Journal of
Risk and Financial Management

MDPI

Article

Semiparametric Time-Series Model Using Local Polynomial: An Application on the Effects of Financial Risk Factors on Crop Yield

Syed Ejaz Ahmed [1], Dursun Aydin [2] and Ersin Yilmaz [2,*]

[1] Department of Statistics, Faculty of Mathematics and Science, Brock University, St. Catharines, ON L2S 3A1, Canada; sahmed5@brocku.ca
[2] Department of Statistics, Faculty of Science, Mugla Sitki Kocman University, Muğla 48000, Turkey; duaydin@mu.edu.tr
* Correspondence: ersinyilmaz@mu.edu.tr

Abstract: This paper proposes a semiparametric local polynomial estimator for modelling agricultural time-series. We consider the modelling of the crop yield variable according to determined financial risk factors in Turkey. The derivation of a semiparametric local polynomial estimator is provided with its fundamental statistical properties to estimate the semiparametric time-series model. This paper attaches importance to precision agriculture (PA) and therefore a local polynomial technique is considered due to some advantages it has over alternative methods. The introduced estimator provides less estimation risk, involving both parametric and nonparametric components that allow the estimator to represent the data structure better. From that, it can be said that the proposed estimator and model is beneficial to agricultural researchers for financial decision-making processes.

Keywords: local polynomial regression; crop yield; financial risk; semiparametric time series

Citation: Ahmed, Syed Ejaz, Dursun Aydin, and Ersin Yilmaz. 2022. Semiparametric Time-Series Model Using Local Polynomial: An Application on the Effects of Financial Risk Factors on Crop Yield. *Journal of Risk and Financial Management* 15: 141. https://doi.org/10.3390/jrfm15030141

Academic Editor: Robert Hudson

Received: 18 February 2022
Accepted: 14 March 2022
Published: 16 March 2022

1. Introduction

In an agricultural context, profitability, sustainability, efficiency in resource usage, quality of production and managing decisions are supported by "precision agriculture" (PA) which involves methodologies, modeling tools and strategies to improve the quality of financial decision marking in the agricultural sector. Different methodologies of data modelling are used for PA, including regression models (Van de Putte et al. 2010), artificial neural networks (Shoshi et al. 2021; Kujawa and Niedbała 2021) and other machine learning techniques (Chlingaryan et al. 2018; Liakos et al. 2018). In addition, there are a lot of studies about PA from other fields of applied science, such as electronical engineering, chemistry, biology and other natural sciences. This paper accounts for the data modelling part through the use of a semiparametric time series model and local polynomial estimator, providing estimates with less risk and a more accurate representation of the data structure.

For the last decade, data science applications and modelling tools have been more frequently used in agricultural studies, leading to extensive literature on the subject. In the context of regression modelling, some important studies are as follows: Gonzalez-Sanchez et al. (2014) focused on accurate yield estimation based on machine learning methods and a linear regression model. Regarding the use of nonparametric methods in agricultural data analysis, Färe et al. (2013) presented a detailed review. Grigorios (2009) introduce a nonparametric regression-based kernel density estimator to represent the production function. Sam (2010) modeled the market risks of the agricultural futures of corn, soybeans and wheat by using a nonparametric kernel estimator. In addition, Zvizdojevic and Vukotic (2015), Ogundari and Brümmer (2011), Wang et al. (2016), Majumdar et al. (2017) and Shoshi et al. (2021) provide important contributions in the modeling of agricultural data using regression models and other machine learning techniques.

The studies given above involve using parametric and nonparametric analysis tools to model agricultural data. Note that parametric methods such as linear regression models require strict assumptions about data structure. Moreover, it should be emphasized that nonparametric estimators, unlike parametric approaches, are very flexible, however, their estimation quality and accuracy diminish greatly if several predictors are added to the model, which is known as the "curse of dimensionality". This paper considers a semiparametric time-series regression model to solve both problems: the curse of dimensionality and the need for strict assumptions regarding data structure. Therefore, avoiding the disadvantages of the two aforementioned regression models, the benefits of using semiparametric regression models, which combine the features of parametric and nonparametric models simultaneously, can be clearly seen. Although there are a number of studies about the use of semiparametric time series model in the literature, in applications pertinent to "precision agriculture", the lack of semiparametric techniques is evident. To go into detail, the parametric component of the model is interpreted as a linear regression model, while the nonparametric component allows flexibility from the strict structural assumptions associated with linear regression. Moreover, interpretation of the semiparametric model is easy and understandable. Some of the important studies on the semiparametric time series model are as follows: Kato and Shiohama (2009), Gao and Phillips (2010), and Aydın and Yılmaz (2021). These studies used kernel and spline-based estimators and applied these estimators to different application fields including econometrics, censored time-series data and medical applications.

In contrast to the studies mentioned above, the main purpose of this study is to contribute to PA by proposing a semiparametric local polynomial estimator (LPE) for modelling agricultural time-series data and to show the effects of the three main financial risk factors (currency exchange rates, foreign investment, and interest rates) on agricultural data. Note that if statistical importance and better qualified estimates are obtained using LPE for the response variable (i.e., crop yield values), it may provide a critical advantage in managing agricultural productivity and financial decision-making processes. It follows that the effective features of the response variable can be shown through the statistical significance of the parametric and nonparametric components of the model. The statistical properties of the LPE are derived in Section 3. Both the semiparametric time-series model and LPE can be easily understood and interpreted, which is beneficial for farm managers and researchers carrying out data analysis for the purpose of predicting sound financial decisions in the agricultural sector.

