Forecasting Black Sigatoka Infection Risks with Latent Neural ODEs

Yuchen Wang [*1] **Matthieu Chan Chee** [*1] **Ziyad Edher** [1] **Minh Duc Hoang** [1] **Shion Fujimori** [1]
Jesse Bettencourt [1]

Abstract

Black Sigatoka is the most widely-distributed and destructive disease affecting banana plants. Due to the heavy financial burden of managing this infectious disease, farmers in developing countries face significant banana crop losses. The spread of black Sigatoka is highly dependent on weather conditions and though scientists have produced mathematical models of infectious diseases, adapting these models to incorporate climate effects is difficult. We present MR. NODE (Multiple predictoR Neural ODE), a neural network that models the dynamics of black Sigatoka infection learnt directly from data via Neural Ordinary Differential Equations. Our method encodes external predictor factors into the latent space in addition to the variable that we infer, and it can also predict the infection risk at an arbitrary point in time. Empirically, we demonstrate on historical climate data that our method has superior generalization performance on time points up to one month in the future and unseen irregularities. We believe that our method can be a useful tool to control the spread of black Sigatoka.

1. Introduction

As one of the most consumed fruits worldwide, bananas had a global retail value between US $20 and $25 billion in 2016 (BananaLink). However, banana production faces a predominant leaf-spot infectious disease called black Sigatoka (Food and Agriculture Organization of the United Nations, 2013). Caused by the fungal plant pathogen *Mycosphaerella fijiensis*, the disease leads to premature ripening and a 30 - 50% loss in yields (Queensland Government, 2021). Between 2007 and 2009, St. Vincent and the Grenadines even faced a 90% decline in banana crop production due to black Sigatoka (Food and Agriculture Organization of the United Nations, 2013).

The current protection measures include applying expensive fungicides regularly, which account for up to 40% of the cost of banana production (approximately $520 million per year globally) (Joint Genome Institute, U.S. Department of Energy, 2013). Nevertheless, many local farmers in developing countries have no access to disease management facilities due to financial boundaries (Food and Agriculture Organization of the United Nations, 2013). Thus, predicting upcoming *M. fijiensis* infections on banana plants would allow farmers to take appropriate preventative measures and mitigate disease management costs.

Summary of Contributions In this study, we predict the spread of black Sigatoka based on varying microclimatic conditions by adopting a latent neural ODE approach. [1]

- We propose Multiple predictoR Neural ODE, a type of ODE-Net that defines a latent generative function. Our method extends the architecture of latent Neural ODEs (Chen et al., 2018) to model multivariate time series with external factors. See section (4).
- We demonstrate our method's effectiveness in learning the dynamics of generated black Sigatoka disease data based on a historical climate database.
- We trained and evaluated RNNs and LSTMs on our dataset as baseline models and demonstrate that our method outperforms them. See section (5).

2. Relevant Work

Mathematical Models In 1925, (M'Kendrick, 1925) invented the first mathematical algorithm for epidemics, consisting of a differential equation model that considers a fixed population with susceptible, infected and recovered individuals (SIR). In 2016, (Ochoa et al., 2016) suggested modeling black Sigatoka using the autoregressive integrated moving average (ARIMA) model. However, these "white box" models have few parameters and strong modelling assumptions,

[*]Equal contribution [1]Department of Computer Science, University of Toronto, Toronto, Canada. Correspondence to: Yuchen Wang <yuchenw@stanford.edu>, Matthieu Chan Chee <matthieu.chanchee@mail.utoronto.ca>.

Tackling Climate Change with Machine Learning workshop at ICML 2021

[1]Our code and datasets used are available at https://github.com/UofTrees/ProjectX2020.

