STATE-SPACE ANALYSIS OF SOIL DATA: AN APPROACH BASED ON SPACE-VARYING REGRESSION MODELS

Luís Carlos Timm[1]*; Emanuel Pimentel Barbosa[2]; Manoel Dornelas de Souza[3]; José Flávio Dynia[3]; Klaus Reichardt[1]

[1]*USP/CENA - Lab. de Física do Solo, C.P. 96 - 13416-000 - Piracicaba, SP - Brasil.*
[2]*UNICAMP/IMECC - Depto. de Estatística, C.P. 6065 - 13083-970 - Campinas, SP - Brasil.*
[3]*Embrapa Meio Ambiente, C.P. 69 - 13820-000 - Jaguariúna, SP - Brasil.*
**Corresponding author <lctimm@esalq.usp.br>*

ABSTRACT: The assessment of the relationship among soil properties (such as total nitrogen and organic carbon) taken along lines called transects is a subject of great interest in agricultural experimentation. This question has been usually approached through standard state-space methods by some authors in the soil science literature. Important limitations of the mentioned procedures used in practice are pointed out and discussed in this paper, specially those related to the model parameters, meaning and practical interpretation. In the standard state-space approach, based on an autoregressive structure, it does not present any parameters that express the variables relationship at the same point in space, but only at lagged points. Also, its model parameters (in the transition matrix) have a global meaning and not a local one, not expressing more directly the soil heterogeneity. Therefore, the objective here is to propose an alternative state-space approach, based on dynamic (space-varying parameters) regression models in order to avoid the mentioned drawbacks. Soil total nitrogen and soil organic carbon samples were collected on a Typic Haplustox. Samples were taken along a line (transect) located in the middle of two adjacent contour lines. The transect samples, totaling 97, were collected in the plow layer (0-0.20 m) at points spaced 2 meters appart. Results show the comparative advantages of the proposed method (based on an alternative state-space approach) in relation to the standard state-space analysis. Such advantages are related to a more adequate incorporation of soil heterogeneity along the spatial transect resulting in a better model fitting, and greater flexibility of the model's building process with an easier interpretability of the local model coefficients.
Key words: dynamic regression, soil properties, spatial heterogeneity, Kalman filter

ANÁLISE DE DADOS DE SOLO VIA MÉTODOS DE ESPAÇO DE ESTADO: REGRESSÃO COM COEFICIENTES VARIÁVEIS

RESUMO: A avaliação da relação entre certas variáveis representando propriedades do solo (tais como nitrogênio total e carbono orgânico) coletadas ao longo de linhas chamadas "transects", é assunto de grande interesse em experimentação agrícola. Este problema tem sido usualmente abordado através de modelos estatísticos padrão de espaço de estado por alguns autores na literatura de ciência do solo. As mais importantes limitações dos procedimentos utilizados na prática são apontados e discutidos neste artigo, sendo relacionadas ao significado dos parâmetros do modelo e a sua interpretação prática. A abordagem padrão de espaço de estado, que é baseada em uma estrutura autoregressiva, não apresenta nenhum parâmetro que expressa a relação entre as variáveis no mesmo ponto do espaço, mas somente em pontos defasados. Além disso, os parâmetros do modelo (na matriz de transição) tem um significado global e não local, não expressando diretamente a heterogeneidade do solo. Desta forma, o objetivo aqui é propor uma abordagem alternativa de espaço de estado, baseada em modelos de regressão com coeficientes variando ao longo do espaço de modo a evitar estas limitações. Dados de nitrogênio total e carbono orgânico do solo foram coletados de um Latossolo. Eles foram medidos na camada de 0 – 0,20 m ao longo de uma transeção de 194 m, totalizando 97 amostras espaçadas entre si de 2 m, entre duas curvas de contorno adjacentes. Os resultados mostram as vantagens comparativas do método proposto em relação ao método de espaço de estados padrão. Tais vantagens estão relacionadas a uma mais adequada incorporação da heterogeneidade do solo ao longo da transeção espacial resultando em um melhor ajuste do modelo e a uma maior flexibilidade no processo de construção do modelo permitindo uma fácil interpretabilidade dos coeficientes estimados.
Palavras-chave: regressão dinâmica, propriedades do solo, heterogeneidade espacial, filtro de Kalman

