Modeling Seasonal Dynamics of Surface Soil Bulk Density in a Forest in West Virginia

Qiuchen Li

Problem Report submitted to the
Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Master of Science
in Statistics

Philip Turk, Ph.D., Committee Chair
E. James Harner, Ph.D.
Yanqing Hu, Ph.D.

Department of Statistics

Morgantown, West Virginia
2014

Keywords: soil bulk density, seasonality, linear mixed model, spline

ABSTRACT

Modeling Seasonal Dynamics of Surface Soil Bulk Density in a Forest in West Virginia

Qiuchen Li

Bulk density is a commonly measured soil property during field investigations of soils. Accurate and reliable bulk density measurements are critical for assessing soil productivity and soil degradation. It is recognized that bulk density is variable both spatially and temporally. However, most attempts to quantify the dynamic nature of bulk density have focused on agricultural fields and the effects of tillage operations. Our objective was to determine if there are significant seasonal changes to the measured bulk density of surface soil horizons (O and A) in a forested ecosystem. The frame method was used to measure bulk density at monthly intervals for 12 months at 10 locations within a forested catchment selected using a generalized random tessellation stratified spatial sample. We report the results of fitting a linear mixed model to bulk density to examine the effects of seasonality and horizons. Moreover, a B-spline expansion on time is used to examine non-linear seasonal effects for a specific soil horizon.

Contents

Acknowledgements

I want to thank Dr.Turk for his help and effort to this paper. He is the mentor of my research and life.

1 Introduction

Bulk density of soil is determined by the mass of soil per volume of soil, which is critical for assessing soil productivity and soil degradation. Moreover, accurate and reliable bulk density measurements are the pivots for converting mass-based measurements to volume-based values. In forested ecosystems, bulk density is used to evaluate the litter layer and forest floor for erosion potential, soil-water relationships, soil pH, etc. It is known that bulk density greatly depends on the mineral makeup and soil layer, or horizon. The differentiation of the bulk density is largely the result of influences such as the existence of rock fragments and the existence of macroscopic vegetal (Nottingham et al. 2013 [1]).

Coopers Rock State Forest is a 12,713 acre state forest located in Monongalia and Preston counties in West Virginia (www.coopersrockstateforest.com [2]). Its southern edge borders the Cheat Lake and the canyon section of the Cheat River. The landscape is similar in the forest, where it consists of even aged, fire tolerant species such as scarlet oak and red maple. Nottingham et al. (2013 [1]) used the frame excavation method to measure bulk density of surface soil horizons (O_i, O_e and A, from top to bottom) at monthly intervals for 12 months at 10 random locations in Coopers Rock (see Figure 1 in Appendix). They fit a linear mixed model to ln(RFBD), where RFBD is rock-free bulk density, using the fixed factor Horizon (O_i, O_e, and A), the predictor Month (1, 2,..., 12) and Plot*Horizon random effects. Based on Akaike's Information Criterion and a likelihood ratio test, they selected a first-order autoregressive (AR(1)) error covariance structure.

They found out that there were large differences in the RFBD among the three horizons, with O horizons having the lowest values and the A horizon is the highest. According to their research, there were no obvious seasonal

effects nor trend observed in the O_e or A horizon RFBD data. However, they discovered that there was mild evidence of seasonal effects for RFBD in the O_i horizon. From the plot of ln(RFBD) in Figure 1, where the x-axis is the month, the association is non-linear. Thus, we chose a B-spline model to address the seasonality.

A linear mixed model assumes that the relationship between the mean of the dependent variable y, which is the response, and the factors, which are fixed and random effects, can be modeled as a linear function (Ruppert et al. 2013 [3]). The error variance is constant and is not dependent on the mean. The random effects follow a normal distribution with mean 0 and the variance of y depends on a set of unknown components. Conceptually the variance components estimation problem can be broken into two parts. The first part is the estimation of the variance components associated with the random effects. The second part is the estimation of the variance components associated with the error distribution.

