Water **2015**, *7*, 1-11; doi:10.3390/w7010001

OPEN ACCESS

water

ISSN 2073-4441
www.mdpi.com/journal/water

Article

Prediction of Biomass Production and Nutrient Uptake in Land Application Using Partial Least Squares Regression Analysis

Vasileios A. Tzanakakis [1,*], Andy Mauromoustakos [2] and Andreas N. Angelakis [3]

[1] Region of Crete, Directorate of Agricultural Economy, Iraklion 71201, Greece
[2] Agricultural Statistics Lab, University of Arkansas, Fayetteville, AR 72701, USA; E-Mail: amauro@uark.edu
[3] National Agricultural Research Foundation (N.AG.RE.F.), Institute of Iraklion, Iraklion 71307, Hellas; E-Mail: info@a-angelakis.gr

* Author to whom correspondence should be addressed; E-Mail: vetzanakakis@gmail.com; Tel.: +30-2815-002-710.

Academic Editor: Miklas Scholz

Received: 10 September 2014 / Accepted: 4 December 2014 / Published: 23 December 2014

Abstract: Partial Least Squares Regression (PLSR) can integrate a great number of variables and overcome collinearity problems, a fact that makes it suitable for intensive agronomical practices such as land application. In the present study a PLSR model was developed to predict important management goals, including biomass production and nutrient recovery (*i.e.*, nitrogen and phosphorus), associated with treatment potential, environmental impacts, and economic benefits. Effluent loading and a considerable number of soil parameters commonly monitored in effluent irrigated lands were considered as potential predictor variables during the model development. All data were derived from a three year field trial including plantations of four different plant species (*Acacia cyanophylla*, *Eucalyptus camaldulensis*, *Populus nigra*, and *Arundo donax*), irrigated with pre-treated domestic effluent. PLSR method was very effective despite the small sample size and the wide nature of data set (with many highly correlated inputs and several highly correlated responses). Through PLSR method the number of initial predictor variables was reduced and only several variables were remained and included in the final PLSR model. The important input variables maintained were: Effluent loading, electrical conductivity (EC), available phosphorus (Olsen-P), Na^{+}, Ca^{2+}, Mg^{2+}, K^{2+}, SAR, and NO_3^{-}-N. Among these variables, effluent loading, EC, and nitrates had the greater contribution to the final PLSR model. PLSR is highly compatible with intensive agronomical practices such as land application, in which

a large number of highly collinear and noisy input variables is monitored to assess plant species performance and to detect impacts on the environment.

Keywords: land application; land treatment systems; biomass production; nutrient uptake; partial least squares regression (PLSR); JMP

1. Introduction

Plant biomass production and nutrient recovery are important management goals in land application associated with treatment efficiency, potential impacts on the environment, and economic benefits [1,2]. Plant species with high biomass potential usually achieve increased nutrient recovery resulting from high biomass yield and nutrient assimilation in plant tissues [3]. On the other hand, such plant species may receive high effluent loading, due to their high water requirements, which in turn may result in changes in soil properties and put the surrounding environment at risk as a result of the increased nutrient and/or pollutant release [4–6]. The latter as well as potential negative impacts on soil properties are undesirable during land application and may have negative influence on vegetation and overall system performance. Considering this close relationship between vegetation performance and soil properties in land application any potential quantitative description of this relationship in the form of a strong prediction model would be valuable and could provide useful information during system design and monitoring. Until now several statistical methods have been used to develop efficient prediction models in crop and soil science, such as principal component regression (PCR), multiple regression, and partial least squares analysis [7–10].

Partial least squares regression (PLSR) has become a popular statistical technique used widely in Chemometrics and other related areas [11–13]. There is also wide application in soil and crop studies providing information either for soil or plant parameters by considering spectroscopic measurements [14–17]. PLSR regression analysis construct models by linking predictors (Xs) and responses (Ys) and this is achieved via a projection procedure which reduce data dimensionality to a small number of important factors (also called *latent variables*). Characteristic of the method, unlike other projection methods (e.g., PCR) is that it integrates the compression and regression steps while selection of the orthogonal factors is carried out to achieve maximum covariance between the predictor and response variables [11]. Because of its principles, PLSR approach fits best in cases when matrix of predictors has more variables than observations and predictor variables are highly collinear [18].