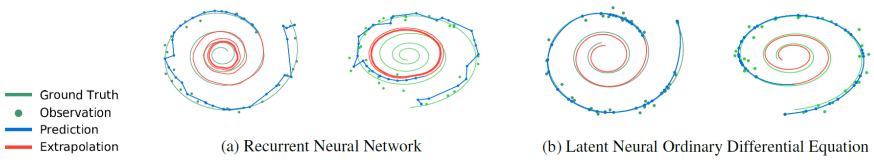

Figure 1. Comparison between RNNs and latent Neural ODE in (Chen et al., 2018). The latent Neural ODE outperforms the RNN when given irregularly-sampled data.

and are often too simplistic to capture the pattern of the disease precisely. Therefore, disease modelling would favor state-of-art neural network models, operated with fewer assumptions and more flexibility.

A Statistical Model For a number of hypothetical cohorts of *M. fijiensis* spores, (Bebber, 2019) modelled the infection of black Sigatoka as a probabilistic survival process depending on three microclimatic condition variables: relative humidity (RH, in %), canopy temperature (T, in kelvin), and moisture storage on canopy (CM, in meters). A cohort of spores germinates and infects its host during wet periods and ceases the process during dry ones. A wet period is a succession of at least three contiguous time points whereby CM > 0 meters or RH > 98%.

$$r(T) = \left(\frac{T_{max} \quad T}{T_{max} \quad T_{opt}}\right)\left(\frac{T \quad T_{min}}{T_{opt} \quad T_{min}}\right)^{\frac{T_{opt}-T_{min}}{T_{max}-T_{opt}}} \tag{1}$$

$$H(t;T) = r(T)\left(\frac{t}{-}\right) \tag{2}$$

$$F(t;T) = 1 \quad e^{-H(t;T)} \tag{3}$$

$$Y(t;T) = \quad F(t;T) \tag{4}$$

Given estimated cardinal temperatures (the minimum T_{min}, the optimum T_{opt} and the maximum T_{max}, in kelvin), (1) can determine a relative rate r for spore growth. With the scale factor and the shape parameter we then calculate a cumulative Weibull hazard function H at each time point t in a wet period (2). Via (3), H further determines F, the fraction of a cohort of spores that has infected a leaf. Finally, Y, the number of cohorts of *M. fijiensis* spores that caused infection, is computed as the product of F with the number of cohorts (4). Thus, the infection risk is defined as the sum of hourly spore cohorts that infect a leaf over a time interval.

Machine Learning Methods for Time Series A recurrent neural network (RNN) is a class of artificial neural networks for sequential modelling. Internal states in an RNN connect to each other in temporal order, enabling the network to process inputs of variable lengths. (Hochreiter & Schmidhuber, 1997) invented long short-term memory (LSTM), which reinforced the RNN architecture by solving its vanishing gradient problem. RNNs and LSTMs are known to perform well on tasks such as language modelling, whereby data is sampled at regular intervals (Lamb et al., 2016). However, as suggested by (Chen et al., 2018), applying RNNs to irregularly-sampled data can be challenging. Such data is typically discretized into bins of fixed duration, thus leading to complications if missing data exists.

Latent Neural ODEs (Chen et al., 2018) introduced Neural ODEs, a family of deep neural networks that parameterize the derivative of the hidden state using a neural network, which is fed into a black-box ODE solver. Latent Neural ODEs adopt Neural ODEs as a critical part to model continuous time series. (Chen et al., 2018) demonstrated that latent Neural ODEs could outperform RNNs in terms of extrapolation as well as modelling irregularities. Their approach is as follows:

(i) Assume that the given time series can be represented by a latent trajectory uniquely defined by an initial hidden state z_{t_0} and a time-invariant dynamics function $f = \frac{dz}{dt}$. f is parameterized by a feed-forward neural network.