A spline function is a piecewise polynomial function in which the same degree individual polynomials are connected "smoothly" at join points from prespecified knots (James et al. 2013 [4]). Basis functions are used to form a linear combination in a regression spline format that allow us to adequately model non-linear behavior between y and x. The basis set is a family of functions, that we can get from statistical software, or transformations that can be applied to a variable X: $b_1(X), b_2(X), ..., b_k(X)$. The points where the coefficients change are called knots. Both the first and second derivatives of the piecewise polynomial must be continuous at knots to keep the function smooth. Thus, cubic splines are popular because the discontinuity cannot be detected by human eyes. In this paper, we used a cubic B-spline basis, which is the default in SAS software.

The West Virginia University Division of Plant and Soil Sciences was

interested to determine if RFBD is influenced by changing seasons. My objective was to use the dataset from the O_i horizon collected by Dr.James Thompson and determine how seasonal effects impacted the RFBD.

2 Methods

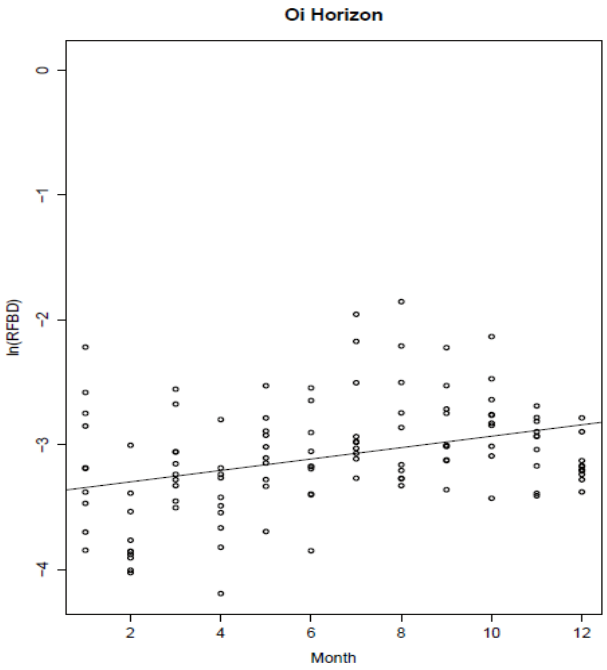
Table 1 below shows some basic statistical summary information. We took the log transformation of RFBD initially because of non-constant variance seen in the model from Nottingham et al. (2013 [1]), which is shown as the last row of Table 1, denoted tRFBD.

Table 1

Variable	N	Mean	Std Dev	Minimum	Maximum
Month	120	6.5000000	3.4665266	1.0000000	12.0000000
RFBD	120	0.0505980	0.0248130	0.0151280	0.1565867
tRFBD	120	-3.0849533	0.4455793	-4.1912080	-1.8541451

Following is a plot of ln(RFBD) versus month of data collection (Month). Using the model from Nottingham et al. (2013 [1]), as we can see in Figure 1, it is not suitable to fit the trend as linear. September was the first observed month, October was the second, so on. August of the next year is the last month. The ln(RFBD) during the fall was the lowest in the year. With seasons changing, the ln(RFBD) goes up in the winter (December to February), reaches a peak in spring (March to May), and then goes down in the summer (June to August). Notice the values of November look unusual due to a big storm that year.