PLSR models can deal effectively the following scenarios: (a) wide data (where number of input variables is much greater than the number of observations); (b) tall data; (c) square data; (d) collinear data; and (e) noisy data. This makes PLSR suitable for studies dealing with land application where usually a great number of highly collinear and noisy input variables is monitored to describe plant species performance and effects on soil properties and environment. However, PLSR and even multivariate techniques in general are either lacking or under-utilized in land application schemes. Thus, the primary objective in the present study was the application of a PLSR approach in order to develop a robust model capable of predicting important management goals in land application (*i.e.*, biomass production and N and P uptake) using as predictor variables critical soil parameters. The information provided here is

expected to help in the development of the appropriate methodology during design and monitoring of intensive agronomical practices, such as land application in quest of appropriate management strategies with respect to vegetation and field practices.

2. Materials and Methods

2.1. LTS Set Up, Sampling, and Chemical Analyses

A three-year-field trial with four different plant species (*Eucalyptus camandulensis*, *Acacia cyanophylla*, *Populus nigra*, and *Arundo donax*), each forming a separated land treatment system (LTS), was carried out at Skalani village, located approximately 5 km south of Iraklion city, Hellas (at 35°16'50.87" N, 25°10'52.61" E). Plant species received septic tank municipal effluents for three consecutive years (2001–2003) at hydraulic loading rate based on crop water requirements and evaporation losses. The soil in which LTS were established was characterized as a clay loam with relatively high calcium content (55% $CaCO_3$). Details about LTS set up, effluent loading and characteristics, soil properties, climatic conditions of the area, and methods used to determine and assess soil data were previously described (soil surface data from 0–7.5 to 55–65 cm obtained from the third irrigation period were included in PLSR as described below) [4]. In brief soil samples prepared and analyzed according to methods referred to the Methods of Soil Analysis [19]. pH, EC, soluble Na^+, Ca^{2+} and Mg^{2+} were assessed in saturation paste extracts with atomic absorption spectrometry (Ca^{2+} and Mg^{2+}) and flame photometer (Na^+). Soil organic matter (SOM) was assessed by the Walkley and Black wet-digestion method and available-P according to the Olsen method after extraction with $NaHCO_3$. Total Kjeldahl Nitrogen (TKN) was assessed by a macro-Kjeldahl device and analysis of NO_3^--N in soil solution samples was carried out using the phenol-disulfonic acid method. In the present work, additional measurements were carried out: Soil C:N ratio was determined as the quotient of organic matter and TKN contents; gravimetric moisture content (ω) was determined by oven drying a representative undisturbed ring of moist soil at 105 °C; bulk density (ρb) was determined by the weight of the soil per unit volume (g/cm^3) at 105 °C. Samplings and measurements regarding biomass and nutrient recovery across plant species are also presented in our previous study [3]. In brief, at the end of every growing season (October), one representative tree from each of four plots was harvested and separated into individual organs. The fresh weight of leaves, shoots and trunk (old wood) were recorded. For reeds, the whole plot surface was harvested each season and separated in leaves and shoots. Dry weights of vegetation were determined by drying (65 °C) to a constant weight. In 2002 and 2003 one replicate plot was harvested from each treatment and the tissue dry weights were determined. The dried samples were ground to 1-mm and used in elemental analysis. Micro-Kjeldahl N-digestion was used to determine total-N content of biomass samples [20] and P content was determined by the vanado-molybdo-phosphoric acid colorimetric method after digesting the samples with a mixture of perchloric-nitric acid.

2.2. Statistical Analysis

The effluent loading and some soil parameters were selected according to their importance in system description and ease of determination as candidate X variables in the PLSR model. This model contained plant biomass and N and P uptake as response variables (Ys). The soil parameters were SOM, dissolved

organic matter (as COD), TKN, pH, EC, soil solution NH_3-N (in soil solution sampler), soil solution P, soil solution EC, NO_3^--N, C:N, Olsen-P, Na^+, Ca^{2+}, Mg^{2+}, SAR, K^+, ρb, and ω. Most of these soil input variables were measured in each location across five depth intervals over the 0.65 m of soil profile and their average values were used in the initial PLSR model to eliminate some of the noise. After the preliminary PLSR regression the important Xs variables were identified and included in the final PLSR prediction model.

The non-linear iterative partial least squares (NIPALS) algorithm was used for computing the first few factors. KFold validation was used to select the number of factors that minimize the Root Mean PRESS statistic. The variable VIP (variable importance for the projection) value measures its influence on the factors that define the model. Wold and others advocated cut-off-values for VIP to separate terms that do not make important contribution to the dimensionality reduction involved in PLSR (VIP < 8) and those that might (VIP ≥ 8). In addition, the percentage of variation explained for X variables and Y responses and the contribution of each of the important factors were assessed. Loadings were also calculated and plotted to give another way to view the relationships between the Xs the Ys and the PLSR factors. For each X variable VIP (variable importance for the projection) was calculated to assess its importance in the determination of the PLSR projection model for both predictors and responses [21]. Also, Xs coefficients in the PLSR model were calculated to assess their contribution to the prediction of the Ys. Based on PLSR model the predicted values for the responses were calculated and plotted *versus* the observed values. Also, validation of PLSR prediction model was performed based on data derived from the previous year (2002). All analyses were carried out the PLS platform of JMP® (SAS Institute Inc.: Cary, NC, USA) Pro Version 11.2.1 [22].