(ii) An encoder RNN takes in data $x_{t_0}; \ldots; x_{t_N}$ for observed time steps $t_0; \ldots; t_N$ and produces the parameters and for a Gaussian posterior over the initial state z_{t_0} in latent space:

$$q(z_{t_0} | f x_{t_i}; t_i g_i; \quad) = N(z_{t_0} | \quad ; \quad)$$

(iii) Sample $z_{t_0} \quad q(z_{t_0} | f x_{t_i}; t_i g_i; \quad)$

(iv) The initial state z_{t_0}, dynamics function f, and the time steps for prediction and extrapolation $t_0; \ldots; t_N$, $t_{N+1}; \ldots; t_M$ are fed into a black-box ODE solver. This ODE solver applies techniques such as the Euler method or the Dormand-Prince method (Dormand & Prince, 1980) to generate values $z_{t_0}; \ldots; z_{t_N}; z_{t_{N+1}}; \ldots; z_{t_M}$.

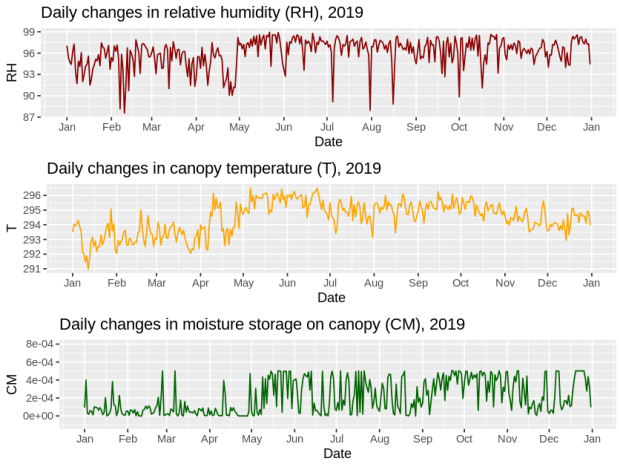

Figure 2. The progression of RH, T, and CM throughout 2019 in Costa Rica (83.812 W, 10.39 N)

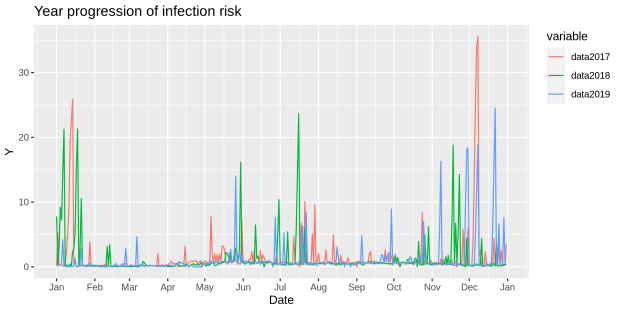

Figure 3. The progression of the generated infection variable throughout 2017, 2018 and 2019 in Costa Rica (83.812 W, 10.39 N)

(v) A decoding neural-net maps the latent space values $z_{t_0}, \ldots, z_{t_N}, z_{t_{N+1}}, \ldots, z_{t_M}$ back to data space, thus giving $\hat{x}_{t_0}, \ldots, \hat{x}_{t_N}, \hat{x}_{t_{N+1}}, \ldots, \hat{x}_{t_M}$.

3. Datasets

Since binding agreements between farmers and companies highly privatize crop disease datasets, we chose to train our model with semi-synthetic data. This kind of dataset has high label consistency and can quickly be produced in scale with minimum labour cost, meanwhile comprising a proportion of real-world data. Although (Bebber, 2019) presents a mechanistic method to estimate black Sigatoka infections (see Section (2)), we apply their algorithm for semi-synthetic data generation only, whose prediction is fed as the "ground truth" to our method.

We generated our dataset based on the Japanese Meteorological Agency 55-Year reanalysis (JRA-55) database (KOBAYASHI et al., 2015). JRA-55 comprises high spatio-temporal resolution climate data, collected from 1958 to the present day. From the vast amount of longitude and latitude coordinates available in JRA-55, we selected the longitude-latitude coordinate (83.812 W, 10.39 N) in Costa Rica which had plentiful banana productions in 2010 as indicated in the Spatial Production Allocation Model (SPAM) dataset (You et al., 2014) of global production. Then we obtained a 6-hourly multivariate time series for years 1958 - 2020 inclusive, which contains 91,556 time points and three microclimatic condition variables: relative humidity (RH), canopy temperature (T), and moisture storage on canopy (CM).

