Data-Driven Defenses: Refining Pathogen Predictive Models for Global Agriculture Via Neural Networks

Theodore John Stronkowsky IV
University of Florida
April 20, 2024

Contents

1 Prologue

In 2023, my research journey began when Dr. Aashish Adhikari, a leading expert in plant pathology, approached me with a pressing question within his field: how could we develop a mathematical model that effectively integrates economic and environmental factors to assess the invasion potential of nodes in an epidemiological network? At Garrett Labs, we have long grappled with sparse data issues. To address this, we employed an inverse power law with the pest extent parameter in the numerator and distance in the denominator, scaled by an attenuation factor typically set at 2. While this model has been broadly effective, it remains incomplete. Given the dominant role of human behavior and environmental factors in pathogen spread, we hypothesized that incorporating trade data and meteorological data would significantly enhance our model. Our chosen test case was Magnaporthe Oryzae, or Rice Blast—a devastating fungus capable of destroying up to 40% of rice crop yields, making it ideal for our study due to the extensive trade data available for rice.

2 Introduction

Our investigation into the optimal mathematical models for assessing invasion potential began with the development of three distinct models, each designed to capture different aspects of epidemiological dynamics. The first model, referred to as the ”tri model,” integrates environmental factors and trade data with our foundational inverse power law. The second, the ”duo model,” incorporates trade data into the inverse power law to assess how commercial exchanges influence pathogen spread. Finally, the ”single model” employs our original inverse power law, serving as a baseline for comparison against the enhanced models.

Following the establishment of these models, we proceed to calculate the invasion potential for each, utilizing these calculations as inputs for a neural network designed to classify the extent of pest infestation. This approach is inspired by significant research in the field, specifically drawing on the structural, computational, and methodological insights provided by Fenu and Malloci [1] and the empirical distribution functions related to climatic influences on Magnaporthe oryzae as detailed by Tabonglek et al. [2]

These foundational papers provided a dual framework for our study: Fenu and Malloci's [1] work guided our approach to model structure and the implementation of artificial neural networks, along with providing crucial preprocessing tips and data references. Simultaneously, Tabonglek et al. offered a detailed examination of how environmental factors such as precipitation and temperature correlate with the dispersion of Magnaporthe oryzae spores, which significantly informed the environmental components of our models. [2]

3 Data Preparations

3.1 Data Scavenging

To streamline our data acquisition process, I consulted Jacobo Robeldo, a PhD student specializing in plant pathology, who is well-versed in the data sources relevant to our research. His recommendations were invaluable in identifying and obtaining the necessary datasets:

- **FAO STAT**: Provided comprehensive trade data, essential for our models that incorporate economic variables.
- **The World Bank**: Offered annual precipitation and temperature data, crucial for analyzing environmental impacts on pathogen spread.
- **CABI and EPPO**: Supplied data on pest extent, which is labeled annually by country and region.

Unfortunately, these datasets for pest extent were not readily accessible in an extractable format. Many were only available on archived web pages, requiring the use of the Internet Archive to access historical data representations. Consequently, data had to be manually transcribed, making this the most labor-intensive aspect of the project, as numerous data rows had to be meticulously entered into our database.

3.2 Data Preprocessing

The next phase involved loading and processing the data within R. A significant challenge was that the data was initially formatted with different years represented as columns. To address this, I performed a "long pivot" transformation, which restructured the data by consolidating yearly data into single columns for each feature. This restructuring not only facilitated easier data manipulation but also required standardizing overlapping columns that contained both country and year data to ensure consistency across all datasets.

3.3 Data Availability

It is crucial to acknowledge the limitations imposed by our reliance on data from developing nations. The precision and frequency of the data available to us are often beyond our control, impacting the robustness of our models. For instance, we were restricted to using only yearly pest extent reports, which complicates the accurate modeling of invasion potential due to the seasonal nature of rice farming. Ideally, our models would benefit from more frequent updates, such as seasonal data, but we were compelled to adjust our models to work with annual meteorological data to align with the available pest extent reports. Additionally, a significant challenge with the pest extent data is its incompleteness and the predominance of certain classes, such as the 'present but no detail' class, denoted by the integer 2, which introduces further complexity into our analysis.