Figure 1



A model with both fixed effects and random effects is called a mixed-effects model or linear mixed model (Gurka 2006 [5]). Mixed-effects models are primarily used to describe relationships between a response variable and some covariates in data. These covariates can contain classification factors. A linear mixed model is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{b}_1 + \boldsymbol{Z}_2\boldsymbol{b}_2 + ... + \boldsymbol{Z}_m\boldsymbol{b}_m + \boldsymbol{\epsilon} \qquad (1)$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\boldsymbol{I})$, $\boldsymbol{X}_{n\times p}$ is a fixed effects design matrix, $\boldsymbol{\beta}$ is a fixed effects $(p \times 1)$ vector of unknown, constant population parameters, and $\boldsymbol{Z}_i$ is an $(n \times r_i)$ random effects design matrix, where $i = 1, \ldots, m$. $\boldsymbol{b}_i \sim N(0, \sigma^2_{plot})$ is a random effects $(r_i \times 1)$ vector associated with a particular plot selected at random from the population. We also assumed the ln(RFBD) of one month would have an effect on the next month, following an AR(1) structure. We

treated the errors as AR(1), $\epsilon_t = \phi\epsilon_{t-1} + \omega_t$, which has two parameters: σ_w^2 and ϕ.

A spline model is:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + ... + \beta_K b_K(x_i) + \epsilon_i \qquad (2)$$

The basis functions $b_1(.), b_2(.), ..., b_K(.)$ are functions of X, which is fixed and known (James et al. 2013 [4]). We can treat $b_1(x_i)$ as x_1^*, $b_2(x_i)$ as x_2^* and so forth. Thus, equation (2) can be written as:

$$y_i = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \beta_3 x_3^* + ... + \beta_K x_K^* + \epsilon_i \qquad (3)$$

As mentioned, spline function is a piecewise polynomial function in which the individual polynomials, the basis functions, have the same degree d (http://support.sas.com [6]). These basis functions are connected smoothly at join prespecified points, referred to as knots. Visually, a cubic spline, a spline of degree 3, is a smooth curve, and it is the most commonly used spline.

As we can see, where we should place the knots is the key of fitting a spline model. Usually there are two ways: one way is to place more knots where we feel the function might change frequently, and to place less knots where it seems to behave in a uniform fashion; another way is to set the desired degrees of freedom in a statistical software, and then the software will automatically select the corresponding amount of knots at uniform quantities of the data (James et al. 2013 [4]). In this paper, we used SAS to do our analysis and to choose knots. Table 2 on the left hand side shows that SAS chose 9 knots and where it chose the knots. Those stars are showing that some knots are outside of the data boundaries, which would not be shown on the graph.

Table 2

Knots for Spline Effect spl		
Knot Number	Boundary	Month
1	*	-4.50000
2	*	-1.75000
3	*	1.00000
4		3.75000
5		6.50000
6		9.25000
7	*	12.00000
8	*	14.75000
9	*	17.50000

Basis Details for Spline Effect spl			
Column	Support		Support Knots
1	-4.50000	3.75000	1-4
2	-4.50000	6.50000	1-5
3	-1.75000	9.25000	2-6
4	1.00000	12.00000	3-7
5	3.75000	14.75000	4-8
6	6.50000	17.50000	5-9
7	9.25000	17.50000	6-9

A B-spline basis can be built by starting with a set of Haar basis functions, which are functions that are 1 between adjacent knots and 0 elsewhere. Then applying a simple linear recursion relationship d times, this yields the $n+d+1$ needed basis functions (http://support.sas.com [7]). Since we used cubic B-spline basis, the d here was 3. For the purpose of building the B-spline basis, the n prespecified knots are referred to as internal knots. Since negative months and the months larger than 12 are not actually in the domain, months 3.75, 6.50 and 9.25 are internal knots. Thus, the n was 3 and we obtained $3+3+1=7$ basis functions. The right hand side of Table 2 tells where the 7 basis functions, $b_i(.) >= 0$, are non-zero. For instance, basis function 1 is non-zero at months from -4.5 to 3.75. This construction requires d additional knots, known as boundary knots, to be positioned to the left of the internal knots, and $MAX(d,1)$ boundary knots to be positioned to the right of the internal knots. Therefore, we obtained $3+3+3=9$ knots.

Since we fit a B-spline expansion on Month, the predictors are functions of Month. Then we fit a linear mixed model to ln(RFBD) with the B-spline expansion of Month and AR(1) errors on the right hand side. For each plot, we treated it as a time series of length 12, so that we have 10 time series total, each of length 12.