3. Results

Preliminary PLSR regression removed several soil parameters with minor contribution to the prediction model. The remaining parameters were included in final PLSR model. These parameters were effluent loading, EC, available phosphorus (Olsen-P), Na^+, Ca^{2+}, Mg^{2+}, K^{2+}, and NO_3^--N. Based on PLSR method, the above X data set was reduced to two principal factors. The first explained the 45.4% of the variation while the second explained the 34.2%. Thus, the cumulative variation explained by two principal factors was 79.6%. The percentage of variation explained for X variables and the contribution of each of the two factors are shown in Figure 1a. Among X variables K^+, SAR, and Na^+ were the greater contributors to explained variation. Also, Mg^{2+}, EC, nitrates, effluent, Ca^{2+}, and P had distinctively higher factor loadings (X loadings) at the first factor compared to K^+, Na^+, and SAR. With regard to second factor SAR, Na^+, EC, and effluent received positive loadings and were higher than the other variables (Figure 1c). In terms of Y data (Y responses) two factors explained the 78.3% of the variation, with the first contributing with 70.4% and the second with 7.9%. Total biomass had the highest percentage of the explained variation followed by N uptake and P uptake (Figure 1b). In the first and strongest factor, plant biomass had the highest factor loading (Y loadings) and N uptake the lowest. In the second factor, N uptake received the highest factor loading over plant biomass and P uptake (Figure 1d).

