sons, 1995.

[9]. F. Yang, K. Ichii, M. A. White, H. Hashimoto, A. R. Michaelis, P. Votava, A.-X. Zhu, A. Huete, S. W. Running, and R. R. Nemani, "Developing a continental-scale measure of gross primary production by combining modis and ameriflux data through support vector machine approach," *Remote Sensing of Environment*, vol. 110, no. 1, pp. 109–122, Sep. 2007.

[10]. S. W. Running, R. R. Nemani, F. A. Heinsch, M. Zhao, M. Reeves, and H. Hashimoto, "A continuous satellite-derived measure of global terrestrial primary production," *BioScience*, vol. 54, no. 6, pp. 547– 560, Jun. 2004.

[11]. D. P. Turner, M. Guzy, M. A. Lefsky, W. D. Ritts, S. V. Tuyl, and B. E. Law, "Monitoring forest carbon sequestration with remote sensing and carbon cycle modeling," *Environmental Management*, vol. 33, no. 4, pp. 457–466, Aug. 2004.

[12]. L. Zhang, B. Wylie, T. Loveland, E. Fosnight, L. L. Tieszen, L. Ji, and T. Gilmanov, "Evaluation and comparison of gross primary production estimates for the northern great plains grasslands," *Remote Sensing of Environment*, vol. 106, no. 2, pp. 173–189, Jan. 2007.