INTRODUCTION

The analysis of soil quality and crop yield data by standard state-space methods has become a relatively common practice in the area of soil science (Wendroth et al., 1992; Dourado-Neto et al., 1999; Nielsen et al., 1999; Wendroth et al., 2001). The assessment of the relationship among certain soil quality variables such as total nitrogen, organic carbon, and others, is a subject of great interest in agricultural experimentation and soil research (Feller, 1993; Beare et al., 1994; Sikora & Stott, 1996; Wendroth et al., 1997, Timm et al., 2000). Usually these soil variables are measured along lines called transects, forming spatial data series, measured with errors, so that the use of Kalman filter based methods (state-space), in principle, sounds appropriate mainly for filtering the measurement noise and the model uncertainty of state equation. By standard state-space (Shumway, 1988) we understand that the state-variables are just a filtered (noiseless) version of the observed variables and that these noise-free quantities follow a (vector) first order autoregressive model with constant parameters. This is the special way by which the state-space approach has been employed in the soil science literature (Wendroth et al., 1992; 1997; 2001; Hui et al., 1998; Nielsen et al., 1999). Although largely used in practice for soil research and related areas as mentioned above, we argue that this basic approach presents limitations or drawbacks, the main ones being the following:

1st) Since the parameters of the auto-regressive process are constant, they do not express directly the spatial heterogeneity (soil variability) present in the transect data. The main model parameters are global or space-independent (just average values) and not local or space-dependent as it should be if we want to express the variables, relationship at each point in space, that is, to express the soil heterogeneity in an explicit way;

2nd) The model is restricted to an auto-regressive process, introducing unnecessary constraints into the relationship among the variables, limiting therefore the model´s building process. For instance, it does not permit to express the variables, relationship at the same point in space, but only at lagged points, which restricts the model´s building process;

3rd) Other limitations are related to the Kalman Filter implementation via the EM (Expectation-Maximization) algorithm (Dempster et al., 1977; Shumway, 1988; Shumway & Stoffer, 2000) which do not provide standard errors of parameter estimates in its basic formulation, unless some extra computations are considered, as for instance, using the SEM algorithm, that is, the EM algorithm with estimation of standard errors (Meng & Rubin, 1991; Shumway & Stoffer, 2000).

In fact, the third mentioned limitation is more related to the filter computational implementation than to the method itself, and therefore, can be avoided in part if we choose carefully the software to be used or if we make some extra calculations in order to include the Hessian matrix, which is related to the standard errors. For instance, the ASTSA software in its first version (Shumway, 1988) does not present standard errors of auto-regressive parameter estimates, but these estimates could be implemented with some extra computations (see, for example, Meng & Rubin, 1991). However, the first and the second mentioned limitations are more important, since they are related to modeling flexibility and parameter interpretation.

In order to overcome the mentioned points about the standard procedures in use, an alternative state-space approach is proposed, based on dynamic (space-varying coefficients) regression models, known as dynamic linear models- DLM´s (West & Harrison, 1997). The proposed models should incorporate the spatial variability (soil heterogeneity) present in the data set through the regression parameter evolution along the transect, according to a Markovian process (random walk).

The consideration of a dynamic model of a regression type as an alternative to the standard state-space approach has the merit of overcoming the two main difficulties just mentioned. The choice of predictors (regressors), in our approach, is more flexible (not restricted to the auto-regressive form) and the regression coefficients are space-dependent. Although theoretically simple, in practice this modeling process requires some data transformations, as it will be illustrated in the next sections.

Therefore, the objective of this study is to propose another methodology using a dynamic regression model with coefficients changing along the space as an alternative to the standard state-space model.

MATERIAL AND METHODS

Soil samples used were collected in Jaguariuna (22° 41' S and 47° W), SP, Brazil, on a Typic Haplustox in May 1999. Samples were taken along a line (transect) located in the middle of two adjacent contour lines. The transect samples, totaling 97, were collected in the plow layer (0-0.20 m) at points spaced 2 meters appart. The transect soil had been limed, received phosphate (broadcasted and incorporated) and was planted to an oat crop, three months before soil sampling. Samples were air dried, granted to pass a 2 mm sieve and analyzed for organic carbon by the Walkey-Black method (Walkey & Black, 1934) and for total nitrogen by the Kjeldhal method (Bremner, 1960).