4 Model Refinement

Through the study of Tabonglek et al.'s research on the modeling of spore dispersal of *Magnaporthe oryzae* [2], we refined our models' functions for precipitation and temperature, key environmental factors influencing invasion potential. The paper utilized differential equations to describe spore dispersal dynamics. Although we lacked specific data or an explicit function to replicate these dynamics completely, we extracted essential principles that guided our model adjustments.

The research indicated that temperatures between 25°C and 28°C are optimal for spore dispersion, enhancing pathogen spread. Similarly, an extended dew period, correlated with higher precipitation levels, significantly increases spore dispersal. In response to these insights, we implemented:

- A Gaussian function for temperature, centered at 26.5°C, to model the optimal temperature range for spore activity.
- A sigmoid function for precipitation, to model the relationship between increasing rainfall and spore dispersal rates.

These function implementations aim to more accurately reflect the natural dynamics of pathogen spread based on environmental conditions, thereby improving the predictive accuracy of our invasion potential model. How we will test the accuracy of these models is how well their output works as a feature to predict pest extent labels. That is the entire point of invasion potential.

4.1 Temperature-Dependent Spread Function

The temperature-dependent infection efficiency is modeled as a Gaussian function:

$$\phi_T(t) = e^{-k\frac{(T(t)-T_{\text{opt}})^2}{2\sigma^2}}$$

where $T_{\text{opt}} = 26.5$°C and σ is the standard deviation.

4.2 Precipitation-Dependent Spread Function

The infection efficiency based on the dew period is modeled using a sigmoid function:

$$\phi_P(t) = \frac{1}{1 + e^{-k(P(t)-P_{\text{crit}})}}$$

where $P_{\text{crit}} = 0$mm (this value has the potential to change in the future the more we discover about this relation) and k determines the curve's steepness.

4.3 Inverse Power Law Function for Invasion Potential

The infection potential is described by an inverse power law function that takes into account the infection rate, host availability, pest presence, extent of pest distribution, and the distance from nodes, defined as:

$$\text{Tri-model Invasion Potential} = (1+\phi_{\text{T}}(t))\times(1+\phi_{\text{P}}(t))\times\sum(\frac{(1+Q(t))\times(1+E(T\leq t))}{(d)^a})$$

$$\text{Duo-model Invasion Potential} = (\sum(\frac{(1+Q(t))\times(1+E(T\leq t))}{(d)^a})$$

$$\text{Single-model Invasion Potential} = (\sum(\frac{(1+E(T\leq t))}{(d)^a})$$

Where each component is defined as follows:

- $\phi_T(t)$ is the invasion rate function dictated by temperature for our import country at time t.
- $\phi_P(t)$ is the invasion rate function dictated by precipitation for our import country at time t.
- $Q(t)$ is the trade quantity from our export country to our import country at time t.
- $E(T \leq t)$ denotes the pest extent of our export country for all countries that have a pest extent for time $T \leq t$.
- d is the distance from the export country to the import country.
- a is the attenuation value, which determines how rapidly the invasion potential decreases with increasing distance from the nodes.

5 Implementation of Multi-Classification Model

Our initial step in the modeling phase involved the development of several key functions to accurately calculate invasion potential. These included:

- **Precipitation Function:** Utilizes precipitation data as a key variable.
- **Temperature Function:** Incorporates temperature data to assess its influence on pathogen dynamics.
- **Invasion Potential Calculator:** This core function takes two inputs—the country in question and the year of interest. It is designed to recognize pest labels from countries from which imports have occurred, for years up to and including the year assessed. This approach leverages historical data to infer pest presence, essential for predicting pathogen dynamics in an epidemiological framework.

For the neural network structure, we initially configured a model with four layers, each containing 100 neurons activated by the ReLU function, and trained over 100 epochs. This setup was intended to capture the high non-linearity inherent in epidemiological data networks. The model also included seven output neurons, corresponding to the multilabel nature of our classification task, utilizing a softmax activation function for output. Upon further evaluation, the model was simplified to two layers with 50 neurons each, undergoing 30 epochs of training. This adjustment was made after discovering that the simpler model achieved comparable results with significantly reduced complexity.