We used the model to obtain the conditional predictions and marginal predictions of ln(RFBD). We used contrasts $\boldsymbol{c}'\boldsymbol{\mu}=\mathbf{0}$, where $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ is the parameter vector and $\boldsymbol{c}$ is the contrast coefficient vector, to check whether there were seasonal differences, for example: fall versus winter. We used the marginal predictions of September, October, and November across the 10 plots to obtain the grand mean for fall. We also obtained the mean of winter in the same way and then estimated the following seasonal contrast.

$$(\mu_1 + \mu_2 + \mu_3)/3 - (\mu_4 + \mu_5 + \mu_6)/3 = 0 \qquad (5)$$

where μ_i is the marginal mean for month i.

We use month 5 as an example to explain how to get the marginal predictions $\hat{\boldsymbol{y}}_m = \hat{\boldsymbol{\beta}}'\boldsymbol{x}^*$ of ln(RFBD) by linear algebra. Table 3 shows the basis values for month 5. Hastie et al. (2008 [8]) introduce the theory about how to calculate these basis functions, but here we just use the result given by SAS. According to Table 2 right hand side, month 5 only gets positive support from basis functions 2, 3, 4, and 5. These four basis functions are then evaluated at the 5th month.

Table 3

Coefficients for Estimate Test case - month 5 marg		
Effect	spl	Row1
Intercept		1
spl	1	
spl	2	0.027
spl	3	0.507
spl	4	0.4503
spl	5	0.0157

Coefficients for Estimate Test case - month 5 marg		
Effect	spl	Row1
spl	6	
spl	7	

We can obtain the $\hat{\beta}$'s of the linear mixed model from Table 4. The marginal prediction of 5th month is the values of the basis functions multiplied by the appropriate $\hat{\beta}$'s, which is:

$$\hat{y}_5 = -3.6190 + (-0.4178)(0.027) + (0.4185)(0.507) + (0.6195)(0.4503) + (0.9466)(0.0157) = -3.1243$$

Table 4

Solutions for Fixed Effects						
Effect	**spl**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept		-3.6190	1.8230	102.6	-1.99	0.0498
spl	1	3.9637	2.3603	104	1.68	0.0961
spl	2	-0.4178	1.9542	102.1	-0.21	0.8311
spl	3	0.4185	1.7702	103.1	0.24	0.8136
spl	4	0.6195	1.9213	102.8	0.32	0.7478
spl	5	0.9466	1.6717	102	0.57	0.5725
spl	6	0.4671	2.2167	103.5	0.21	0.8335
spl	7	0	.	.	.	.

We can get the marginal predictions of 12 months in the same way. Then, the predicted ln(RFBD) of fall is the average of $\hat{y}_1$, $\hat{y}_2$, and $\hat{y}_3$ that is:

$$\hat{y}_{fall} = (-3.16712 - 3.5136 - 3.45165)/3 = -3.377457$$

The other three season's predictions were calculated in the same way. Based on equation (5), the estimated difference between fall and winter is:

$$\hat{y}_{fall} - \hat{y}_{winter} = -3.377457 - (-3.13706) = -0.240397$$

Then, we checked whether this difference is significant using a t-test.

3 Results

Table 5 is the result of using a likelihood ratio test on the plot variance component and on the autocorrelation parameter ϕ. The first row tests whether there is meaningful plot-to-plot variation; that is, $\sigma_{plot} = 0$ or $\sigma_{plot} > 0$. The second row tests whether ϕ is 0 or not. If ϕ is 0, then we do not

need AR(1) which means $\epsilon_t = \omega_t$, where ω_t is white noise $(0, \sigma^2)$. From the P-values we can see that there is no plot-to-plot variation and it is not necessary to fit an AR(1) model for the errors. However, the AR(1) model was retained for consistency with the previous model used earlier in this study.