The proposed method to build and implement a model for soil quality data analysis (along a spatial transect) is based on the following main assumptions: Along the transect with inherent soil variation, local characteristics are better represented by a local model (and

not a global one) with space-varying coefficients expressing the heterogeneity of the site; Different variables measured by different scales or units, in order to be analyzed or modeled in conjunction and in an efficient way, could be conveniently transformed into a dimension-free scale, as suggested by Hui et al. (1998). However, a method that does not involve data transformation is preferable since it allows an easier interpretation; Some soil quality variables (as for instance, soil total nitrogen) are time-consuming and expensive to be measured, but can be well correlated to other variables easier to be measured (as for example, the soil organic carbon).

The proposed method for soil quality data analysis through space-varying regression models is based on the following procedure:

1st stage : Data transformation

The data should be transformed from x to x' through x' = [x – (m – 2s)] / 4s, where m and s are the mean and standard deviation of the original x data as suggested in the mentioned soil science literature (Hui et al., 1998; Nielsen et al., 1999; Wendroth et al., 2001).

2nd stage : Model Building and Fitting

Step 2.1 – Define which variable is the basic regressor (in our case, the "easy to measure" variable is soil organic carbon in its filtered version) and which one is the response or dependent variable (in our case, the "expensive to measure" variable is soil total nitrogen in its original version). The lagged version of the dependent variable can eventually be used as a second regressor;

Step 2.2 – Once the model variables have been previously defined and prepared (transformation, etc.), then fit the space-varying coefficient regression model as a special dynamic model with regression components, using any implementation of the DLM, as for instance, the BATS system (**B**ayesian **A**nalysis of **T**ime **S**eries; West & Harrison, 1997), or the PRVWIN System (**PR**ediction with **V**arying coefficients models for **WIN**dows; PRVWIN User´s Guide, 2000) or STAMP (**S**tructural **T**ime Series **A**nalyzer, **M**odeller and **P**redictor; Koopman et al., 2000).

Model Formulation

The *standard state-space model* for the soil data (nitrogen N_i and carbon C_i, i = 1, 2, ...n, where the subscript i indicates the position along the transect) is formulated by an observation equation which relates the observed data $y_i = (N_i, C_i)'$ to a non-observable state-vector $x_i = (N_i^*, C_i^*)'$, and a system equation that describes the state-vector evolution in space, as follows,

observation (*measurement*) *equation*:

$$y_i = x_i + v_i , \qquad v_i \sim N(0; V)$$

system (*space evolution*) *equation*:

$$x_i = G\, x_{i-1} + w_i , \quad w_i \sim N(0; W) ,$$

where G is the system evolution matrix (its elements are the auto-regressive coefficients), v_i is the observation (measurement) error with variance matrix V, and w_i is the system perturbation with variance matrix W. The key element is the state-vector x_i formed by the filtered version (noise free) of the observables, which is obtained sequentially by the Kalman filter updating equations. For further details, see for example, Shumway & Stoffer (2000), where the full implementation aspects are presented.

The alternative modeling approach to the soil data is the *space-varying coefficient regression (state-space) model* in which the state-vector is formed by the dynamic regression coefficients β_i, as follows,

observation equation:

$$y_i = F_i\, \beta_i + v_i , \qquad v_i \sim N(0; V)$$

system evolution equation:

$$\beta_i = \beta_{i-1} + w_i , \qquad w_i \sim N(0; W) ,$$

where $y_i = N_i$ is the response variable, $F_i = (1, N_{i-1}, C_i)$ are the considered regressors, v_i and w_i are defined as before (with the exception that v_i is now a scalar), and the regression coefficient state-vector follows a random-walk type of evolution. This non-standard state-space model, known as dynamic regression model (which is a special case of the so called dynamic linear model) can be implemented as a DLM, for instance, using the BATS system (Pole et al., 1994 ; West & Harrison, 1997) or the PRVWIN System (PRVWIN User´s Guide, 2000) or STAMP (Koopman et al., 2000; Durbin & Koopman, 2001) or SsfPack (State-space form Package; Koopman, et al., 1999; Durbin & Koopman, 2001).