The data analyzed in this paper involves agricultural data and financial risk factors obtained from all over Turkey. The dataset contains data points from 1962 to 2020. Cereal yield (kg per hectare) is considered as the response variable. The determined predictors are official exchange rate (USD), foreign direct investment (% of GDP) and interest rate, as decided by the Central Bank of the Republic of Turkey. The nonparametric covariate of land used for cereal production is determined by comparing the relationship between land (km^2) and yield.

The organization of the paper is as follows: Section 2 offers a detailed overview of both the semiparametric time-series model and LPE. Section 3 provides the finite sample properties of the introduced LPE as well as the evaluation metrics to measure the quality of the LPE in modeling crop yield data. Section 4 is comprised of the analysis and results for the estimation of the semiparametric time-series model for the crop yield data. The parametric and nonparametric components of the model are presented individually. Finally, conclusions are described in Section 5.

2. Materials and Methods

2.1. Semiparametric Time-Series Model

Consider the semiparametric time-series model of the form

$$Y_t = \mathbf{X}_t\boldsymbol{\beta} + f(z_t) + \varepsilon_t,\ t = 1 \leq t \leq n \tag{1}$$

where Y_ts are the values of response variable, which is stationary time-series, $\mathbf{X}_t = (X_{1t}, \ldots, X_{pt})$ is a $(n \times p)$-dimensional matrix of predictors in time t, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $(p \times 1)$-dimensional vector of regression coefficients, $f(z_t)$ is an unknown smooth function to be estimated based on values of nonparametric variable z_ts and finally, ε_t's are the stationary autoregressive error terms given by

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \tag{2}$$

where ρ is an autocorrelation parameter and u_ts are independent and identically distributed (i.i.d.) random error terms with $u_t \sim N(0, \sigma^2_{u_t})$ and $|\rho| < 1$. Note that if $\rho = 0$, model (1) becomes an ordinary semiparametric regression model.

2.2. Local Polynomial Estimator (LPE)

Assuming that $\boldsymbol{\beta} = 0$ in model (1) and $\rho = 0$ in (2), model (1) turns into a classical nonparametric regression model. Following from Fan et al. (1997), by applying the local polynomial regression technique in a neighborhood of z_0, a regression function $f(z_t)$ can be approximated locally by a polynomial of order q. Based on Taylor's expansion in z_t at a neighbourhood of z_0, the q^{th} degree polynomial approximation of $f(z_t)$ can be expressed as

$$f(z_t) \approx \sum_{j=0}^{q} \frac{f^{(j)}(z_0)}{j!}(z_t - z_0)^j = \sum_{j=0}^{q} b_j(z_t - z_0)^j \tag{3}$$

From (3), this polynomial expression is locally fitted using weighted least squares based on kernel methods, meaning local polynomial regression. However, in this paper, a semiparametric time-series model is considered. Therefore, the LPE based on weighted least squares should be obtained by minimizing the following criterion (4) due to autocorrelation adjustment and parametric component terms. Accordingly, the local weighted least squares criterion can be defined as follows:

$$\min_{b,\boldsymbol{\beta}} \sum_{i=1}^{n} \mathbf{R}\left\{ Y_i - \sum_{j=0}^{q}(z_t - z_0)^j b_j - \mathbf{X}_t'\boldsymbol{\beta} \right\}^2 K\left(\frac{z_t - z_0}{h}\right) \tag{4}$$

where $\boldsymbol{\Sigma}$ is $(n \times n)$-dimnesional symmetric and a positive-definite covariance matrix required to solve the autocorrelation problem between the autoregressive error terms $\boldsymbol{\varepsilon}_t \sim \boldsymbol{N}_n(0, \boldsymbol{\Sigma})$ given in (2). The elements of $\boldsymbol{\Sigma}$ are calculated as follows:

$$\boldsymbol{\Sigma} = \frac{\sigma_u^2}{1-\rho^2}\mathbf{R}, \;\; R_{i,j} = \rho^{|i-j|}, \;\; 1 \le (i,j) \le n. \tag{5}$$

In practice, $\boldsymbol{\Sigma}$ is unknown, but it is needed for the estimation. In order to make simple the illustration of the estimation procedure, we assume that $\boldsymbol{\Sigma}$ is known. $K(.)$ is a kernel function to compute the weights of each z_t point and h is the bandwidth parameter that controls the size of the local neighborhood of z_0. Accordingly, semiparametric LPEs $(\hat{b}_j, \hat{\boldsymbol{\beta}})$ of $(b_j, \boldsymbol{\beta})$ can be obtained by minimizing (4). It should be noted that in vector and matrix notation (4) can be written as follows:

$$\min_{b,\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{B}\mathbf{b} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{B}\mathbf{b} - \mathbf{X}\boldsymbol{\beta})^T \tag{6}$$

where $\mathbf{Y} = (Y_1, Y_2 \ldots, Y_n)^T$, $\mathbf{b} = (b_0, \ldots, b_q)^T$, $\mathbf{B}$ and $\mathbf{X}$ matrices are written as

$$\mathbf{B} = \begin{bmatrix} 1 & (z_1 - z_0) & \cdots & (z_1 - t_0)^p \\ 1 & (z_2 - z_0) & \cdots & (z_2 - t_0)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (z_n - z_0) & \ldots & (z_n - z_0)^p \end{bmatrix} \text{ and } \mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}$$

After solving Equation (6) with respect to the derivative of $\mathbf{b}$, some algebraic operations denote that the estimate $\hat{\mathbf{b}}$ is given by

$$\hat{\mathbf{b}} = \left(\mathbf{B}'\boldsymbol{\Sigma}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{7}$$

