 doi:10.20944/preprints202010.0255.v1

Article

Characterization of pathogen air-borne inoculum density by information theoretic analysis of spore trap time series data

Robin A. Choudhury [1] and Neil McRoberts [2,*]

[1] School of Earth, Environmental, and Marine Sciences, University of Texas, Rio Grande Valley, Edinburg TX 78541; robin.choudhury@utrgv.edu
[2] Quantitative Biology and Epidemiology Group, Plant Pathology Department, University of California, Davis, Davis CA 95616, USA ; nmcroberts@ucdavis.edu
* Correspondence: nmcroberts@ucdavis.edu

Abstract: Air sampling using vortex air samplers combined with species specific amplification of pathogen DNA, was carried out over two years in four or five locations in the Salinas Valley of California. The resulting time series data for the abundance of pathogen DNA trapped per day displayed complex dynamics with features of both deterministic (chaotic) and stochastic dynamics. Methods of nonlinear time series analysis developed for the reconstruction of low dimensional attractors provided new insights into the complexity of the pathogen abundance data, but also indicated that practicality may limit the capacity for definitively classifying the dynamics of air borne plant pathogen inoculum. Over the two years of the study five location/year combinations were classified as having stochastic linear dynamics and four were not. Calculation of entropy values for either the number of pathogen DNA copies or for a binary string indicating the pathogen abundance data were increasing or not, revealed (1) some robust differences in the dynamics between seasons that were not obvious in the time series data themselves, and also (2) that the series were almost all at their theoretical maximum entropy value when considered from the simple perspective of whether instantaneous change along the sequence is positive or not.

Keywords: time series; entropy, average mutual information, stochastic processes, deterministic dynamics

1. Introduction

> *"We now have to look at apparently random time series of data, be they from the stock market, or currency exchanges, or in ecology and ask are we seeing "random walks down Wall street" or deterministic chaos, or, often more likely, some mixture of the two."* —Sir Robert May [1].

The study of disease dynamics in plant pathology has been dominated by analysis of situations where disease increases monotonically within single growing seasons, or over several seasons [2]. Reflecting this focus, the literature on the use of monotonic growth curve models is voluminous and methodology is well developed. In contrast, the literature on how to handle long, oscillating data series for plant pathogen populations is rather thin, with only isolated case studies [3-7] employing a range of statistical approaches. To date there has been no serious effort in the botanical epidemiology literature to establish general properties of time-series data associated with disease. This is due in part, no doubt, to the fact that time series methods have been considered to be relevant mostly to multi-season data, and mult-season datasets are scarce in plant pathology. However, with the advent of molecular probes for studying the airborne inoculum of plant pathogens, it has become much easier to capture time-series data within single growing seasons [5, 6, 8].

 doi:10.20944/preprints202010.0255.v1

The increasing number of studies monitoring airborne inoculum offers a promise to epidemiologists who have an interest in developing evidence-based decision rules for managing interventions during the growing season of the crop. Given that potential application in disease management for spore traps and quantification of target nucleic acid sequences, it is important that efforts are made to develop an analytical approach which takes account of the relevant statistical properties of such data. What can experimenters expect to see when they collect such data? What types of dynamical behaviour are likely to be apparent, and how should the results be interpreted in relation to the use of the data in disease management?

The work we report here falls into the broad theme on decision-making that runs through several of the contributions in this special issue [refs from special issue]. In the case of the current work, our effort is aimed more at understanding the basic properties of the data, than in deriving decision-rules from them. Nonetheless, it is important to be aware of any informational limitations inherent in the data, so that efforts to use air sampling as a means of forecasting intervention occur with realistic expectations.

Airborne concentrations of pathogen inoculum have been recently monitored using vortex (spinning rod) air samplers combined with species-specific quantitative polymerase chain reaction (qPCR), in some cases used commercially for crop management decisions. Carisse and colleagues pioneered the approach in one of the first examples of commercial use in managing fungicide applications to control Botrytis leaf blight in onion in Quebec Province, Canada [5, 9, 10]. This work (along with characterization of effective fungicide regimes and conducive weather conditions) helped to improve monitoring and reduce disease outbreaks in onion crops in Quebec. The use of spore traps linked with qPCR assays has been developed successfully for disease monitoring in several other pathosystems, including monitoring for early season inoculum for grape powdery mildew [11], where mitigating early season inoculum can reduce yield losses in susceptible varieties. These studies show that managing disease based on the binary presence or absence of detection of pathogen inoculum can be quite successful, especially when monitoring for primary inoculum. The use of these systems for mitigating secondary inoculum is challenging.