RESULTS AND DISCUSSION

The available data consist, therefore, of two spatial series with 97 observations each : the nitrogen series and the carbon series. These two series are plotted against the transect points (Figure 1). The series do not present any exceptional or extreme points such as outliers or likewise (Figures 1A and B). At both series along the transect points there is a detectable soil heterogeneity previously assumed (assumption i), since the process level and the process variability are not totally stable through the space. This is true since both series present a slow growth in level from transect point 1 to 80 and a decreasing one from point 80 to the end of the series, which can be seen as a process with at least two different regimes. Also, the series variability, mainly for nitrogen, presents some visible intervals with very low dispersion (transect points 30 until 40, and points 80 until 97), contrasting with a higher variability in the other parts of the transect. Therefore, our assumption **i** that the process presents local characteristics changing in space seems to be reasonable.

The data information is initially explored through an analysis of its correlation structure. The correlation structure between the two series (Cross-Correlation Function – CCF) is shown at Figure 2C, where the high value (near 0.80) for CCF at lag 0 suggests that the carbon series can be a good predictor for the nitrogen series. This is in accordance with our assumption **iii** about using a variable easy to measure such as the soil organic carbon to predict a more expensive one such as the soil total nitrogen. Also, the significant values for the Auto-Correlation Function – ACF at the first three lags for both series (Figures 2A and 2B) show the presence of spatial correlation and an autoregressive structure for observations distant until 6 m in this study. This data can be analyzed through a state-space model, since it is based on a vector AR structure (Figures 2A, 2B and 2C).

Model Implementation and Numerical Results

The two classes of models just presented are here implemented in different versions. The standard state-space model presented (called from now on *Model I*) is implemented in three different versions: (a) without data transformation; (b) with standard transformation (variables centered on the mean and divided by its standard deviation); and (c) with the data transformed from x to x'.

The dynamic (space-varying) regression model for the nitrogen series presented in the last sub-session is implemented in two different versions: in the first one (called *Model II*) the regressors are the same as in model I (lagged nitrogen and lagged carbon), and in the second one (called *Model III*) the regressors are carbon and lagged nitrogen. Since soil carbon is more correlated to soil nitrogen than lagged carbon (as shown in Figure 2C), we expect that model III will fit better to the data than model II.

In practice, all the models have the lagged nitrogen as the second regressor and the carbon (or lagged carbon) as the first regressor, what is in accordance with any preliminary data exploration (Figure 2). The implementation of model I, apart from the data transformations in versions (b) and (c), are practically automatic, since most algorithms for the Kalman filter sequential updating equations consider a non-informative initialization (it is not necessary to specify a mean vector and a variance-covariance matrix for the state-vector distribution at the initial point) as a possible option, usually as default. The basic difference among the several implementations (softwares) available for both state-space approaches are related to the treatment given to the hyper-parameter matrices (G,V,W), for example, if estimated by the E-M algorithm coupled with the KF as in the ASTSA software

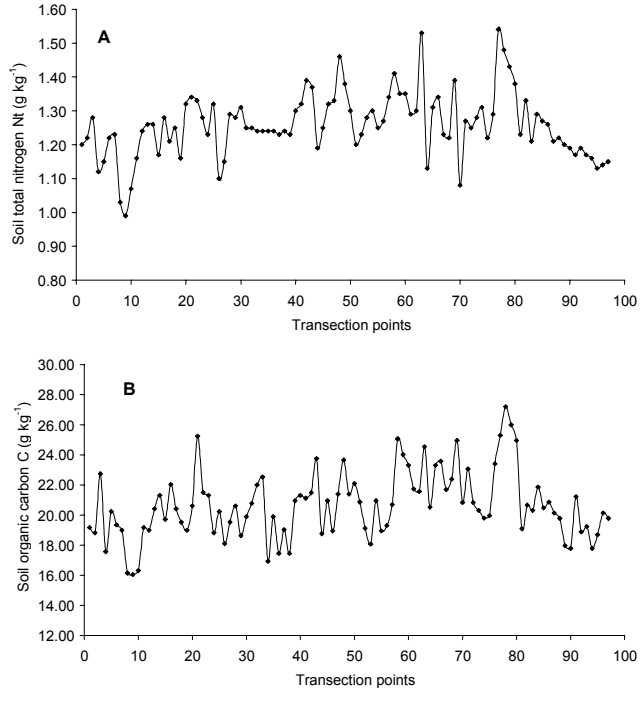

Figure 1 - Spatial data series: soil total nitrogen Nt (1A) and soil organic carbon C (1B) along the 97 point transect.