Following from the Taylor expansion in (3), it is necessary to select the first element of the vector $\hat{\mathbf{b}} = (\hat{b}_0, \ldots, \hat{b}_q)$ for obtaining $\hat{f}(z_0) = \hat{b}_0$. Note that this leads to the semiparametric LPE of the nonparametric time-series function for an arbitrary point z_0. Thus, similar to Speckman (1988), the estimate of function can be defined as

$$\hat{f}_h(z_0; h) = \sum_{i=1}^{n} \omega_i\prime\left(B_i{}'\Sigma_i\mathcal{Q}_i\right)^{-1}\mathcal{Q}_\mathrm{i}{}'\Sigma_i^{-1}\left(Y_i - \mathbf{X}_i'\boldsymbol{\beta}\right) = \mathbf{S}_h(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{8}$$

where $\mathbf{S}_h = \boldsymbol{\omega}_1{}'\left(\mathbf{B}'\boldsymbol{\Sigma}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}'\boldsymbol{\Sigma}^{-1}$ denotes the local polynomial smoother matrix (or penalty matrix), and $\boldsymbol{\omega}_1' = (1, \mathbf{z}, \ldots, \mathbf{z}^q)$ is a $\mathbb{R}^{(p+1)}$ dimensional matrix having 1 in the first position.

By using the local polynomial smoothing matrix $\mathbf{S}_h$ given after (8), the partial residuals can be computed to estimate the regression coefficients of the parametric component of the semiparametric time-series model. Hence, minimization criterion (6) is rewritten as a weighted least squares based on the partial residuals, as follows:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \Sigma_i^{-1}\left\{\tilde{Y}_i - \tilde{\mathbf{X}}_i'\boldsymbol{\beta}\right\}^2 = \sqrt{\boldsymbol{\Sigma}^{-1}}\left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)^2 \tag{9}$$

where $\boldsymbol{\Sigma}$ is defined as in (5), $\tilde{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{S}_h)\mathbf{Y}$ and $\tilde{\mathbf{X}} = (\mathbf{I}_n - \mathbf{S}_h)\mathbf{X}$. Also note that $\mathbf{I}_n$ is an $(n \times n)$ identity matrix. Thus, the LPE of $\hat{\boldsymbol{\beta}}$ can be obtained by minimizing (9). It is defined as follows:

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{Y}} \tag{10a}$$

After this step, when the vector $\hat{\boldsymbol{\beta}}$ found in (10a) is substituted into (8), the estimate of $\hat{\mathbf{f}}$ corresponding to the nonparametric component of the model based on the LPE can be obtained by

$$\hat{\mathbf{f}} = \mathbf{S}_h\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \tag{10b}$$

where $\mathbf{S}_h$ is defined in (8). The vector of fitted values is defined as

$$\mu = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}} = \mathbf{H}_h\mathbf{Y} = \hat{\mathbf{Y}} \tag{10c}$$

where $\mathbf{H}_h = \mathbf{S}_h + (\mathbf{I}_n - \mathbf{S}_h)\mathbf{X}\left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}(\mathbf{I}_n - \mathbf{S}_h)$.

3. Statistical Properties and Evaluation Metrics

In this section, the finite sample properties of the proposed LPE are discussed. Regarding the parametric component of the model, Equation (10a) is expanded to show the bias and variance of $\hat{\boldsymbol{\beta}}$. Note that partial residuals are needed here; these are defined after Equation (9). Because the model involves autoregressive error terms, both bias and variance involve the $\boldsymbol{\Sigma}$ matrix. Before the calculations are made, some assumptions are needed in order to obtain accurate bias and variance of the regression coefficients. These assumptions are as follows:

A1. Regression function $f(.)$ is bounded its second partial derivative.

A2. Matrix of parametric covariates $\left(X_1, \ldots, X_p\right)^T \in \mathbb{R}^p$ have a continuous density function $d(.)$.

A3. $Cov(\hat{\boldsymbol{\beta}})$ is bounded as $\sup\left(Cov(\hat{\boldsymbol{\beta}})\right) < \infty$.

A4. Standard assumptions of Kernel function $K(.)$ are ensured. These are: $K(.)$ is a continuous bivariate kernel function and $\int K(u)du = 1$.

A5. To provide the asymptotic normality, $Cov(\hat{\boldsymbol{\beta}})$ is bounded away from zero as $0 < \inf\left(Cov(\hat{\boldsymbol{\beta}})\right)$, which is considered together with A3. The bias and variance of the regression coefficients are presented in Theorem 1 under these assumptions.

Theorem 1. *Assume that (A1)–(A5) are ensured. The expanded form of* $\hat{\boldsymbol{\beta}}$ *is thus given by:*

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{Y}}^{*}_{\hat{G}} = \left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}(\mathbf{I}_n - \mathbf{S}_h) \\ \mathbf{Y} &= \boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}} + \left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\left(\mathbf{I} - \mathbf{S}^{DL}_{hi}\right)\boldsymbol{\varepsilon} \end{aligned} \tag{11}$$

where $\tilde{\mathbf{f}} = (\mathbf{I} - \mathbf{S}_{\mathrm{h}})\mathbf{f}$ *is the partial residuals for* $\mathbf{f}$*. From (11), the bias and covariance matrix of* $\hat{\boldsymbol{\beta}}$ *can be inferenced easily as follows:*

$$Bias\left(\hat{\boldsymbol{\beta}}\right) = E\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}} \tag{12}$$

$$Cov\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left[\left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}(\mathbf{I}_n - \mathbf{S}_h)^2\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}\right)^{-1}\right] \tag{13}$$

Also, if $n \to \infty$, $\left|Bias(\hat{\boldsymbol{\beta}})\right| \equiv O_p\left(n^{-1/2}\right)$ *and* $Cov(\hat{\boldsymbol{\beta}}) \equiv O_p(n^{-1})$.