Spinach downy mildew, caused by the obligate oomycete pathogen *Peronospora effusa*, is the most important threat to spinach production worldwide. Choudhury et al. [6] analyzed several sets of qPCR-based spore trap data collected from the Salinas Valley in California. The resulting time-series data were analyzed by fitting a series of statistical models to characterize both trend and periodicity. While the approach was successful in producing a description of the observed dynamics and linking important statistical features to plausible biological mechanisms, it offered little in the way of general understanding of inoculum dynamics. Analyses of the coefficients of prediction and the Lyapunov exponents of the resulting time series suggested that the datasets were quasi-chaotic. Further analyses of this example dataset could reveal general dynamics of airborne inoculum for plant pathogens.

Recent developments in time-series analysis [12], based on information-theoretic quantities offer some promise in being able to extract more generic properties from the available data and to provide a first example for the botanical epidemiology literature. Our objectives in this paper are to revisit the data originally studied by Choudhury et al. [6] and apply the methods suggested by [12] in order to describe the dynamics in information theoretic terms. The analyses also place our data from botanical epidemiology in the wider context of the analysis of dynamical systems allowing interdisciplinary comparison and our primary intended audience is plant pathologists and epidemiologists who might be interested in an introduction to these topics. For that reason our approach is somewhat pedagogical but does not delve deeply into the underlying technical details. We provide R code and data necessary to replicate a full set of analyses for one of the nine time series analyzed, in the repository at this URL: https://github.com/robchoudhury/spore_trap_information_theory.

2. Materials and Methods

 doi:10.20944/preprints202010.0255.v1

2.1 Data collection. Air-borne inoculum of *P. effusa* was sampled at four locations in the Salinas Valley of California in 2013 and 2014 using vortex air samplers constructed by Dr. Walt Mahaffee (USDA-ARS Corvallis, OR) and operated by Dr. Steven Klosterman (USDA-ARS Salinas, CA). Presence of the inoculum and quantification were achieved using qPCR amplification of a species-specific DNA sequence in the total DNA extract from the sampler rods. Details of the sampling procedure, qPCR primers, reaction conditions, and translation to pathogen DNA copy number for any day, t (N_t) from the qPCR cycle threshold number are described in Klosterman et al., (2014).

2.2 Data preparation. Samples were recovered from the air samplers on an irregular sampling interval of two to three days depending on the availability of technical staff. In the original 2017 study we accommodated the irregular sampling interval by fitting flexible sine function to the observations, having first removed any temporal trend by linear regression. In the current work, in order to utilize methods based on information theoretic properties, we interpolated the raw data to produce transformed time series with a regular time step of one day. The interpolation was achieved by linear interpolation to fill in missing data points. All nine data series were subjected to the same set of analyses so that we could compare their statistical properties in light of additional information characterizing each individual location-season combination. The interpolation method will have the effect of somewhat smoothing the data and the interpretation of the results of the analyses takes that into account: we avoid over-interpretation of fine-grain aspects of the analyzed series and focus on the major dynamic features that are unlikely to be strongly influenced by the interpolation.

2.3 Basic time series analysis. After interpolation of the data to a daily time step, each of the nine time series consisted of 129 observations of estimated target DNA copy number of *P. effusa* trapped for a 24 hour period. The nine time series were first inspected for evidence of an overall trend in copy number with time. Increasing trends were detected in seven of the nine series; and the series were tagged as to whether an increasing trend was apparent or not. Irrespective of whether the initial inspection suggested a trend to be present, in order to standardize the pre-treatment of the data, a simple linear regression with time (*i.e.* Julian day of observation) was fitted to the natural logarithm of the copy number. The residuals from the regression were then exponentiated to produce the detrended series that were subsequently used for time series analysis. In what follows we refer to the series analysed as N_t, indicating the (detrended) copy number on day t. When log-transformed values are analyzed they are denoted n_t.