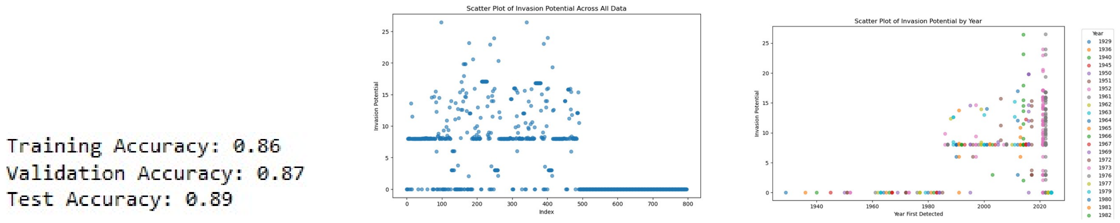
5.1 Results of Multi-Classification Model

The testing phase yielded surprising results. All model variants, despite varying complexities, demonstrated similar accuracy levels, and we identified a persistent, irreducible error of 11%. Initially, this was perplexing to both my advisors and myself. However, further analysis revealed an underlying pattern: The distribution of our calculated invasion potentials, when plotted against indices and years, showed consistent groupings of data points with multiple separations. This pattern was most distinct in the single model, suggesting that simpler models might more effectively capture the core dynamics we sought to model.

Moreover, models with additional features did not improve separation but introduced 'free-floating' data points, suggesting that these features might be adding noise rather than informative variability. The fact that our model achieved an 89% testing accuracy suggests it has captured a sufficient variance to be practical, yet still offers room for improvement, particularly in generalizing to less biased datasets. Notably, the prevalent representation of the 'pest extent class' labeled by the integer 2 highlighted potential data bias.

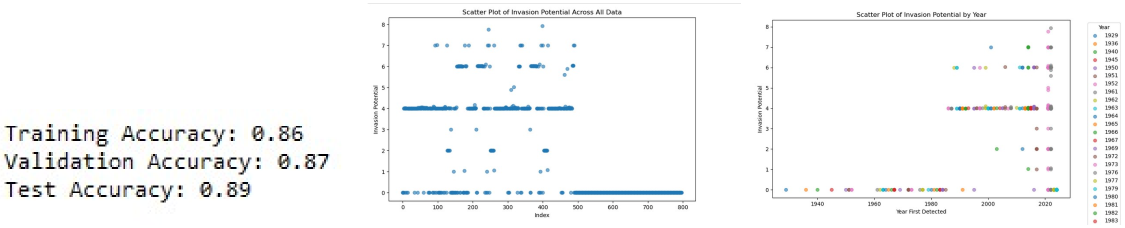
A critical insight from our multi-label classification testing was the realization that fewer data points could yield similar predictive results. This finding is particularly valuable given the challenges of sourcing extensive agricultural data from developing regions. While these results did not fully resolve our initial questions, they prompted us to consider alternative modeling approaches to further enhance our understanding and prediction of pathogen spread.

5.2 Multi-label Classification figures



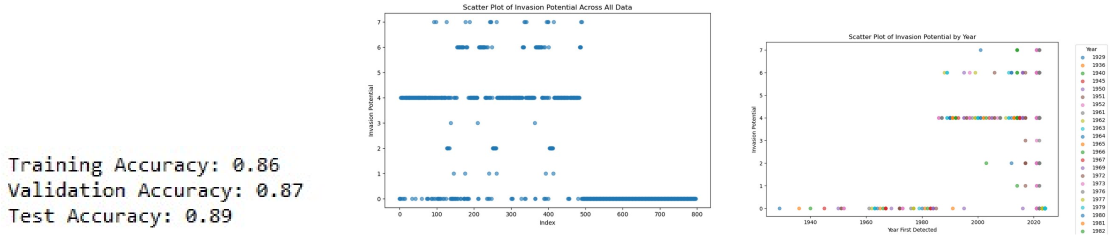
```
Training Accuracy: 0.86
Validation Accuracy: 0.87
Test Accuracy: 0.89
```

Figure 1: Visualizations for the Tri Model



```
Training Accuracy: 0.86
Validation Accuracy: 0.87
Test Accuracy: 0.89
```

Figure 2: Visualizations for the Duo Model



```
Training Accuracy: 0.86
Validation Accuracy: 0.87
Test Accuracy: 0.89
```

Figure 3: Visualizations for the Single Model

6 Implementation of Regression Model and Rounded Classification

Our approach to developing the regression model mirrored the initial steps used in our classification model, starting with defining meteorological functions and designing the invasion potential. However, for the regression model, we tempered our expectations and began with a simpler structure: two layers, each with 50 neurons. We incrementally increased complexity until reaching four hidden layers with 100 neurons each. During this process, we observed that the mean squared error (MSE), mean error (ME), and mean absolute error (MAE) plateaued, indicating minimal gains in model performance with increased complexity, leading us to halt further expansions.