Proof of Theorem 1 is given in Appendix A. Theorem 2 is provided below to demonstrate the distribution of $\hat{\boldsymbol{\beta}}$.

Theorem 2. *Assume that (A1)–(A5) are confirmed. Let* $\phi(.)$ *be the distribution function of the standard normal distribution and* $M = \sqrt{Cov(\hat{\boldsymbol{\beta}})}$*. Accordingly, the following expressions can be written:*

$$P\left(\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\mathbf{M}} \leq \eta\right) = \phi(\eta) + o_p(1), \text{ when } n \to \infty$$

Here, this result shows that whether smooth function $f(.)$ *does or does not exist in the model, the estimate of* $\hat{\boldsymbol{\beta}}$ *has a* $\sqrt{n}$*-convergence to* $\boldsymbol{\beta}$.

Proof of Theorem 2 is given in Appendix B.

Note that the bias and variance–covariance matrix of the regression coefficients given in (12) and (13) are used a measurement tool to evaluate the behaviors of the LPE in crop yield data modelling. Moreover, note that model variance σ^2 is generally unknown. Therefore, an estimate of σ^2 is used, calculated as follows:

$$\hat{\sigma}^2 = \frac{\left(\mathbf{Y} - \hat{\mathbf{Y}}\right)^T\left(\mathbf{Y} - \hat{\mathbf{Y}}\right)}{tr(\mathbf{H})}$$

where $\mathbf{H}$ and $\hat{\mathbf{Y}}$ are given in (10c).

In addition, root mean squared (RMSE) scores for the nonparametric component estimation are calculated as:

$$RMSE\left(\mathbf{f}, \hat{\mathbf{f}}\right) = \sqrt{n^{-1}\sum_{j=1}^{n}\left[f(z_j) - \hat{f}(z_j)\right]^2} = \sqrt{\left(\mathbf{f} - \hat{\mathbf{f}}\right)^T\left(\mathbf{f} - \hat{\mathbf{f}}\right)} \tag{14}$$

After the parametric and non-parametric components, two criteria popular in the time-series literature are introduced to show the performance of the LPE for the semiparametric time-series model. These criteria are given below:

$$MARE = n^{-1}\sum_{t=1}^{n}|Y_t - \hat{Y}_t|/|Y_t|,\ MAPE = n^{-1}\sum_{t=1}^{n}|Y_t - \hat{Y}_t|/Y_t \tag{15}$$

Hence, the performance of the LPE can be evaluated by using fitted values from the time-series model. Note that semiparametric LPE estimator shows its difference by involving both parametric and nonparametric components. These feature makes LPE more flexible than its conventional alternatives such as linear estimators or autoregressive models. In this context, from our point of view, the effects of the financial risk factors on the crop yield are represented by LPE estimator better than existing methods.

In addition, Table 1 is presented below to provide some basic information about the data of interest. Detailed information about the data is given in Section 4. Table 1 involves the descriptive statistics of the variables.

Table 1. Descriptive statistics of cereal yield dataset.

	$\log(Yield_t)$	$Exchange_t$	$Foreign_t$	$Interest_t$	$Land_t$
Min	7.02	0.00	0.02	8.00	16.20
Max	8.10	1.9	3.62	87.79	16.46
Range	1.08	1.9	3.60	79.79	0.26
Median	7.66	0.00	0.54	22.78	16.40
Mean	7.62	0.52	1.06	33.68	16.38
SE of mean	0.04	0.09	0.12	2.93	0.01
Std. Deviation	0.31	0.69	0.92	22.49	0.07
Coef. Variation	0.04	1.32	0.88	0.67	0.00

4. Analysis of the Effects of Financial Risk Factors on Crop Yield

As mentioned in Section 1, a cereal yield dataset collected between years 1962 and 2020 is modelled by the introduced LPE. The dataset was collected from the following website: https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?locations=TR (accessed on 28 January 2022). The cereal yield (kg/hectare) (yield) variable is considered as a response variable to be explained using a multiple predictor time-series model. Note here that there are lots of potential predictors to model the yield variable. However, this study focuses on some of the main financial risk factors that are explained below in detail. In this section, the important predictors are determined according to linear correlation between the predictors and the response variable. The nonparametric covariate of the model is decided by observing its scatter plot versus the yield. Accordingly, explanatory variables for both the parametric and nonparametric components of the model are listed as follows:

Covariates generating the parametric components:

- Official exchange rate (USD annually average)—*Exchange*;
- Foreign direct investment (% of GDP)—*Foreign*;
- Interest ratio—*Interest*.

Notice that the variable names used in the semiparametric time-series model are given on the right side of the list.

Figure 1 is obtained by using R software. Figure 1 displays the correlations between the parametric covariates and yield, which is used as the response variable. It should also be noted that panel (a) in Figure 1 shows the scatter plots for each combination of variables, as well as the density plots for each variable and the graph giving the correlations between the variables, while panel (b) displays the correlogram showing the strength of the correlations between the variables.

