Peer Review Report

REVIEWERS:

ⓘD Paulo Hubert Júnior (Fundação Getulio Vargas, EAESP, Brazil)
And one anonymous reviewer.

ROUND 1

Reviewer 1 report

Reviewer 1 for this round chose not to disclose his/her review report.

Reviewer 2 report

Reviewer: Paulo Hubert Júnior

Date review returned: February 19, 2021

Recommendation: Major revision

Comments to the authors

The paper is very interesting and proposes a novel method to calculate insurance premiums for soybean crops in Brazil. The methods, however, are not clearly stated throughout the work, and clarification is needed, specially for publication in a journal that aims at a wide audience. Comments below.

The abstract introduces an acronym (SPI) without first defining it, which would be the best practice.

Page 7, Line 53: first sentence ("If $p_t(z)$ os the check loss function...") seems to be incomplete

Page 7, Line 56: "For a location (u_t, v_t), let d_{it}..." - actually d_{it} depends on thwo locations, not one.

Page 8, Line 58: "obtaining the solution of" - obtaining the minimum point of

Page 10, Line 40: "for any dp" - it is not clear what does this mean. Any dp = any infinitesimal variation in p?

Page 10 line 44: "which become more serious in the case of distributions". It is unclear what is this case of distributions, it would be good to clarify this sentence.

Page 11, first paragraph: the authors mention a Bayesian bootstrap procedure, but it is not clear from context what procedure is this that must be done before computing the semi-deviation. It would be important to clarify this paragraph.

Page 11, equation for U_{it} and V_i: please explicitly define variables W_{it} and W_i. What do they represent? Also, it is not clear what authors mean by "semi-deviation method" being "expressed by σ_{ssd} and U_{it}". Which of these values is taken as the semideviation?

Page 12, line 43: what interest units are being clustered? Why?

Page 13, line 36: it is actually the probability distribution that is fitted ("adjusted") to the data, not the other way around. Also, the authors point out that McKLee used the Gamma. Did the authors also use the Gamma, or adopted another model?

Page 14: "insurance premium is derived from the probability distributino function of indemnities". How are indemnities obtained?

Page 14: maybe it would be nice to provide time series plots of SPI and also soybean yields. The correlation was calculated from the original time series, or from differentiated time series? It is not adequate to calculate correlation between non-stationary time series, and it is not possbile to understand from the text wether the series were stationary or not. It would also be illustrative to see the values of these correlations for different time ranges of the SPI.

Page 14, line 43: "probability distribution function (pdf)". Are the indemnities a discrete random variable? If not it would be more adequate to refer to a probability density function.

Page 14: please clarify how the HBA method works and how it differs from HDA and Monte Carlo based methods.

Page 15: "thus it may misrepresent variability at the farm level.' It is not clear why this is the case, and what data will be used at farm level of aggregation.

Page 15: how are the indemnities I_{it} calculated?

Page 15: "a cross-validation method would be ideal". How then are the hyperparameters tuned? Please clarify how does the Bayesian bootstrap replace the cross validation procedure, and how exactly it was used.

Page 16: it is unclear what are the scenarios tested. First one WII + GWQLASSO x no insurance. Second scenario: WII insurance designed with GWQLASSO or quantile regression (which one?) against what? Third scenario: what is an yield insurance and how does it compare to parametric insurance?

Page 16: optimal number of clusters is two, but it is not clear what individuals are being clustered here.

It would be good to include the model's equations around page 16 so the reader can better understand the model's form. It it not clear, for instance, what is the response variable for these models.

I am not completely convinced that the usage of detrended time series allows the model to be built without regard for temporal structure. Are the time series stationary? If the response variable are the yields in a given year, why not use as explanatory the values of each month's SPI at that same year? It is difficult to review this part of the paper as it is unclear what exactly is the model form. Also, the boxplots show high variability in the coefficients' values, which indicates that they cannot be reasonably understood as different estimations of the same coefficient as the paper seems to suggest.

All these issues prevent us to adequately understand and review the papers' conclusions.

Additional Questions:

Does the manuscript contain new and significant information to justify publication?: Yes

Does the Abstract (Summary) clearly and accurately describe the content of the article?: Yes

Is the problem significant and concisely stated?: No

Are the methods described comprehensively?: No

Are the interpretations and conclusions justified by the results?: No

Is adequate reference made to other work in the field?: Yes

Is the language acceptable?: Yes

Does the article have data and / or materials that could be made publicly available by the authors?: Yes

Please state any conflict(s) of interest that you have in relation to the review of this paper (state "none" if this is not applicable).:

Rating:

Interest: 1. Excellent

Quality: 2. Good

Originality: 1. Excellent

Overall: 2. Good

Authors' Responses

Dear Mr. Mendes-da-Silva,

We thank you and the reviewers for the contributions made to this article. Our responses (R) to each suggestion (S) follow:

Reviewer: 1

The authors' responses to the comments of Reviewer 1 for this round were omitted from this report, since the reviewer did not authorize the disclosure of his/her report.

Reviewer: 2

(S) The abstract introduces an acronym (SPI) without first defining it, which would be the best practice.

(R) Altered.

(S) Page 7, Line 53: first sentence ("If $p_t(z)$ os the check loss function...") seems to be incomplete

(R) Altered.

(S) Page 7, Line 56: "For a location (u_t, v_t), let d_{it}..." - actually d_{it} depends on thwo locations, not one.