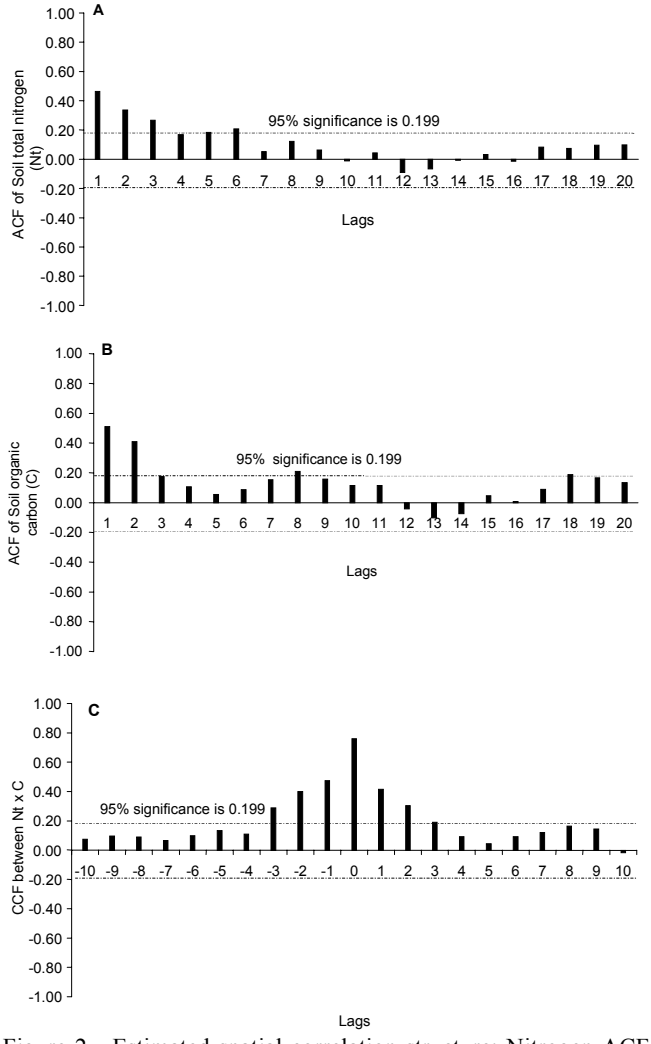

Figure 2 - Estimated spatial correlation structure: Nitrogen ACF (2A), Carbon ACF (2B) and CCF of Nt versus C (3C).

or by Bayesian methods (using Markov Chain Monte Carlo – MCMC algorithms) as in the PRVWIN software, or other method. For more details about the combination of the Kalman Filter KF and EM algorithm see, for instance, Shumway & Stoffer (2000), and for the Bayesian approach see, for instance, West & Harrison (1997). Since we have mentioned a few model implementation possibilities (softwares), with different characteristics, related to availability (free or commercial), estimation methods (Bayesian or Classic) and type of models (Standard State Space – SSS or Dynamic Linear Models – DLM´s), summary is presented of this information in Table 1. In fact, the final results are practically invariant of the particular implementation (at least the main softwares we have mentioned).

The two approaches or classes of models (total of three models in five different versions) have been fitted to the nitrogen-carbon data, where some parameter estimation (point and interval) and goodness of fit results are obtained. In fact, the improvement in goodness of fit with the R^2 measure ranging from 0.752 (Model IA), 0.806 (Model IB), and 0.991 (Model IC) shows that the performance of the standard state-space model depends, in this study, on the particular data transformation. For the Model II, R^2 measure is 0.943 and for the proposed Model III is 0.997, i.e., the proposed method (Model III) has better fitting performance when compared to Model II and Model I. However, there are some other advantages of qualitative nature for the proposed method since: (i) it does not require data transformation; (ii) the predictor variables are chosen more freely; and (iii) the model parameters have a more direct interpretation since it is a local regression model. With respect to (i), in general, for simplicity reasons, we should prefer procedures that do not depend on data transformations. Regarding (ii), the fact that the proposed Model III presents a parameter relating soil carbon and soil nitrogen at each point (the key quantity of interest) which the standard Model I does not present, is an important differential in favor of the alternative procedure. Also, with respect to (iii), since the data present local characteristics (level and variability) that change along the transect points, it is not surprising that the soil nitrogen-carbon model parameter estimates change in space (Figure 3). In this figure, the interval parameter estimates are also presented for each point, giving a more complete information on the soil carbon-nitrogen relationship.