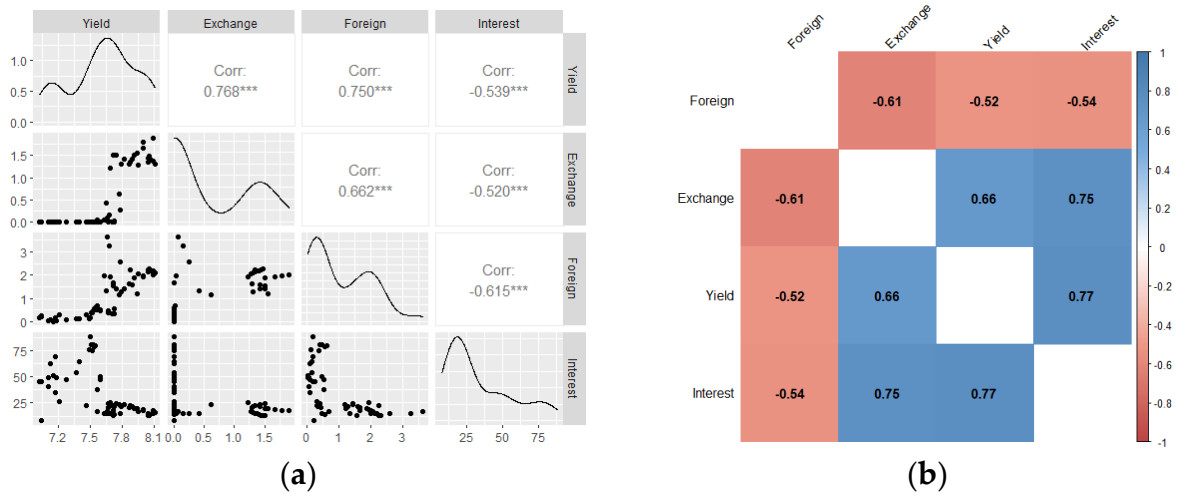

Figure 1. Pairs of parametric covariates and correlogram with response variable $\log(Yield)$. (**a**) Pairs with densities and correlations. (**b**) Correlogram. Asterisks (***) denote that correlations are statistically significant in 95% confidence level.

Nonparametric covariate:

- Land under cereal production (km^2)—*Land*.

The reason for choosing "Land under cereal production" as a nonparametric covariate can be clearly seen in Figure 2. It seems that there is a clear nonlinear relationship between response variable yield.

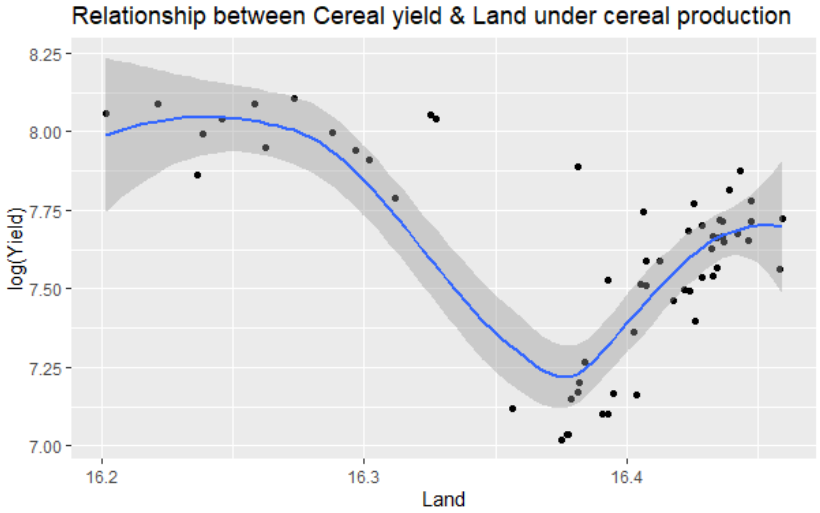

Figure 2. Relationship between *Land* and $\log(Yield)$ with a smooth curve.

From the information given above, the semiparametric time-series model is written as follows:

$$\log(Yield_t) = \beta_1 Exchange_t + \beta_2 Foreign_t + \beta_3 Interest + f(Land_t) + \varepsilon_t \quad (16)$$

where $t = 1, \ldots, 59$ and ε_t are autoregressive error terms, as defined in Equation (2). Here, the vector of regression coefficients can be notated as $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ and their LPE estimate is then $\hat{\boldsymbol{\beta}}$. Similarly, if $\mathbf{f} = (f(land_1), \ldots, f(land_{59}))^T$ is specified as a vector, and its LPE estimate is expressed as $\hat{\mathbf{f}}$. Note that one of the commonly used methods in the time-series literature for obtaining a model for $yield_t$ responses is an autoregressive (AR) model. Therefore, an AR model is used as a benchmark method and the quality of each of the two models is compared. Note that Aydın and Yılmaz (2021) have previously discussed a similar comparison. The Dickey–Fuller test is applied to determine the optimum lag for the AR model and the results are shown in Table 2.

Table 2. Augmented Dickey–Fuller test for stationary of the *yield*.

No. Lag	ADF Statistic	*p*-Value
0	−4.9713	$p < 0.01$ *
1	−3.2139	0.0938
2	−2.7427	0.2706
3	−2.187	0.4898

* The null hypothesis that the series is non-stationary is rejected at 95% confidence level.

It can be seen from Table 1 that the yield series are stationary without lag when the trend coefficient is added to the model. In order to represent data, this paper considers an $AR(2)$ model according to AIC criterion, and model coefficients are estimated as $\gamma_1 = 0.662$ and $\gamma_2 = 0.264$. Thus, the $AR(2)$ model is given in (17) as

$$yield_t = 0.662(yield_{t-1}) + 0.264(yield_{t-2}) + u_t \quad (17)$$

where u_ts are normally distributed with a constant variance (white noise). Additional results of the analysis are provided in following figures and tables.

The following tables (Tables 3 and 4) contain the scores for measuring the quality of both the semiparametric time series (based on LPE) and the $AR(2)$ models, respectively. The minimum scores are indicated with an asterisk. Obviously, the LPE gives the minimum mean absolute percentage error (MAPE), mean absolute relative error (MARE) and model variances. This means that the best estimates are obtained with LPE, not $AR(2)$. Note that the performance values of $AR(2)$ are not too distant from those of LPE. However, regarding overall model performance, LPE appears to be superior.