For this regression task, we utilized a single output node with the ReLU activation function to capture the nonlinear behavior of our features. We chose MSE, ME, and MAE as our metrics for a comprehensive analysis of model performance. MSE helped penalize larger errors significantly, MAE provided insight into the typical size of errors, and ME was instrumental in detecting any bias in our predictions. The split of training, validation, and testing data remained consistent with the multi-label classification setup.

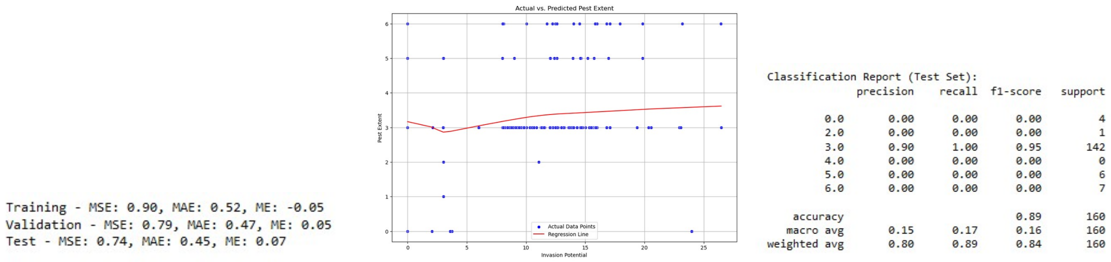
Furthermore, to add an additional layer of analysis, we rounded the predicted values from the regression model and then treated the task as a classification. This approach allowed us to compare performance across different modeling strategies more directly.

6.1 Results of Regression Model and Rounded Classification

As illustrated in the figures b, the single model and duo model marginally outperformed the tri-model in the regression tasks. However, the differences were minimal, and regression lines indicated similar performance across all models. The most notable findings emerged when we treated the rounded regression outcomes as classification results. Here, the tri-model and the single model both achieved a testing accuracy of 89%, hitting the upper limit of our irreducible error.

Interestingly, the tri-model slightly excelled in this rounded classification setting, showing a marginal 1% improvement in F1-score, recall, and precision over the other models. This improvement suggests that the tri-model could potentially offer better performance in terms of true positives and correct positive rates as our datasets scale. But essentially all the mathematical models have approximately the same predictive power.

6.2 Regression Model figures



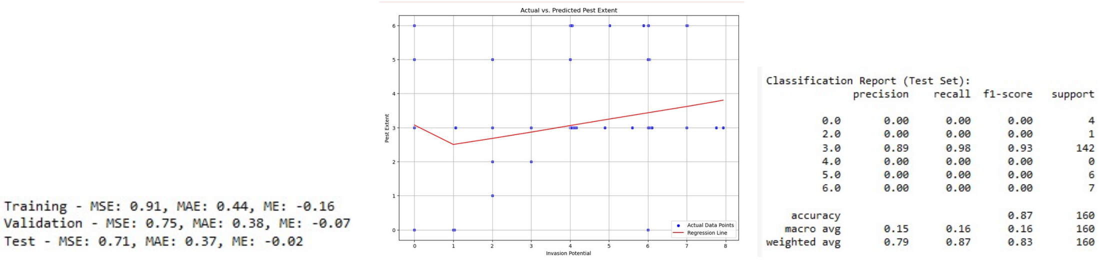
```
Training - MSE: 0.90, MAE: 0.52, ME: -0.05
Validation - MSE: 0.79, MAE: 0.47, ME: 0.05
Test - MSE: 0.74, MAE: 0.45, ME: 0.07
```

```
Classification Report (Test Set):
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00         4
         2.0       0.00      0.00      0.00         1
         3.0       0.90      1.00      0.95       142
         4.0       0.00      0.00      0.00         0
         5.0       0.00      0.00      0.00         6
         6.0       0.00      0.00      0.00         7

    accuracy                           0.89       160
   macro avg       0.15      0.17      0.16       160
weighted avg       0.80      0.89      0.84       160
```