To generate the infection variable Y as laid out in Section (2), we utilized the best-fitting model parameters from the simulation experiments in (Bebber, 2019), where $T_{min} = 289{:}75$, $T_{opt} = 300{:}35$, $T_{max} = 303{:}45$, $= 32{:}6$, $= 1{:}76$ and $= 37{:}6$. Hence we obtained a 6-hourly four-dimensional time series dataset (three microclimatic conditions and the infection variable). We split the training, validation, and test sets with a ratio of 0.80 : 0.15 : 0.05.

4. Methodology: Multiple predictoR Neural ODE

We introduce the Multiple predictoR Neural ODE (MR. NODE), an architecture suitable to model time series data with external predictors. Two key innovations enable this. Firstly, we implemented a look-up function in the Neural ODE dynamics. A naive application of the latent Neural ODE system would learn latent dynamics of all the variables, which has high computation cost and departs from our goal to predict only the infection risks. Therefore, we instead feed external predictors into the latent space as given. In particular, we concatenate a continuous function $w(t)$ (the progression of external conditions through time t) to the encoded inputs, in the dynamics neural network f before solving the ODE. Secondly, instead of extrapolating the entire input space in the predictions, we train our latent Neural ODE system as a "partial" autoencoder, which only outputs the disease variable Y (see Figure 4).

5. Experiments

We conducted a series of experiments to validate the following questions:

1. Does our method extrapolate the number of cohorts that caused infection well into the future compared to baseline models? (long extrapolation)
2. Does our method interpolate the number of infections well at unseen irregularities? (irregular interpolation)

Long Extrapolation For MR. NODE, we used an LSTM encoder in all experiments. In the training phase, MR.

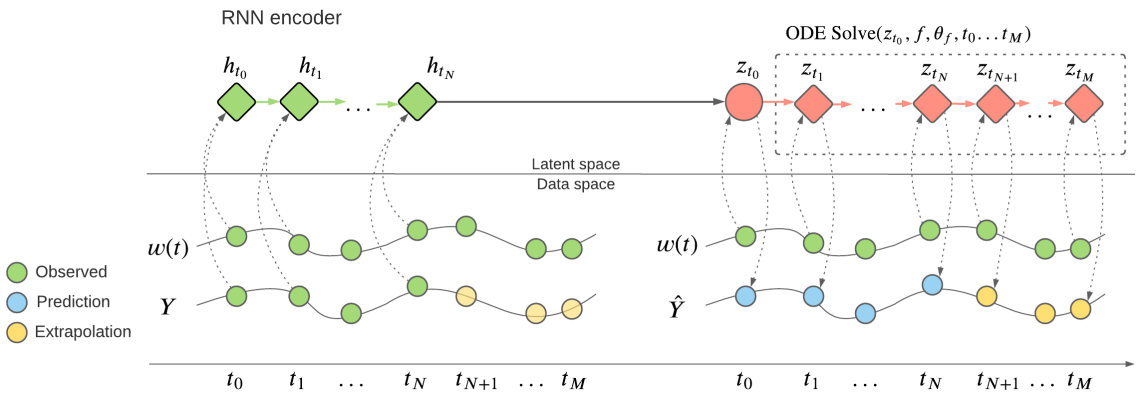

Figure 4. Computational graph of our model.

Method	Drop rates	Avg Test MSE
RNN	0	13.55
	0.3	13.51
	0.5	12.62
	0.7	13.70
LSTM	0	12.76
	0.3	12.71
	0.5	17.94
	0.7	14.12
MR. NODE	0	**12.16**
	0.3	**12.29**
	0.5	**12.40**
	0.7	**12.68**

Table 1. Average MSE results over the test data windows.