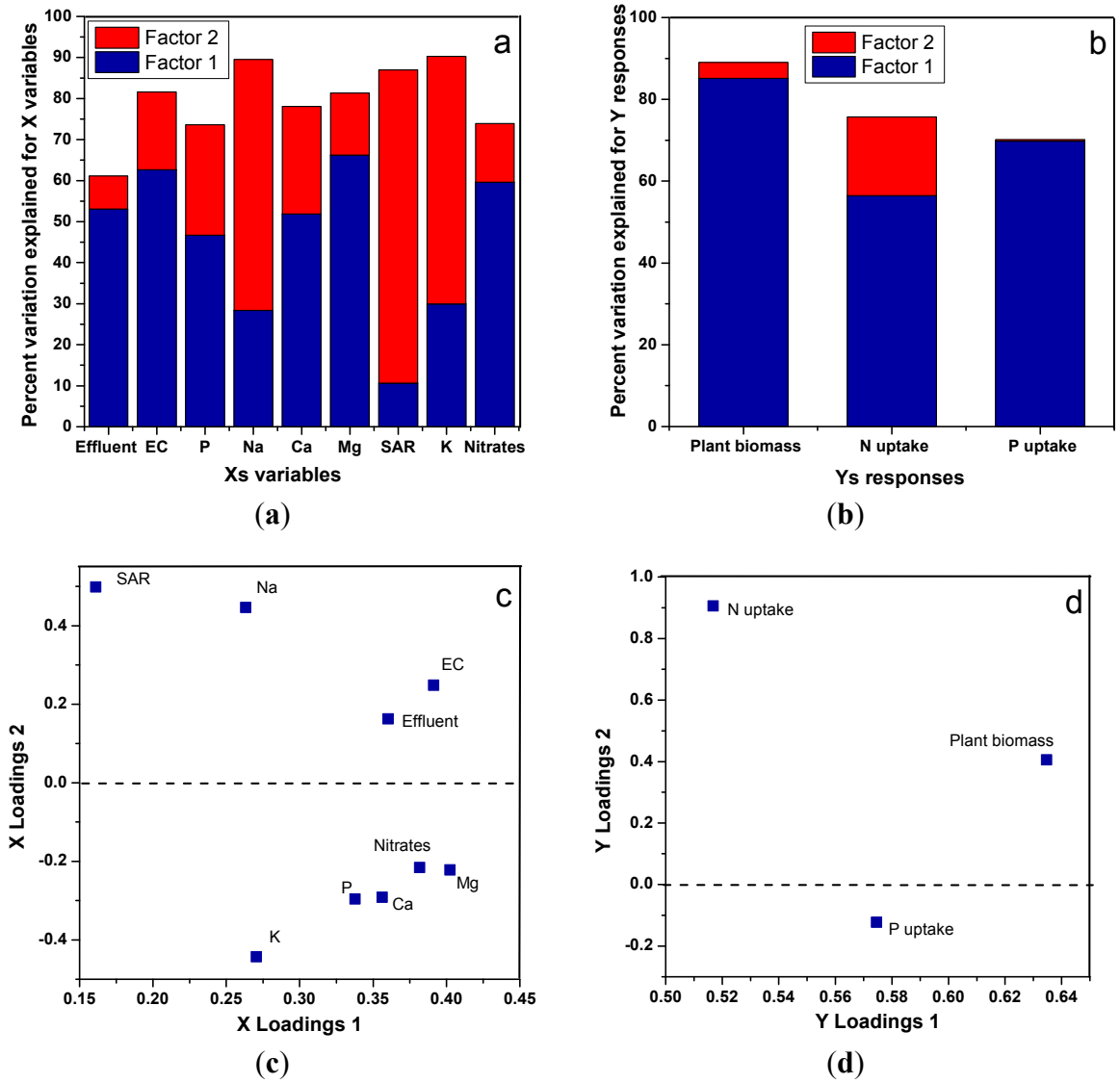

Figure 1. Explained variation and factor loadings for X variables (**a**,**c**) and Y responses (**b**,**d**).

The amount of the effluent, EC, nitrates and Mg^{2+} were among the most powerful X variables in the determination of PLSR model (Figure 2). Based on 0.8 threshold two X variables (SAR and K^{+}) were less influential in the final PLSR model (Figure 2), however, because of their importance for Xs factors (Figure 1a) and contribution in the prediction of Ys (Figure 3) were retained within the model. Effluent had the greater coefficient values across Ys prediction models followed by EC, Na^{+}, SAR, and/or nitrates dependent on the Y response (Figure 3). Specifically, apart from the amount of effluent, EC had also high coefficient values either for the prediction of biomass produced or prediction of the nutrients. Na^{+} and SAR also received high values with exception of the P recovery prediction model. Interestingly, nitrates received high values in both biomass and P recovery, but mid values were registered in terms of N recovery model. Predicted values derived from PLSR model were plotted over the observed values showing close relationship across all Y responses (Figure 4).

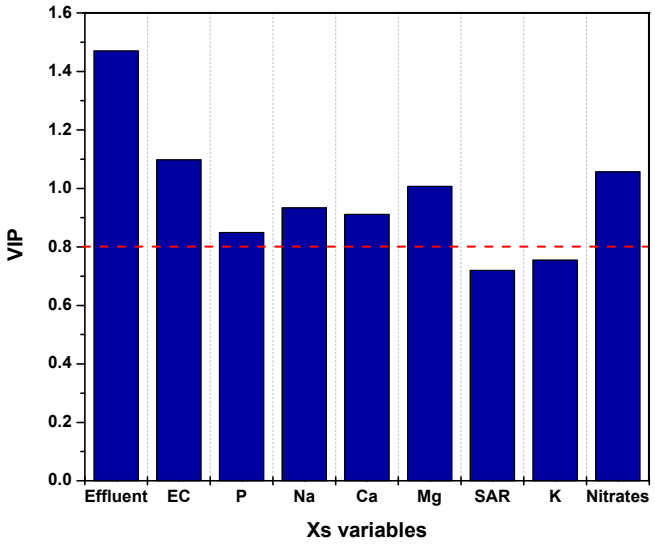

Figure 2. Important X variables in the Partial Least Squares Regression (PLSR) model.