Table 3. Values of evaluation metrics form the estimated semiparametric and $AR(2)$ models.

	MAPE	MARE	$\hat{\sigma}^2$
LPE	0.1404 *	0.5177 *	0.2280 *
$AR(2)$	0.1597	0.5357	0.4397

*: The best performance scores.

Table 4. Bias and variances of regression coefficients obtained from LPE.

	$\hat{\beta}_1 = 2.376$	$\hat{\beta}_2 = 1.678$	$\hat{\beta}_3 = 0.114$
$Var\left(\hat{\boldsymbol{\beta}}\right)$	0.105	0.034	0.594
$Bias\left(\hat{\boldsymbol{\beta}}\right)$	0.573	0.396	0.055

Table 4 shows the bias and variances of the parametric component estimated using the LPE. Here, $\hat{\beta}_1$ denotes the effect of $Exchange_t$ to the model, $\hat{\beta}_2$ shows the contribution of $foreign_t$ and $\hat{\beta}_3$ denotes the effect size of $Interest_t$ on the variables. According to the values of the regression coefficients, it can be seen that $Exchange_t$ affects the response variable the most. This variable is highly dependent on the sensitive economical structure of Turkey. Interestingly, foreign investment seems to have an effect on the yield, however, we cannot say for certain what is the correlation between these two variables. In Turkey, higher foreign investment may indirectly affect the price of fertilizers, animal feed and pesticides, which are the main agricultural expenditures. Therefore, the yield may be positively affected by foreign investment. On the other hand, while the interest ratio in the country seems to have little effect on the yield, this is not significant enough proof to reject its long-term effect on the yield or other important agricultural indicators. Therefore, the relationship between agricultural indicators and the interest ratio should be closely inspected.

Figure 3 contains the bar graph of the scores given in Table 3. With the exception of the MARE criterion, the performance of the LPE and $AR(2)$ models are similar. However, in general, the LPE solves the targeted modeling problem more efficiently than the traditional $AR(2)$ model, which is an expected result because the semiparametric time-series model includes the non-parametric component. The success of the LPE in this context is illustrated in Figure 4. An estimated curve of $f(Land_t)$ and its 95% confidence interval are also given in this figure. The RMSE value for the estimated curve is 0.4192. Figure 4 also shows that the estimate obtained from LPE satisfactorily represents the data.

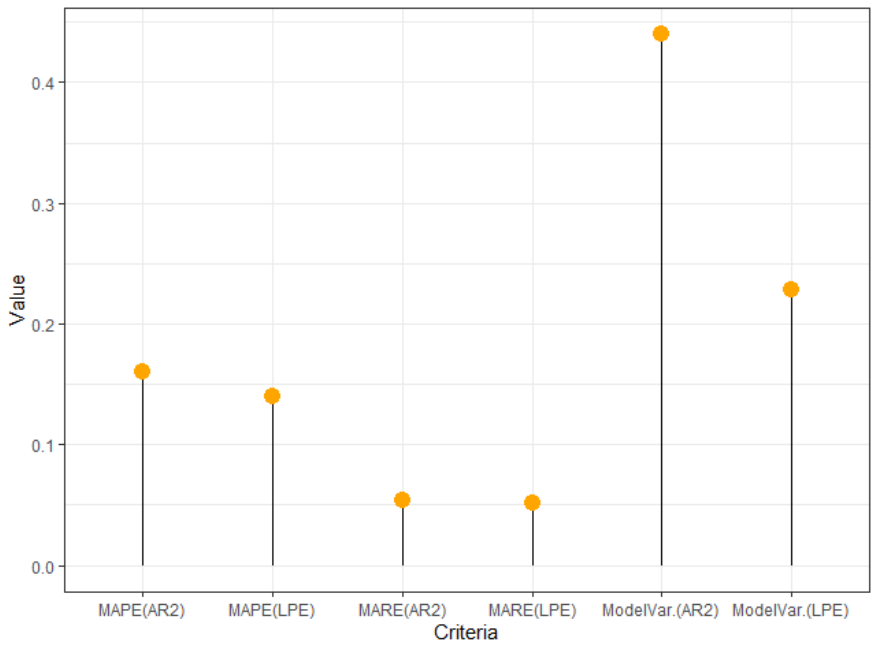

Figure 3. Bar graph representation of the values in Table 3.

As can be seen in Figure 4, the relationship between $\log(Yield_t)$ and $Land_t$ is clearly nonlinear. Due to the AR model using linear modelling structure, it cannot catch the pure nonlinear relationships such as $\log(Yield_t)$ and $Land_t$. In this context, the merit of the introduced LPE estimator should be emphasized because it can represent both linear and nonlinear relationships between the variables.