Figure 4: Visualizations for the Tri Model



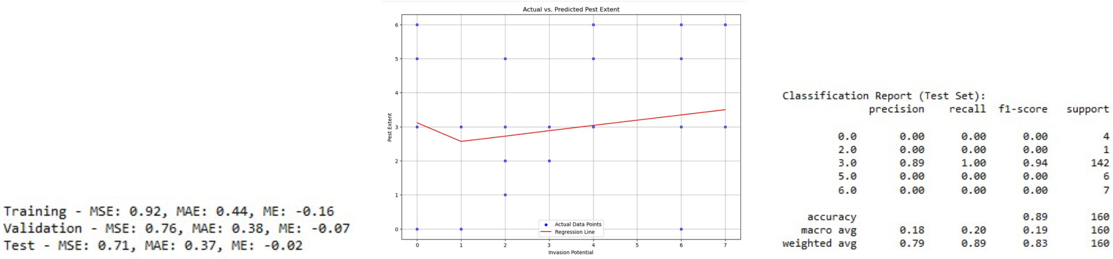
```
Training - MSE: 0.91, MAE: 0.44, ME: -0.16
Validation - MSE: 0.75, MAE: 0.38, ME: -0.07
Test - MSE: 0.71, MAE: 0.37, ME: -0.02
```

```
Classification Report (Test Set):
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00         4
         2.0       0.00      0.00      0.00         1
         3.0       0.89      0.98      0.93       142
         4.0       0.00      0.00      0.00         0
         5.0       0.00      0.00      0.00         6
         6.0       0.00      0.00      0.00         7

    accuracy                           0.87       160
   macro avg       0.15      0.16      0.16       160
weighted avg       0.79      0.87      0.83       160
```

Figure 5: Visualizations for the Duo Model



```
Training - MSE: 0.92, MAE: 0.44, ME: -0.16
Validation - MSE: 0.76, MAE: 0.38, ME: -0.07
Test - MSE: 0.71, MAE: 0.37, ME: -0.02
```

```
Classification Report (Test Set):
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00         4
         2.0       0.00      0.00      0.00         1
         3.0       0.89      1.00      0.94       142
         5.0       0.00      0.00      0.00         6
         6.0       0.00      0.00      0.00         7

    accuracy                           0.89       160
   macro avg       0.18      0.20      0.19       160
weighted avg       0.79      0.89      0.83       160
```

Figure 6: Visualizations for the Single Model

7 Final Results and Applications

At the outset of this project, my anticipation was high, fueled by the prospect of developing a sophisticated mathematical model that would not only impress my advisors but also pave the way for publication. Initially, the realization that all the models tested offered similar predictive power was a source of disappointment. However, a deeper understanding of the implications revealed a significant and practical utility in the findings.

The parity in predictive accuracy across the models indicates that we can rely on the simplest data inputs—pest extent and geographical coordinates—for effective pathogen spread prediction. Geographical coordinates, which remain constant, combined with readily available pest extent data, form a robust basis for modeling. This simplicity is particularly advantageous for developing nations, where the defense of the agricultural economy is crucial yet often hampered by limited resources.

In regions where advanced sensory equipment and consistent time-series data collection (encompassing rainfall, humidity, and temperature) are financially unfeasible, the reliance on a more complex model, despite its theoretical advantages, would be impractical. Thus, the finding that a simpler model performs equally well, if not better, under these constraints is not just a consolation but a strategic advantage.

Implementing the simpler model requires only basic data that can be gathered without expensive technology. Instead, the data can be collected through periodic surveys conducted by local experts—an approach far more feasible for countries with limited technological infrastructure. This method allows for increased reporting frequency without the need for continuous, costly data collection systems.

Moreover, the adoption of this model can significantly aid global efforts to safeguard agricultural outputs against pathogen threats by making predictive tools more accessible and cost-effective. This accessibility could transform how developing nations anticipate and mitigate crop diseases, ultimately contributing to more stable food supplies and economic resilience.

These insights not only redefine the success of this project but also underscore the broader applications of our research. They highlight the importance of adapting scientific innovations to the practical realities of their intended settings, ensuring that advancements in agricultural epidemiology reach and benefit the global community, not just technologically advanced regions.