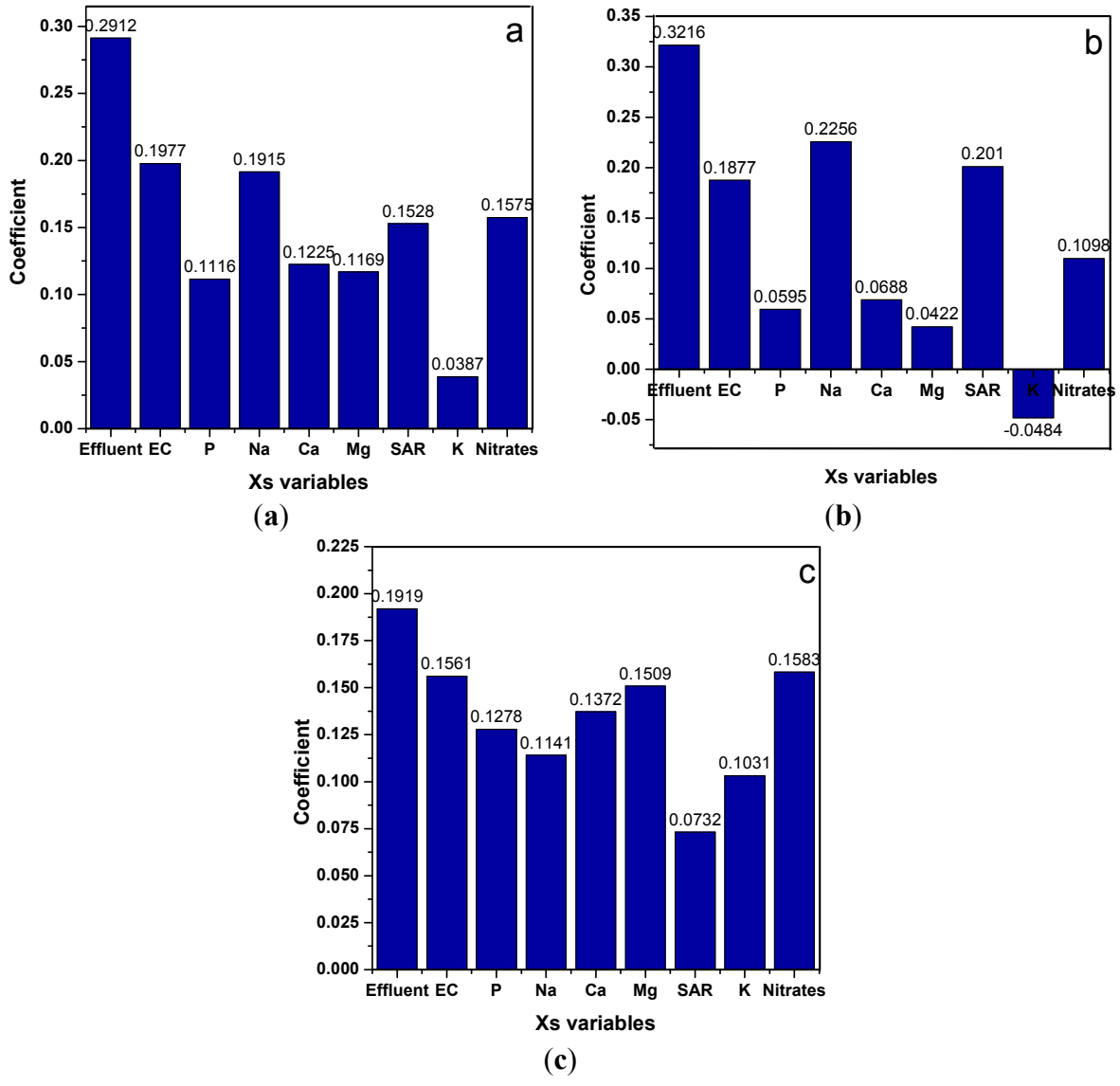

Figure 3. Coefficients of the Xs in the PLSR regression for the models of: (**a**) plant biomass; (**b**) N uptake; and (**c**) P uptake, based on the centered and scaled data.