Table 1 - Software Basic Characteristics.

Software	Availability	Estimation	Type of Models
BATS	Free	Bayesian	DLM
PRVWIN	Commercial	Bayesian	DLM and others
ASTSA	Free	Classic	Standard State Space
SAS/ETS*	Commercial	Classic	Standard State Space
STAMP	Commercial	Classic	DLM / Structural
SsfPack	Both	Both	DLM / SSS and others

*Proc State Space

The model (III) on line fitting for the nitrogen series with its two standard deviation confidence intervals are presented in Figure 4, together with the original data, showing clearly the good fitting performance, with all the data points inside the confidence intervals.

The fact that both state-space approaches provide very good fitting performance (R^2 coefficient greater than 0.99) is not surprising since they have a local characteristic, having the state variable adapted to the data at each point via Kalman Filter. Therefore, both approaches show a similar behavior with respect to fitting and predicting characteristics. On the other hand, they are very different regarding to the extraction of qualitative information from the data, where the two main differences are related to parameter definition and interpretation. As explained in this paper, the key parameters for the standard state-

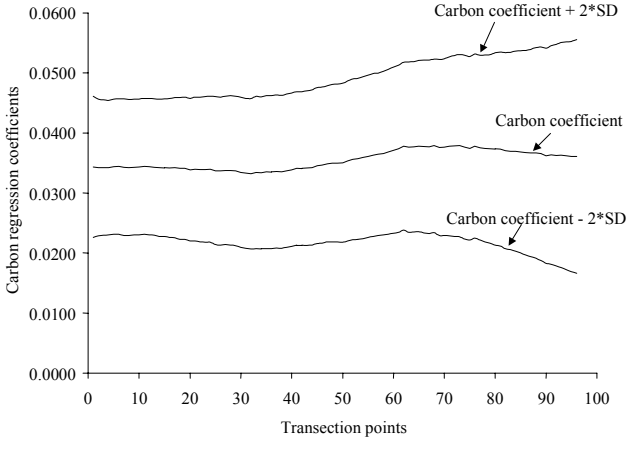

Figure 3 - Space-varying parameter estimates (carbon coefficient).

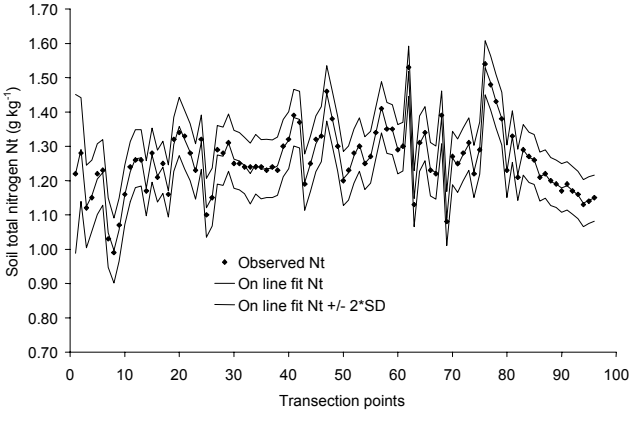

Figure 4 - Observed Nitrogen series and its on-line model fitting with 2*SD confidence interval from Model III.

space are the autoregressive coefficients (with the mentioned shortcomings) and the key quantities for the alternative state-space approach are the space-varying soil carbon-nitrogen relationship expressing the soil spatial heterogeneity along the transect.

REFERENCES

BEARE, M.H.; CABRERA, M.L.; HENDRIX, P.F. Aggregate-protected and unprotected organic matter pools in conventional and no-tillage soils. **Soil Science Society of America Journal**, v.58, p.787-795, 1994.

BREMNER, J.M. Determination of nitrogen in soil by the Kjeldahl method. **Journal of Agricultural Science**, v.55, p.11-33, 1960.

DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society Series B**, v.39, p.1-38, 1977.

DOURADO-NETO, D.; TIMM, L.C.; OLIVEIRA, J.C.M.; REICHARDT, K.; BACCHI, O.O.S.; TOMINAGA, T.T.; CASSARO, F.A.M. State-space approach for the analysis of soil water content and temperature in a sugarcane crop. **Scientia Agricola**, v.56, p.1215-1221, 1999.