Table 5

Tests of Covariance Parameters Based on the Restricted Likelihood					
Label	**DF**	**-2 Res Log Like**	**ChiSq**	**Pr > ChiSq**	**Note**
Parameter list	1	110.99	0.02	0.4480	MI
Parameter list	1	111.07	0.09	0.7604	DF

The estimate of the intercept in Table 6 is the estimate of σ_p^2 and it is near 0. Since we already know that there is no significant plot-to-plot variation, we would not pay attention to it. The estimate for AR(1) is the estimate of ϕ so the auto-correlation parameter of the AR(1) model is estimated to be -0.03301. In the third row, σ^2 is estimated to be 0.1496, the estimate of $Var[\epsilon_t]$.

Table 6

Covariance Parameter Estimates			
Cov Parm	**Subject**	**Estimate**	**Standard Error**
Intercept	Plot	0.000843	0.006616
AR(1)	Plot	-0.03301	0.1078
Residual		0.1496	0.02088

We also checked other results from fitting the B-spline model. Type 3 F-tests of fixed effects were highly significant for the B-spline expansion of

Month (P-value < 0.0001). Thus, using the B-spline model to address the non-linearity of the longitudinal profiles is needed and correct.

After fitting the linear mixed model to ln(RFBD) with a B-spline expansion of Month and AR(1) errors, we obtained marginal predictions of y. The results of seasonal contrast hypothesis tests are shown in Table 7. Since we did six tests at once, we had to adjust these P-values to make sure $\alpha_f = 0.05$ was preserved. We used the Holm-Bonferroni procedure in this paper. From the P-values in Table 7, we can tell fall versus all other seasons is significant and winter versus spring is also significant. However, there is no difference between winter and summer, and spring and summer. Overall, seasonality does exist.

Table 7

Estimates Adjustment for Multiplicity: Holm						
Label	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Adj P**
Fall vs Winter	0.2404	0.09056	45.18	2.65	0.0109	0.0328
Fall vs Spring	0.5197	0.09208	37.29	5.64	<.0001	<.0001
Fall vs Summer	0.3815	0.09448	39.22	4.04	0.0002	0.0012
Winter vs Spring	0.2793	0.08643	44.51	3.23	0.0023	0.0093
Winter vs Summer	0.1411	0.09208	37.29	1.53	0.1339	0.2675
Spring vs Summer	0.1383	0.09056	45.18	1.53	0.1338	0.2675

Looking at the left top panel of Figure 2, the scatter is generally evenly dispersed vertically about zero so the variance in the residuals appears constant. In the left bottom plot, the assumption of normal errors is plausible due to residuals following the line approximately. Since every studentized residual is approximately between -3 and 3, as they follow an approximate N(0,1) distribution, there are no outliers, as shown in the right top panel. Thus the assumptions of the model are met.