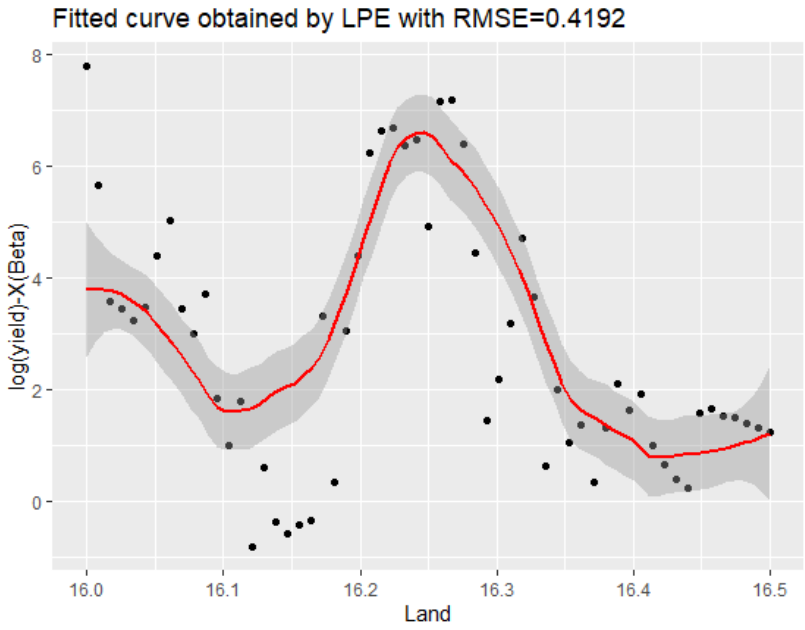

Figure 4. Fitted curve $f(Land_t)$ against $(\log(Yield) - \mathbf{X\beta})$ for the nonparametric component with RMSE score.

5. Conclusions

This paper discusses modelling crop yield data using a new semiparametric estimator based on a local polynomial estimation model, LPE. A Turkish cereal yield dataset is considered as the real-world data example. The results are given in Section 4. In order to determine the accuracy of the LPE's performance, the AR model, which is the traditional method used in modeling time-series data in the literature, is used as a benchmark and the results of a comparison between the two methods are presented. The results given in Tables 3 and 4 and Figures 3 and 4 show that the proposed LPE estimator gives satisfactory estimates for both the parametric and nonparametric components. In this context, it can be said that using the LPE estimator for agricultural data successfully models crop yield with less risk. Although this paper considers the AR model as a benchmark method, crop yield prediction is studied by several authors based on different estimation techniques. For instance, Chandio et al. (2020) investigated the effects of climate change factors in cereal yield in Turkey between the dates 1968–2014. Differently from our paper, they considered linear regression to estimate the cereal yield. Although linear models are a widely used method for modelling time series, they cannot catch the nonlinear effects of the explanatory variables, which means the semiparametric estimator reduces the risk. Some similar studies can be ordered as follows: Çakır et al. (2014) used artificial neural networks to estimate the cereal yield in Turkey. Moreoever, Chandio et al. (2021) inspected modelling cereal production for different phenomena. By considering the methods given in the mentioned studies, a detailed comparison study can be made to show explicitly the behaviors of the introduced semiparametric estimator and the alternative nonparametric and semiparametric estimation methods in future research. On the other hand, the LPE estimator involves only one nonparametric component, and its performance depends on the optimal bandwidth parameter. Accordingly, more than one nonparametric component limits the LPE estimator, and its calculation process is more complicated than the AR and linear models. However, the performance of LPE can tolerate these disadvantages.

Author Contributions: Conceptualization, S.E.A. and D.A.; methodology, S.E.A. and D.A.; software, E.Y.; validation, S.E.A., D.A. and E.Y.; formal analysis, E.Y.; investigation, E.Y.; resources, D.A.; data curation, E.Y. writing—original draft preparation, S.E.A. and D.A.; writing—review and editing, D.A. and E.Y.; visualization, E.Y.; supervision, S.E.A.; project administration; funding acquisition, S.E.A. All authors have read and agreed to the published version of the manuscript.

Funding: The research of S. Ejaz Ahmed was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset was collected from the following website: https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?locations=TR (accessed on 28 January 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Theorem 1

As mentioned in Section 2, $\hat{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^T\widetilde{\mathbf{Y}}$. Assume that assumptions (A1)–(A5) are ensured and $\|(\mathbf{I}_n - \mathbf{S})\mathbf{f}\| = \|\tilde{\mathbf{f}}\| = O_p(n^{-1/2})$ By using this information and Equation (11), if $n \to \infty$ the following expression can be written as

$$\left|Bias(\hat{\boldsymbol{\beta}})\right| = \left|(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}}\right| \leq (\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}} \left|Bias(\hat{\boldsymbol{\beta}})\right| \equiv O_p\|n^{-1/2}\|^2 \quad \text{(A1)}$$

According to A3, $Cov(\hat{\boldsymbol{\beta}})$ depends on a constant v and can be written as follows:

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2[(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\boldsymbol{\Sigma}(\mathbf{I}_n - \mathbf{S}_h)^2\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{X}})^{-1}] \leq vn^{-2}(\mathbf{I}_n - \mathbf{S})\tilde{\mathbf{X}}^2 \quad \text{(A2)}$$

Here, $\|(\mathbf{I}_n - \mathbf{S})\tilde{\mathbf{X}}^2\| \leq 2\|\tilde{\mathbf{X}}\|^2 + 2\|\mathbf{S}^T\tilde{\mathbf{X}}\|^2$. If it is assumed that $\|\tilde{\mathbf{X}}\|^2 = O_p(n)$, then the right side of (A2) is given by

$$n^{-2}\|(\mathbf{I}_n - \mathbf{S})\tilde{\mathbf{X}}\|^2 = O_p(n^{-1}) \quad \text{(A3)}$$

Thus, proof of Theorem 1 is completed

Appendix B. Proof of Theorem 2

Using the inequality $\|(\mathbf{I}_n - \mathbf{S})\tilde{\mathbf{X}}\|^2 \leq 2\|\tilde{\mathbf{X}}\| + 2\|\mathbf{S}^T\tilde{\mathbf{X}}\|^2$ and (A3), the following is obtained

$$\|(\mathbf{I}_n - \mathbf{S})\tilde{\mathbf{X}}\|^2 \leq \frac{1}{2}\|\tilde{\mathbf{X}}\|^2 + 2\|\mathbf{S}^T\tilde{\mathbf{X}}\|^2 = \frac{1}{2}\|\tilde{\mathbf{X}}\|^2\left[1 + o_p(1)\right] \quad \text{(A4)}$$

by using (A2) and the following equation:

$$\lim_{n\to\infty} P(Cov(\hat{\boldsymbol{\beta}})) \geq v(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1} = 1 \quad \text{(A5)}$$