(R) Altered.

(S) Page 8, Line 58: "obtaining the solution of" - obtaining the minimum point of

(R) Altered.

(S) Page 10, Line 40: "for any dp" - it is not clear what does this mean. Any dp = any infinitesimal variation in p?

(R) Altered.

(S) Page 10 line 44: "which become more serious in the case of distributions". It is unclear what is this case of distributions, it would be good to clarify this sentence.

(R) Removed.

(S) Page 11, first paragraph: the authors mention a Bayesian bootstrap procedure, but it is not clear from context what procedure is this that must be done before computing the semi-deviation. It would be important to clarify this paragraph.

(R) Removed without impairing the understanding of the formula in question, its use, as well as the need for the Bayesian bootstrap is detailed on page 17 in the first and last paragraphs.

(S) Page 11, equation for U_{it} and V_i: please explicitly define variables W_{it} and W_i. What do they represent? Also, it is not clear what authors mean by "semi-deviation method" being "expressed by σ_{ssd} and U_{it}". Which of these values is taken as the semideviation?

(R) Clarified. The semideviation must be utilized within a utility framework. In this paper we utilized the mean-variance base utility function, deriving the mean-semideviation utility function.

(S) Page 12, line 43: what interest units are being clustered? Why?

(R) Added to the article. In all cases municipalities are being clustered in order to identify spatial patterns of climate and yield risk.

(S) Page 13, line 36: it is actually the probability distribution that is fitted ("adjusted") to the data, not the other way around. Also, the authors point out that McKLee used the Gamma. Did the authors also use the Gamma, or adopted another model?

(R) Altered. Yes, we also use the Gamma distribution, added to the article.

(S) Page 14: "insurance premium is derived from the probability distributino function of indemnities". How are indemnities obtained?

(R) Added specific section "Payout structure" to the article.

(S) Page 14: maybe it would be nice to provide time series plots of SPI and also soybean yields. The correlation was calculated from the original time series, or from differentiated time series? It is not adequate to calculate correlation between non-stationary time series, and it is not possbile to understand from the text wether the series were stationary or not. It would also be illustrative to see the values of these correlations for different time ranges of the SPI.

(R) While we also find it useful for the reader to see time plots of the series the sheer amount of data, 42 time series of 30 years for each variable, prevent us from doing so. However, the interested reader may contact us in order to get our dataset.

Regarding the correlation, we calculated it from the detrended yields (which were detrended as previously detailed) and the SPI time series, which are all stationary according to the Augmented Dickey-Fuller test.

(S) Page 14, line 43: "probability distribution function (pdf)". Are the indemnities a discrete random variable? If not it would be more adequate to refer to a probability density function.

(R) Altered.

(S) Page 14: please clarify how the HBA method works and how it differs from HDA and Monte Carlo based methods.

(R) Clarified.

(S) Page 15: "thus it may misrepresent variability at the farm level.' It is not clear why this is the case, and what data will be used at farm level of aggregation.

(R) Fitting probability distributions to historical values of the proposed index generally tends to underestimate the real variability of the index and thus the correct value of the premium.

(S) Page 15: how are the indemnities I_{it} calculated?

(R) Added.

(S) Page 15: "a cross-validation method would be ideal". How then are the hyperparameters tuned? Please clarify how does the Bayesian bootstrap replace the cross-validation procedure, and how exactly it was used.

(R) Misplaced, it should be in the yield-index modeling section. A cross-validation method would be ideal for not only tuning the hyperparameters but also ensuring the yield-index relationship is consistent, however, the available computational resources are not up to the task.

(S) Page 16: it is unclear what are the scenarios tested. First one WII + GWQLASSO x no insurance. Second scenario: WII insurance designed with GWQLASSO or quantile regression (which one?) against what? Third scenario: what is an yield insurance and how does it compare to parametric insurance?

(R) Corrected, description of yield insurance added.

(S) Page 16: optimal number of clusters is two, but it is not clear what individuals are being clustered here.

(R) Municipalities are being clustered in order to identify the spatial behavior of both precipitation and yield risk.

(S) It would be good to include the model's equations around page 16 so the reader can better understand the model's form. It it not clear, for instance, what is the response variable for these models. I am not completely convinced that the usage of detrended time series allows the model to be built without regard for temporal structure. Are the time series stationary? If the response variable are the yields in a given year, why not use as explanatory the values of each month's SPI at that same year? It is difficult to review this part of the paper as it is unclear what exactly is the model form. Also, the boxplots show high variability in the coefficients' values, which indicates that they cannot be reasonably understood as different estimations of the same coefficient as the paper seems to suggest. All these issues prevent us to adequately understand and review the papers' conclusions.

(R) Regarding the stationarity of the time series, yes, all of them are stationary as yields are detrended and SPI series despite assuming stationarity were tested with the Augmented Dickey-Fuller test.

Yes, each yield value is matched with the respective SPIs for the same crop year, however, the model itself does not accommodate error correlation structures in its covariance matrix or other types of structures regarding the time component.

The general model form is presented at the beginning of section 2.1 "Geographically Weighted Quantile LASSO".

The high variability in coefficients is expected given that each one represents the relationship between the index and the yield in a given year for all municipalities in the same cluster. The objective here is to identify if the coefficients are different from zero and positive or negative.