Figure 2

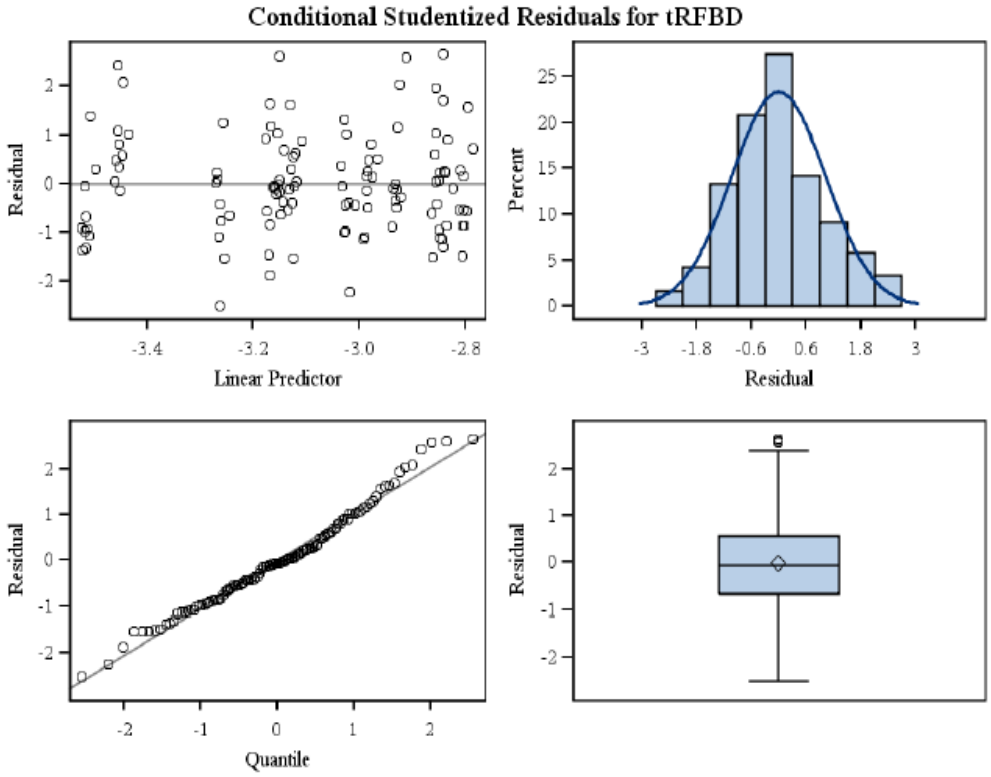
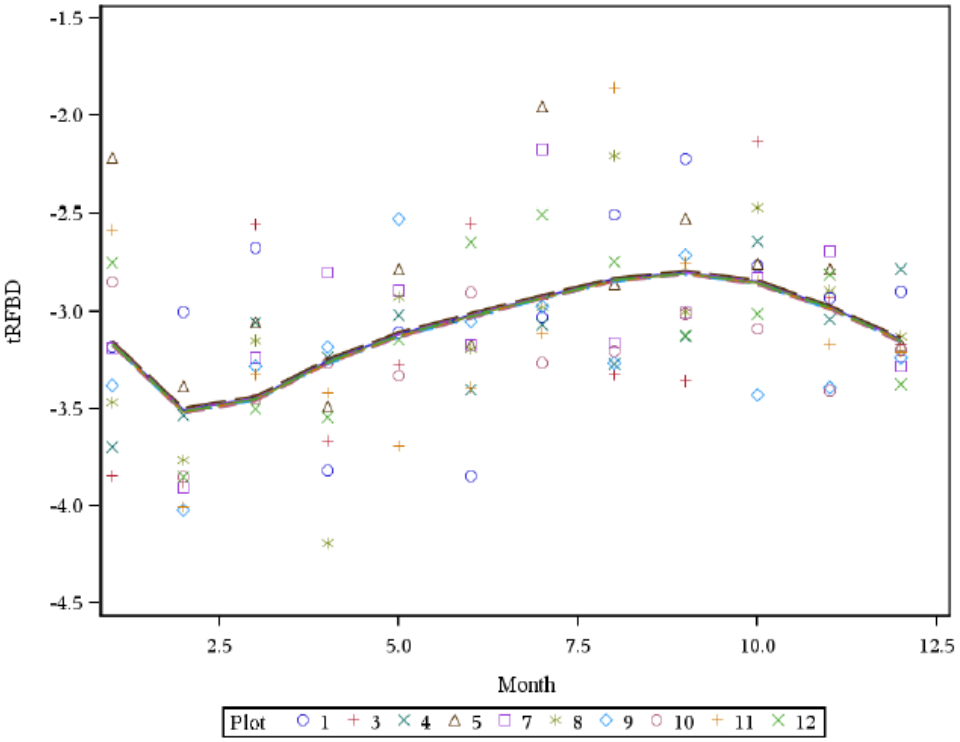


Figure 3 is the scatter plot of ln(RFBD), where the x-axis is the month, and the fitted longitudinal profiles for all the plots constructed using interpolated conditional predictions. We used different characters to represent each plot. Since there is no plot-to-plot variation, they stack up on top of one another.