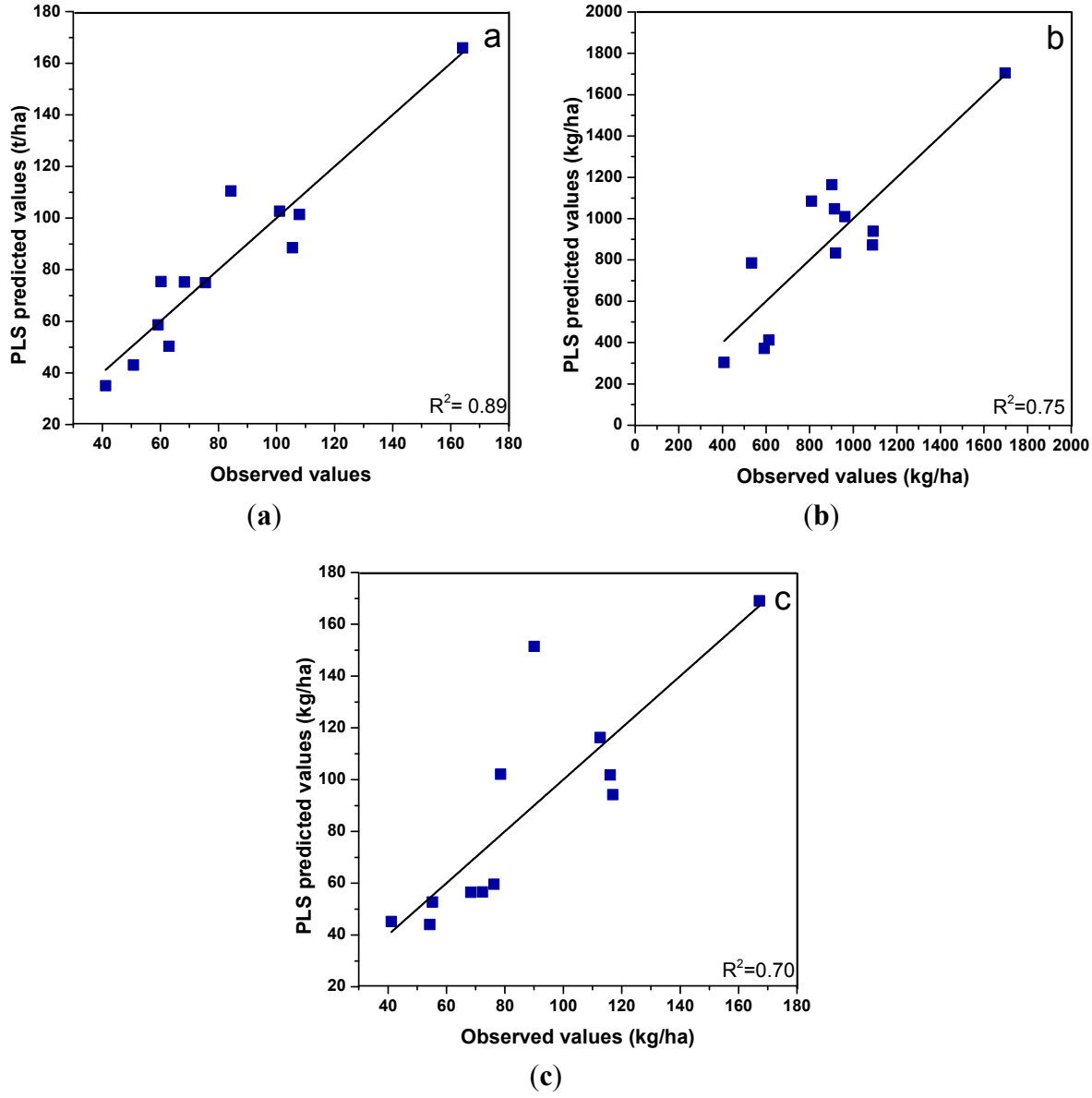

Figure 4. Observed values *versus* PLSR predicted values for: (a) plant biomass; (b) N uptake; and (c) P uptake.

4. Discussion

The PLSR model in this study is well fitted to the experimental data and successfully predicted the established Y responses (*i.e.*, plant biomass and P and N uptake). Important X variables were effluent loading and certain soil parameters such as soil salinity and nitrates. Effluent loading is considered essential component in the design of land application associated with organics and nutrient loading, growth of plant biomass, and potential environmental impacts [2]. In the present study it was applied at rates equivalent to the evapotranspiration requirements of plant species, which reasonably justify its close relationship with the response variables.

Soil salinity, expressed by EC, was also an important contributor to PLSR model showing strong and positive relationship to the response variables. Soil salinity is closely associated with effluent loading which in turn is associated with plant biomass. For example, species with great biomass potential receive

high effluent loading that may result in salts accumulation in the rhizosphere [4,6]. This effect, however, is temporary because of the effect of rainfall which removes salts from the rhizosphere reducing the possible negative impacts on soil and vegetation. EC also has been included in a previous PCR model predicting plant biomass from several soil parameters [23]. With regard to Na^+, it had also significant contribution to the prediction model, particularly for biomass and N uptake, as shown in Figures 2 and 3. SAR had lower contribution than Na^+ indicated also by lower VIP and relative smaller standardized regression coefficients. Increase in Na^+, and SAR values do not necessary imply risk for soil physical properties and/or crop yield considering the high salt and organic matter contents in the effluent. Indeed, in this field experiment effluent application resulted in enhancement in soil physical properties, presented in a previous study [4], an effect that is in agreement with the findings of previous studies [24,25].

Interestingly nitrate was important predictor in our model being among the most important X variables particularly for plant biomass and P uptake. This confirms previous arguments linking nitrates with plant biomass and effluent loading [1,3]. Inorganic N availability in the soil is considered important driver of plants growth with significant effect on biomass yield, tissues nutrient content, and total nutrient/pollutants assimilation [26–29]. The lower effect of nitrates in the model compared to effluent could be attributed to the existence of additional factors with significant influence on fate of inorganic N in the soil. In this field study inconsistent results were registered between the produced biomass and nitrates in *A. donax* species compared to other species [3,4], which is in agreement with the findings of a recent study [30]. Authors suggested the potential effect of species on soil microbial communities with significant role in N cycling. Moreover, N overloading, up to a critical threshold, and the favorable environmental conditions that induced high nitrification rates may have mitigated the contribution of nitrates in the model.