NODE encodes 128 6-hourly time points into the initial latent state and reconstructs the infection risk for those 128 time points. Loss is calculated using negative log likelihood of ground truths under Gaussian distributions with the predictions as means. In the validation phase, the model encodes 100 time points. It then reconstructs the infection risk for those 100 time points and extrapolates for 150 further time points, which is equivalent to extrapolating 37.5 days into the future. (Table 2).

To simulate irregular time conditions, we randomly dropped a proportion ($p = \{0; 0.3; 0.5; 0.7\}$) of data points for each data window (Table 2). Notably, we performed data dropping for MR. NODE only during the testing phase because the model naturally learns a continuous dynamics throughout all the time points while training. In contrast, we trained an RNN and an LSTM as baseline models which, being unable to model continuous trends in the same manner as Neural ODEs, had to be retrained for every dropping rate, with each input being concatenated with the time difference from the previous time step, as suggested in (Chen et al., 2018).

We compared our method with the baseline models by calculating the mean squared error (MSE) on extrapolated points in the testing phase. The results for our model trained on Costa Rica data are summarized in (Table 1). The plots for extrapolated data windows in the testing phase against the ground truth can be found in (Figure 7).

Irregular Interpolation To showcase MR. NODE's generalization capabilities to unseen irregularities, we tested our model for interpolation at irregular times. When encoding a data window of size 100, we randomly dropped data points with rates $p = \{0; 0.3; 0.7; 0.9\}$ (Table 2). Then, we interpolated the disease risk at time points either seen or unseen by the model for this window. We then plotted the interpolated windows against the ground truth (Figure 8).

Discussion Our method MR. NODE showed remarkable advantages over RNNs and LSTMs. Firstly, it consistently achieves the lowest extrapolation errors across 37.5 days in the future, even as the irregularity in encoded data windows increases (Table 1). The current method utilized by farmers to control black Sigatoka consists of spraying the crops with fungicides at regular intervals, thus incurring high costs. With a prediction window length of 37.5 days, farmers can distribute fungicides more efficiently (for example, decreasing the dosage when there are lower infection risks) hence reducing overall expenditure.

Secondly, even when observing only 30% and 10% of the data respectively (Figure 8c, 8d, using data from 2020 as an example), the model still predicts very similar trends as when observing the full data windows (Figure 8a). In practice, missing values frequently appear in agricultural datasets, especially those recorded in developing countries. Thus, our model's ability to handle irregularity adds great value to predicting black Sigatoka's infection risks.

	# encoded	# reconstructed	# extrapolated
Training	128	128	0
Validation	100	100	150
Extrapolation Test	100/70/50/30 (drop rate 0/0.3/0.5/0.7)	100	150
Interpolation Test	100/70/30/10 (drop rate 0/0.3/0.7/0.9)	100	0

Table 2. Training, validation and test settings of MR. NODE

	# encoded	# extrapolated
Training	100/70/50/30 (drop rate 0/0.3/0.5/0.7)	1
Validation	100/70/50/30 (drop rate 0/0.3/0.5/0.7)	1
Extrapolation Test	100/70/50/30 (drop rate 0/0.3/0.5/0.7)	150

Table 3. Training, validation and test settings of the baseline RNN and LSTM. Only extrapolation is performed.

6. Future Work

Since we fixed our extrapolation window size during experiments, we envision future researchers to extend the extrapolation windows of the model and to develop an alert system for peaks in the number of black Sigatoka infections. Reasonable thresholds for Y can give different types of alerts, thus helping farmers manage the disease. Furthermore, our model can be applied to many more agricultural time series forecasting tasks, or other fields. For instance, medical practitioners can apply our model to predict disease onsets for patients.

7. Conclusion

We propose a new architecture Multiple predictoR Neural ODEs (MR. NODE), which learnt the dynamics of infections of the black Sigatoka disease directly from data. Successfully modelling time series with multiple predictors, our method enlarged the problem space that latent Neural ODEs can solve. We conducted experiments using semi-real toy datasets and showed our method's outstanding generalization capacities in forecasting peaks of infections up to 37.5 days into the future. Importantly, if trained on real good-quality data relevant to the disease, our model can help farmers combat black Sigatoka with preventative actions, thus reducing the cost of banana crop production and protecting the most traded fruit worldwide.