Figure 3



4 Discussion

Since the seasonal effects of RFBD do exist, soil scientists should be aware that it is better to measure the O_i horizon in the same season to eliminate these effects when they are doing other effects comparisons.

Following is an explanation of the non-linear behavior we observed from fitted model. The leaves fall on the ground in the fall, yet initiation of decomposition has not started and the RFBD is at its annual lowest annual level. During the period of winter, decomposition of leaves on the ground results in the increasing amount of RFBD. In spring, the temperature starts to

increase, accelerating the decomposition and eventually we see the maximum amount of RFBD in a year. However, RFBD starts to decrease in summer.

Comparing Figure 1 with Figure 3, the spline model is a much better fit for ln(RFBD). It is easy to see the trend and seasonality of RFBD from the fitted profiles where Month in this model is a numerical predictor. Segmented regression and ANOVA approach are other alternative models we could have considered. However, splines are more flexible than segmented regression to fit curves. Moreover, we can keep Month as a numerical predictor rather than setting 12 levels in ANOVA to check seasonal effects.

We could also use the same statistical method to analyze whether O_e and A horizons are affected by seasonal effects. This would be a meaningful topic for the soil sciences.

5 Bibliography

[1] Nottingham, A.C., Thompson, J.A., Cook, J.L., Turk, P.J., & Connolly, S.J. (2013). *Seasonal dynamics of surface soil bulk density in a forested catchment.* ASA, CSSA, & SSSA International Annual Meetings, Tampa, FL.

[2] Coopers Rock State Forest. (2009). *Welcome.* Retrieved February 25, 2014 from http://www.coopersrockstateforest.com

[3] Ruppert, D., Wand, M.P., & Carroll, R.J. (2013). *Semiparametric regression.* New York City, the United States: Cambridge.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* New York City, the United States: Springer.

[5] Gurka, M.D. (2006). Selecting the best linear mixed model under REML. *The American Statistician, 60,* 19-26. doi:10.1198/000313006X90396

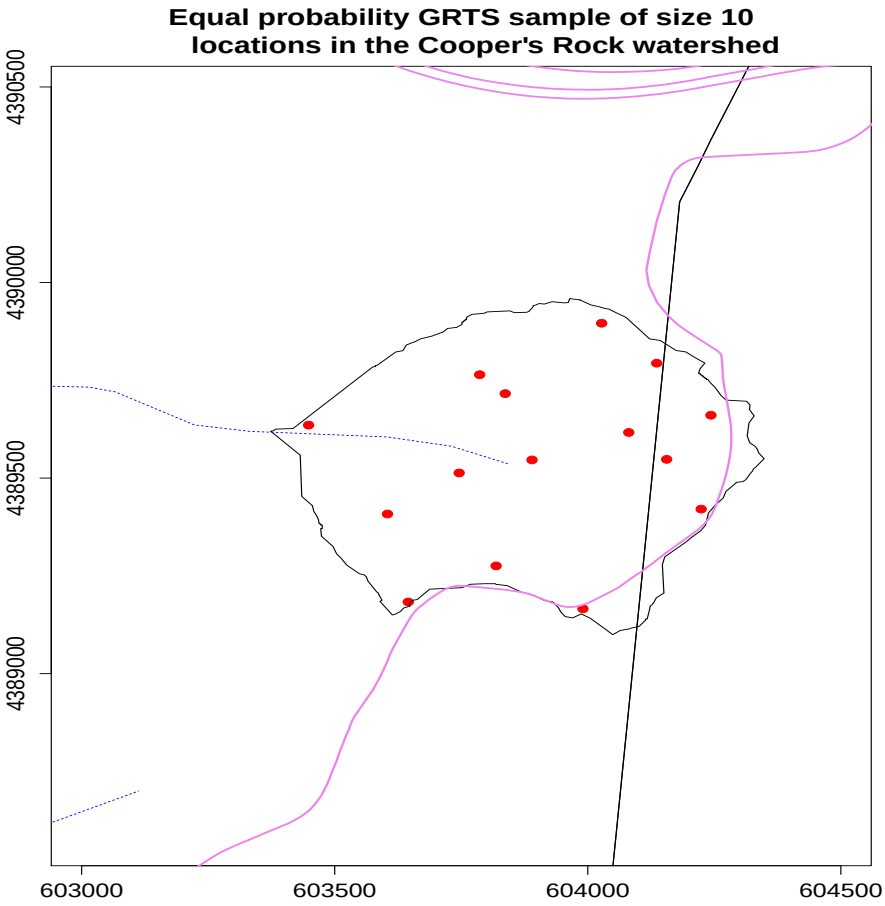
[6] SAS. (2014). *Splines and Spline Bases.* Retrieved March 27, 2014 from http://support.sas.com/documentation/cdl/en/statug/63962 /HTML/default/viewer.htm# statug_ introcom_ a0000003344.htm