In this PLSR model important, yet easily measurable, parameters were included as predictor variables. These parameters were capable of providing important information for three specified response variables. In addition, PLSR model was validated by the data of the previous year, a fact, that increases further its potential for future use. The successful development of the prediction model in this study arises from the advantages of the PLSR method. In comparison to other statistical techniques PLSR can overcame the small or large number of highly collinear and noisy input variables and still provide reasonable prediction of multiple collinear plant response variables. Moreover, PLSR has greater reliability compared to other techniques (single multiple regression or combination of multiple regression with other multivariate methods) when identifying relevant variables and their magnitudes of influence, independently of the sample size used in the analysis [10]. On the other hand, probably there are limitations with regard to the applicability of the suggested model to other areas or different plant species [31]. Because of its advantages, PLSR method could also be ideal for the prediction of important management goals involving soil bio (chemical) parameters such as those involved in C and N cycles [30,32].

5. Conclusions

PLSR is highly compatible with intensive agronomic practices, such as land application, in which a large number of highly collinear and noisy input variables is monitored to assess plant species performance and detect potential impacts on soil and surrounding environment. In the present study a

robust PLSR model was developed to predict important management goals in land application (*i.e.*, plant biomass and N and P uptake), associated with treatment potential, environmental impacts, and benefits. Effluent loadings and several soil parameters commonly monitored in effluent irrigated lands, yet easily measurable, were included in the final PLSR model as predictor variables. Among these variables effluent, soil salinity (as EC), and nitrates had the greater effect on the model. PLSR model suggested in this study could help during the system design procedure and monitoring providing valuable information in terms of system performance and suggesting adjustments according to the management objectives. Monitoring of a large scale (spatial and temporal) data set of the pre-defined soil parameters can strengthen the reliability of the PLSR-drawn results, relevant recommendations, and future extrapolations.

Acknowledgments

This work was financed by the EU Coretech project: ICA 3-CT 1999-00012.

Author Contributions

Vasileios Tzanakakis performed the experimental work, the data collection, statistical analysis, and prepared the manuscript; Andy Mauromoustakos performed statistical analysis and reviewed the manuscript; and Andreas N. Angelakis had the original idea and supervised the research.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Tzanakakis, V.A.; Chatzakis, M.K.; Angelakis, A.N. Energetic environmental and economic assessment of three tree species and one herbaceous crop irrigated with primary treated sewage effluent. *Biomass Bioenergy* **2012**, *47*, 115–124.
2. Paranychianakis, N.V.; Angelakis, A.N.; Leverenz, H.; Tchobanoglous, G. Treatment of wastewater with slow rate systems: A review of treatment processes and plant functions. *Crit. Rev. Environ. Sci. Technol.* **2006**, *36*, 187–259.
3. Tzanakakis, V.A.; Paranychianakis, N.V.; Angelakis, A.N. Nutrient removal and biomass production in land treatment systems receiving domestic effluent. *Ecol. Eng.* **2009**, *35*, 1485–1492.
4. Tzanakakis, V.A.; Paranychianakis, N.V.; Londra, P.A.; Angelakis, A.N. Effluent application to the land: Changes in soil properties and treatment potential. *Ecol. Eng.* **2011**, *37*, 1757–1764.
5. Lado, M.; Ben-Hur, M. Treated domestic sewage irrigation effects on soil hydraulic properties in arid and semiarid zones: A review. *Soil Tillage Res.* **2009**, *106*, 152–163.
6. Leal, R.M.P.; Herpin, U.; Fonseca, A.F.D.; Firme, L.P.; Montes, C.R.; Melfi, A.J. Sodicity and salinity in a Brazilian Oxisol cultivated with sugarcane irrigated with wastewater. *Agric. Water Manag.* **2009**, *96*, 307–316.
7. Ruffo, M.L.; Bollero, G.A. Residue decomposition and prediction of carbon and nitrogen release rates based on biochemical fractions using principal-component regression. *Agron. J.* **2003**, *95*, 1034–1040.