Then, from Equations (12) and (A2),

$$\frac{\left|\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{f}}\right|}{\mathbf{M}} = o_p(1) \quad \text{(A6)}$$

Thus, when $n \to \infty$, the asymptotic normality of $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ can be written in (A7),

$$P\left(\frac{(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T(\mathbf{I}_n - \mathbf{S}_h)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f})}{\mathbf{M}} \leq \eta\right) = P\left(\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\mathbf{M}} \leq \eta\right) = \phi(\eta) + o_p(1) \quad \text{(A7)}$$

References

Aydın, Dursun, and Ersin Yılmaz. 2021. Semiparametric modeling of the right-censored time-series based on different censorship solution techniques. *Empirical Economics* 61: 2143–72. [CrossRef]

Çakır, Yüksel, Kırcı Mürvet, and Güneş Ece Olcay. 2014. Yield prediction of wheat in south-east region of Turkey by using artificial neural networks. Paper presented at the 2014 The Third International Conference on Agro-Geoinformatics, Beijing, China, August 11–13; pp. 1–4.

Chandio, Abbas Ali, Jiang Yuansheng, Akram Waqar, Adeel Sultan, Irfan Muhammad, and Jan Inayatullah. 2021. Addressing the effect of climate change in the framework of financial and technological development on cereal production in Pakistan. *Journal of Cleaner Production* 288: 125637. [CrossRef]

Chandio, Abbas Ali, Ozturk Ilhan, Akram Waqar, Ahmad Fayyaz, and Mirani Aamir Ali. 2020. Empirical analysis of climate change factors affecting cereal yield: Evidence from Turkey. *Environmental Science and Pollution Research* 27: 11944–57. [CrossRef] [PubMed]

Chlingaryan, Anna, Sukkarieh Salah, and Whelan Brett. 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture* 151: 61–69. [CrossRef]

Fan, Jianqing, Theo Gasser, Irène Gijbels, Michael Brockmann, and Joachim Engel. 1997. Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics* 49: 79–99. [CrossRef]

Färe, Rolf, Shawna Grosskopf, and Gerald Whittaker. 2013. Directional output distance functions: Endogenous directions based on exogenous normalization constraints. *Journal of Productivity Analysis* 40: 267–69. [CrossRef]

Gao, Jiti, and Peter Phillips. 2010. *Semiparametric Estimation in Simultaneous Equations of Time-Series Models (No. 2010–26)*. Adelaide: University of Adelaide, School of Economics.

Gonzalez-Sanchez, Alberto, Frausto-Solis Juan, and Ojeda-Bustamante Waldo. 2014. Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research* 12: 313–28. [CrossRef]

Grigorios, Iordanou. 2009. Flat-Plate Solar Collectors for Water Heating with Improved Heat Transfer for Application in Climatic Conditions of the Mediterranean Region. Ph.D. thesis, Durham University, Durham, UK; pp. 71–73.

Kato, Risa, and Takayuki Shiohama. 2009. Model and variable selection procedures for semiparametric time-series regression. *Journal of Probability and Statistics* 2009: 487194. [CrossRef]

Kujawa, Sebastian, and Gniewko Niedbała. 2021. Artificial Neural Networks in Agriculture. *Agriculture* 11: 497. [CrossRef]

Liakos, Konstantinos G., Busato Patrizia, Moshou Dimitrios, Pearson Simon, and Bochtis Dionysis. 2018. Machine learning in agriculture: A review. *Sensors* 18: 2674. [CrossRef] [PubMed]

Majumdar, Jharna, Naraseeyappa Sneha, and Ankalaki Shilpa. 2017. Analysis of agriculture data using data mining techniques: Application of big data. *Journal of Big Data* 4: 1–15. [CrossRef]

Ogundari, Kolawole, and Bernhard Brümmer. 2011. Technical efficiency of Nigerian agriculture: A meta-regression analysis. *Outlook on Agriculture* 40: 171–80. [CrossRef]

Sam, Abdoul G. 2010. Nonparametric estimation of market risk: An application to agricultural commodity futures. *Agricultural Finance Review* 70: 285–97. [CrossRef]

Shoshi, Humayra, Hanson Erik, Nganje William, and SenGupta Indranil. 2021. Stochastic Analysis and Neural Network-Based Yield Prediction with Precision Agriculture. *Journal of Risk and Financial Management* 14: 397. [CrossRef]

Speckman, Paul. 1988. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)* 50: 413–36. [CrossRef]

Van de Putte, Kobeke, Nuytinck Jorinde, Stubbe Dirk, Le Huyen Thanh, and Verbeken Annemieke. 2010. Lactarius volemus sensu lato (Russulales) from northern Thailand: Morphological and phylogenetic species concepts explored. *Fungal Diversity* 45: 99–130. [CrossRef]

Wang, Ruoyu, Laura C. Bowling, and Keith A. Cherkauer. 2016. Estimation of the effects of climate variability on crop yield in the Midwest USA. *Agricultural and Forest Meteorology* 216: 141–56. [CrossRef]

Zvizdojevic, Jelena, and Milica Vukotic. 2015. Application of statistical methods in analysis of agriculture-correlation and regression analysis. *Poljoprivreda I Sumarstvo* 61: 309. [CrossRef]