Acknowledgments

This paper and the research behind it would not have been possible without Haotian Cui. He provided the fabulous idea of generating synthetic datasets using statistical methods. We thank William Navarre for offering the idea of studying infectious disease of plants and Ricardo Barros Lourenço for confirming the lack of public crop disease datasets. We thank Zhiyong Dou and Ricky T. Q. Chen for helpful discussions.

References

BananaLink. All about bananas. https://www.bananalink.org.uk/all-about-bananas/.

Bebber, D. P. Climate change effects on Black Sigatoka disease of banana. May 2019. https://doi.org/10.1098/rstb.2018.0269.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural Ordinary Differential Equations. 2018. https://arxiv.org/abs/1806.07366.

Dormand, J. R. and Prince, P. J. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19 – 26, 1980. ISSN 0377-0427. doi: https://doi.org/10.1016/0771-050X(80)90013-3. URL http://www.sciencedirect.com/science/Article/pii/0771050X80900133.

Food and Agriculture Organization of the United Nations. Battling Black Sigatoka Disease in the banana industry. July 2013. http://www.fao.org/3/a-as087e.pdf.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 1997.

Joint Genome Institute, U.S. Department of Energy. Mycosphaerella fijiensis v2.0. 2013. https://web.archive.org/web/20140228220104/http://genomeportal.jgi-psf.org/Mycfi2/Mycfi2.home.html.

KOBAYASHI, S., OTA, Y., HARADA, Y., EBITA, A., MORIYA, M., ONODA, H., ONOGI, K., KAMAHORI, H., KOBAYASHI, C., ENDO, H., MIYAOKA, K., and TAKAHASHI, K. The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, 93(1):5–48, 2015. doi: 10.2151/jmsj.2015-001.

Lamb, A., Goyal, A., Zhang, Y., Zhang, S., Courville, A., and Bengio, Y. Professor Forcing: A New Algorithm for Training Recurrent Networks. Oct 2016. https://arxiv.org/abs/1610.09038.

M'Kendrick, A. G. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130, 1925. doi: 10.1017/S0013091500034428.

Ochoa, A., Abaunza, F., and Rey, V. Forecasting black sigatoka in banana crops with stochastic models. *XXI International Meeting ACORBAT*, April 2016. https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.13140%2FRG.2.1.1109.5443.

Queensland Government. Black sigatoka. June 2021. https://www.business.qld.gov.au/industries/farms-fishing-forestry/agriculture/crop-growing/priority-pest-disease/black-sigatoka.

You, L., Wood-Sichra, U., Fritz, S., Guo, Z., See, L., and Koo., J. Spatial production allocation model (spam) 2005 v2.0. *MapSPAM*, 2014. https://www.mapspam.info/.

Appendix: Figures

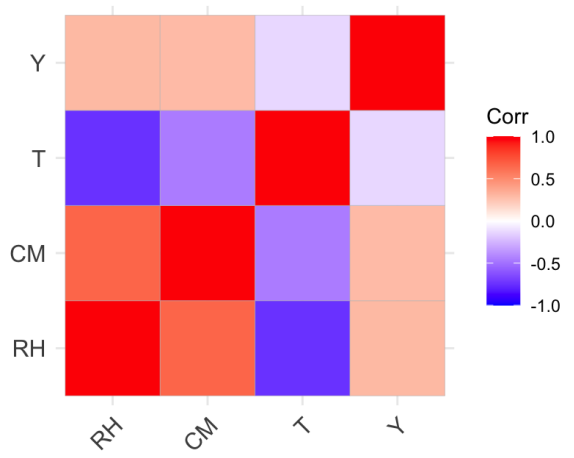

Figure 5. Correlation matrix of data variables in Costa Rica. RH has strong positive correlation with CM and strong negative correlation with T. However, no signs show explicit correlation between Y and any other variables.