8. Jabro, J.D.; Sainju, U.; Stevens, W.B.; Evans, R.G. Carbon dioxide flux as affected by tillage and irrigation in soil converted from perennial forages to annual crops. *J. Environ. Manag.* **2008**, *88*, 1478–1484.
9. Kulmatiski, A.; Beard, K.H.; Stevens, J.R.; Cobbold, S.M. Plant-soil feedbacks: A meta-analytical review. *Ecol. Lett.* **2008**, *11*, 980–992.
10. Carrascal, L.M.; Galván, I.; Gordo, O. Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos* **2009**, *118*, 681–690.
11. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.
12. Kehimkar, B.; Hoggard, J.C.; Marney, L.C.; Billingsley, M.C.; Fraga, C.G.; Bruno, T.J.; Synovec, R.E. Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis. *J. Chromatogr. A* **2014**, *1327*, 132–140.
13. Fonville, J.M.; Richards, S.E.; Barton, R.H.; Boulange, C.L.; Ebbels, T.M.D.; Nicholson, J.K.; Holmes, E.; Dumas, M.-E. The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *J. Chemom.* **2010**, *24*, 636–649.
14. Vohland, M.; Emmerling, C. Determination of total soil organic C and hot water-extractable C from VIS-NIR soil reflectance with partial least squares regression and spectral feature selection techniques. *Eur. J. Soil Sci.* **2011**, *62*, 598–606.
15. Nocita, M.; Stevens, A.; Toth, G.; Panagos, P.; van Wesemael, B.; Montanarella, L. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biol. Biochem.* **2014**, *68*, 337–347.
16. Fu, Y.; Yang, G.; Wang, J.; Song, X.; Feng, H. Winter wheat biomass estimation based on spectral indices, band depth analysis and partial least squares regression using hyperspectral measurements. *Comput. Electron. Agric.* **2014**, *100*, 51–59.
17. Li, F.; Mistele, B.; Hu, Y.; Chen, X.; Schmidhalter, U. Reflectance estimation of canopy nitrogen content in winter wheat using optimised hyperspectral spectral indices and partial least squares regression. *Eur. J. Agron.* **2014**, *52*, 198–209.
18. Wold, S.; Ruhe, A.; Wold, H.; Dunn, W.J., III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735–743.
19. Page, A.L. (Ed.). *Methods of Soil Analysis: Chemical and Microbiological Proerpties*; American Society of Agronomy: Madison, WI, USA, 1982.
20. Markus, D.K.; McKinnon, J.P.; Buccafuri, A.F. Automated analysis of nitrite, nitrate, and ammonium nitrogen in soils. *Soil Sci. Soc. Am. J.* **1985**, *49*, 1208–1215.
21. Cox, I.; Gaudard, M. *Discovering Partial Least Squares with JMP*; SAS Institute Inc.: Cary, NC, USA, 2013.
22. *JMP® Pro*, version 11.2.1; J. SAS Institute Inc.: Cary, NC, USA.
23. Mandal, U.K.; Warrington, D.N.; Bhardwaj, A.K.; Bar-Tal, A.; Kautsky, L.; Minz, D.; Levy, G.J. Evaluating impact of irrigation water quality on a calcareous clay soil using principal component analysis. *Geoderma* **2008**, *144*, 189–197.

24. Guo, L.B.; Sims, R.E.H. Soil response to eucalypt tree planting and meatworks effluent irrigation in a short rotation forest regime in New Zealand. *Bioresour. Technol.* **2003**, *87*, 341–347.
25. Yaron, B.; Dror, I.; Berkowitz, B. Contaminant-induced irreversible changes in properties of the soil-vadose-aquifer zone: An overview. *Chemosphere* **2008**, *71*, 1409–1421.
26. Christersson, L. Biomass production of intensively grown poplars in the southernmost part of Sweden: Observations of characters, traits and growth potential. *Biomass Bioenergy* **2006**, *30*, 497–508.
27. Labrecque, M.; Teodorescu, T.I. High biomass yield achieved by Salix clones in SRIC following two 3-year coppice rotations on abandoned farmland in southern Quebec, Canada. *Biomass Bioenergy* **2003**, *25*, 135–146.
28. Adegbidi, H.G.; Volk, T.A.; White, E.H.; Abrahamson, L.P.; Briggs, R.D.; Bickelhaupt, D.H. Biomass and nutrient removal by willow clones in experimental bioenergy plantations in New York State. *Biomass Bioenergy* **2001**, *20*, 399–411.
29. Guo, L.B.; Sims, R.E.H.; Horne, D.J. Biomass production and nutrient cycling in Eucalyptus short rotation energy forests in New Zealand. I: Biomass and nutrient accumulation. *Bioresour. Technol.* **2002**, *85*, 273–283.
30. Tsiknia, M.; Tzanakakis, V.A.; Paranychianakis, N.V. Insights on the role of vegetation on nitrogen cycling in effluent irrigated lands. *Appl. Soil Ecol*. 2013, 64, 104–111.
31. Goodhue, D.; Lewis, W.; Thompson, R. Small sample size, and statistical power in MIS research. In Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Kauai, HI, USA, 4–7 January 2006; Volume 8, p. 202b.
32. Tsiknia, M.; Tzanakakis, V.; Oikonomidis, D.; Paranychianakis, N.; Nikolaidis, N. Effects of olive mill wastewater on soil carbon and nitrogen cycling. *Appl. Microbiol. Biotechnol.* **2014**, *98*, 2739–2749.