8 Places for Improvements and Future Works

In discussions with my research advisor, we identified several areas for potential refinement, particularly concerning the multiplier functions used in our models. One concern is that our current functions may not accurately reflect the complex relationships between meteorological data and their influence on pathogen spread. While temperature and precipitation are known predictors for the spread of *Magnaporthe oryzae*, the precision and granularity of our data might not be adequate. Specifically, considering rice's growth as a seasonal crop, incorporating seasonal data might provide a more accurate reflection of environmental impacts than annual data aggregates.

Furthermore, the study by Tabonglek et al. [2] emphasizes the role of dew period rather than just precipitation in the spread of *Magnaporthe oryzae*. This suggests that substituting precipitation for dew period in our models could result in missing critical aspects of moisture's effect on pathogen propagation. This realization points to a possible area for model adjustment that could enhance the accuracy of our predictions.

Looking ahead, there are several avenues for extending this research:

- **Exploring Underlying Dynamics:** I plan to investigate whether the dynamics of spore distribution, as observed in Korea by Tabonglek et al., can be generalized to other regions. This could involve adapting the model to accommodate local variations in climate and agricultural practices, thereby enhancing its applicability and usefulness in diverse settings.

- **Utilizing More Precise Data:** Employing more detailed meteorological and pest data could reveal patterns that are not discernible at the generalized yearly scale. This finer resolution of data could potentially uncover new insights into pathogen behavior and spread mechanisms.

- **Integrating Findings with Ongoing Research:** Applying the insights gained from this project to augment existing research endeavors within our lab could significantly boost the predictive power of our risk indices. By refining our models with enhanced data and insights, we aim to improve the robustness and reliability of our predictions, contributing to more effective disease management strategies in agriculture.

These improvements and future research directions not only aim to refine the technical aspects of our models but also seek to make them more adaptable and effective in real-world applications, particularly in resource-constrained environments where advanced data collection may not be feasible.

9 Conclusion

This research initiative was embarked upon with the objective to develop an advanced mathematical model for predicting the spread of Magnaporthe oryzae in rice crops. Despite the initial goal to engineer a complex system capable of capturing intricate non-linear dynamics, the study culminated in an acknowledgment of the efficacy and practicality of simpler models, particularly in contexts with limited resources.

The comparative analysis of various models demonstrated that complexity does not necessarily enhance predictive accuracy. Simpler models, employing minimal but essential data inputs such as pest extent and geographical coordinates, proved to be just as effective. This finding is crucial for regions where technological and data acquisition capabilities are constrained, suggesting that effective pathogen monitoring can be achieved without sophisticated tools.

The broader implications of this research are significant for global agricultural practices. By validating the effectiveness of less complex predictive models, the study provides a foundation for more accessible pathogen defense strategies, applicable across diverse environments. Such models offer the potential to significantly enhance agricultural resilience by facilitating timely and reliable disease prediction, enabling proactive management and mitigation efforts.

Future research directions include refining these models through the integration of more precise data and exploring the effects of additional environmental factors, such as dew periods. The aim is to evolve these preliminary models into standard tools for agricultural planning and disease prevention worldwide, particularly in regions most in need of such capabilities.

Ultimately, this investigation reaffirms the notion that in the search for solutions to global challenges, simplicity can be both a viable and valuable approach. The research not only achieved its foundational goals but also illuminated paths toward broader applicability and scalability of straightforward modeling techniques in agricultural epidemiology.

10 Citations

References

[1] Gianni Fenu and Francesca Maridina Malloci, *Forecasting Plant and Crop Disease: An Explorative Study on Current Algorithms.* Big Data and Cognitive Computing, 2021, 5(2), 10.3390/bdcc5010002.

[2] Saharat Tabonglek, Usa Wannasingha Humphries, and Amir Khan, *Mathematical Model for Rice Blast Disease Caused by Spore Dispersion Affected from Climate Factors.* Symmetry, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand.

[3] Food and Agriculture Organization of the United Nations, *FAO Statistical Databases*, Available online: `http://www.fao.org/faostat/en/` (accessed on date).

[4] The World Bank, *Climate Change Knowledge Portal*, Available online: `https://climateknowledgeportal.worldbank.org/` (accessed on date).

[5] Centre for Agriculture and Bioscience International, *CABI Database*, Available online: `https://www.cabi.org/` (accessed on date).

[6] European and Mediterranean Plant Protection Organization, *EPPO Database*, Available online: `https://www.eppo.int/` (accessed on date).