[7] SAS. (2014). *B-Spline Basis.* Retrieved March 27, 2014 from http://support.sas.com/documentation/cdl/en/statug/63962 /HTML/default/viewer.htm# statug_ introcom_ a0000003346.htm

[8] Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of satistical learning.* New York City, the United States: Springer.

6 Appendix

Figure 1


Equal probability GRTS sample of size 10
locations in the Cooper's Rock watershed
4390500
4390000
4389500
4389000
603000
603500
604000
604500

SAS code

```
options nodate ls = 80 ps = 56 pageno = 1;

ods pdf;
ods graphics on;

data thompson01;
infile 'CRdata.csv' dsd firstobs = 2 lrecl = 500;
input Month Plot $ Horizon $ TotalBD RFBD CWDBD;
run;

data thompson01;
set thompson01;
if Horizon = 'Oi';
catMonth = Month;
tRFBD = log(RFBD);
run;

proc print data = thompson01;
run;

proc means data = thompson01;
run;

proc glimmix data = thompson01 plots = studentpanel;
class Plot catMonth;
effect spl = spline(Month / details);
model tRFBD = spl / solution ddfm = kr;
```

```
random int / subject = Plot;
random catMonth / subject = Plot residual type = ar(1) vcorr;
covtest 0 . .;
covtest . 0 .;
output out = glmmout
predicted(noblup) = mPred lcl(noblup) = mLPred ucl(noblup) = mUPred
predicted = cPred lcl = cLPred ucl = cUPred;

estimate 'Test case - month 5 marg' int 1 spl [1, 5] / e;

estimate 'Fall vs Winter' spl [-1, 1][-1, 2][-1, 3]
                              [1, 4][1, 5][1, 6],
         'Fall vs Spring' spl [-1, 1][-1, 2][-1, 3]
                              [1, 7][1, 8][1, 9],
         'Fall vs Summer' spl [-1, 1][-1, 2][-1, 3]
                              [1, 10][1, 11][1, 12],
         'Winter vs Spring' spl [-1, 4][-1, 5][-1, 6]
                                [1, 7][1, 8][1, 9],
         'Winter vs Summer' spl [-1, 4][-1, 5][-1, 6]
                                [1, 10][1, 11][1, 12],
         'Spring vs Summer' spl [1, 7][1, 8][1, 9]
                                [-1, 10][-1, 11][-1, 12]
         / divisor = 3 adjdfe = row adjust = bon stepdown e;

run;

proc print data = glmmout;
run;
```

```
proc sgplot data = glmmout;
scatter y = tRFBD x = Month / group = Plot;
series y = cPred x = Month / group = Plot;
yaxis min = -4.5 max = -1.5;
run;

ods graphics off;
ods pdf close;

quit;
```