DURBIN, J.; KOOPMAN, S.J. Time series analysis by state space methods. Oxford: Oxford University Press, 2001. 253p.

FELLER, C. Organic inputs, soil organic matter and functional soil organic compartments in low-activity clay soils in tropical zones. In: MULONGOY, K.; MERCKX, R. (Ed.) **Soil organic matter dynamics and sustainability of tropical agriculture**. Chichester: John Willey & Sons; Baffins Lane, 1993. p.77-88.

HUI, S.; WENDROTH, O.; PARLANGE, M.; NIELSEN, D.R. Soil variability – Infiltration relationships of agroecosystems. **Journal of Balkan Ecology**, v.1, p.21- 40, 1998.

KOOPMAN, S.J.; HARVEY, A.C.; DOORNIK, J.A.; SHEPHARD, N. STAMP. Structural time series analyser, modeller and predictor. London: Timberlake Consultants Press, 2000.

KOOPMAN, S.J.; SHEPHARD, N.; DOORNIK, J.A. Statistical algorithms for models in state-space using SsfPack 2.2 (with discussion). **Econometrics Journal**, v.2, p.107-160, 1999.

MENG, X.L.; RUBIN, D.B. Using EM to obtain assymptotic variance-covariance matrices: The SEM algorithm. **Journal of the American Statistical Association**, v.86, p.899-909, 1991.

NIELSEN, D.R.; WENDROTH, O.; PIERCE, F.J. Emerging concepts for solving the enigma of precision farm research. In: INTERNATIONAL CONFERENCE ON PRECISION AGRICULTURE, 4., Madison, 1999. **Proceedings.** Madison: ASA; SSSA; CSSA, 1999. p.303-318.

POLE, A.; WEST, M.; HARRISON, P.J. **Applied bayesian forecasting and time series analysis**. New York: Chapman & Hall, 1994. 409p.

PRVWIN User´s Guide. Rio de Janeiro: Institute of Applied Economic Research - IPEA, 2000. http://www.ipea.gov.br. (10/08/2001).

SAS/ETS **User´s Guide**. Version 8. Cary: Statistical Analysus System Institute, 2000.

SHUMWAY, R.H. **Applied statistical time series analysis**. Englewood Cliffs: Prentice-Hall, 1988. 379p.

SHUMWAY, R.H.; STOFFER, D.S. **Time series analysis and its applications**. New York: Springer-Verlag, 2000.

SIKORA, L.J.; STOTT, D.E. Soil organic carbon and nitrogen. In: DORAN, J.W.; JONES, A.J. (Ed.) **Methods for assessing soil quality**. Madison: SSSA, 1996. p.157-167. (Special Publication, 49).

TIMM, L.C.; FANTE JR., L.; BARBOSA, E.P.; REICHARDT, K.; BACCHI, O.O.S. A study of the interaction soil – plant using state-space approach. ***Scientia Agricola***, v.57, p.751-760, 2000.

WALKEY, A.; BLACK, I.A. An examination of the Degtjareff method for determining soil organic matter and a proposed modifications of the chromic acid titration method. **Soil Science**, v.37, p.29-38, 1934.

WENDROTH, O.; AL OMRAN, A.M.; KIRDA, K.; REICHARDT, K.; NIELSEN, D.R. State-space approach to spatial variability of crop yield. **Soil Science Society of America Journal**, v.56, p.801-807, 1992.

WENDROTH, O.; REYNOLDS, W.D.; VIEIRA, S.R.; REICHARDT, K.; WIRTH, S. Statistical approaches to the analysis of soil quality data. In: GREGORICH, E.G.; CARTER, M. R. (Ed.) **Soil quality for crop production and ecosystem health**. Amsterdam: Elsevier, 1997. 448p.

WENDROTH, O.; JÜRSCHIK, P.; KERSEBAUM, K.C.; REUTER, H.; VAN KESSEL, C.; NIELSEN, D.R. Identifying, understanding, and describing spatial processes in agricultural landscapes – four case studies. **Soil & Tillage Research**, v.58, p.113-127, 2001.

WEST, M.; HARRISON, J. **Bayesian forecasting and dynamic models**. 2.ed. London: Springer-Verlag, 1997. 681p.

Received May 20, 2002