For each series we obtained the autocorrelation function (ACF), the partial autocorrelation function (PACF), and the phase plot of the series with $\ln(N_{t+1})$ on the ordinate and $\ln(N_t)$ on the abscissa. The PACF differs from the standard autocorrelation function in that it considers only the direct effect of observations at one point in the series on observations separated by lag τ; indirect effects, operating through the interposing points in the series are removed.

2.4 Nonlinear time series analysis. To characterize the time series in terms of nonlinear dynamics we follow an approach suggested by Kantz and Schreiber [13] and Huffaker et al. [12]. The various quantities estimated for each series were obtained using functions provided in the R packages "*nonlinearTseries*" [14] or "*TseriesChaos*" [15], "*TseriesEntropy*" [16]. Additional calculations to obtain empirical entropy values used the package "*entropy*" [17], or were coded directly in R. As with many other aspects of applied data analysis for several of the steps in nonlinear time series analysis, there is no single method that is guaranteed to provide optimal results under every circumstance, and for many of the procedures there are no formal test statistics to indicate that a "significant" result has been obtained; we followed the approaches suggested in the references. Code and data for example analyses are provided in the supplementary online material and, in addition can be downloaded from https://github.com/robchoudhury/spore_trap_information_theory. The R code is provided as is and we offer no guarantee that it will work when adapted to other data sets.

2.4.1. Surrogate testing for nonlinear dependence. Since nonlinear analysis can be time-consuming, an initial step should be to test for lack of linear dependence in the observed data. An agreed approach for performing this is to perform surrogate tests [12, 14]. Different versions of the surrogate test are implemented in *nonlinearTseries* and *TseriesChaos*. The basic idea in both cases is to construct an empirical hypothesis test by resampling from the observed data, with the test statistic being a suitable property of the data that will hold under linear dependence but not under nonlinear

 doi:10.20944/preprints202010.0255.v1

dependence. One of the simplest approaches relies on the idea that a Gaussian linear process will show time reversibility. Randomized permutations are obtained using an approach in which the phases of the Fourier transform of the observed data are randomized. A two-sided hypothesis test is implemented to examine whether there is evidence that the observed value differs from the set of surrogates generated in the data resampling routine. We set the "significance level" option at 0.02 which results in the observed data being treated as one observation in a set of 100, with the two-sided test examining whether the observed data are in the $p = 0.02$ upper or lower tail of the sample. The supplied function includes a built-in diagnostic plot of the resampling test. We implemented our own diagnostic graphical representation for the test.

TseriesEntropy implements a more complex surrogate testing procedure. First, the best-fitting linear autoregressive (AR) model is selected on the basis of the Akaike Information Criterion (AIC). The residuals of the AR model are resampled (with replacement). For each resampled series, a metric entropy measure (the Bhattacharya–Matusita–Hellinger measure, S_p) [18] is calculated at different lags. Based on the relevant properties of the resampled data, the 95% confidence band for S_p can be calculated and the values for the observed series compared with the confidence band. If S_p for the observed series falls outside the band, the series can be considered to show nonlinear, as opposed to linear, dependence at the relevant lags. The entropy-based approach in *TseriesEntropy* is computationally more demanding than the expectation-based approach in *nonlinearTseries*. In initial work we examined both approaches. The results reported here are for the time-reversibility approach implemented in *nonlinearTseries*. The code supplied in the supplemental materials includes an example of the regression-based approach.

2.4.2. Characterizing nonlinear properties. Assuming that the surrogate tests indicate sufficient reason to proceed with NLTS the characterization of the dynamics in terms of their tendency to chaotic versus stochastic uncertainty is an important component of the ensuing effort. Following the pioneering work of Takens [19] one widely accepted approach to NLTS analysis proceeds by attempting to reconstruct important features of the complete (and only partially observed) phase space of the whole system by using the methods of time delay embedding to characterize the time series of a single component of the system.

In the current context, where we want to understand the dynamics of the observed series in order to be able to use similar time series data in disease management, the capacity to reconstruct the phase portrait of the whole system is of secondary importance to characterizing the dynamics of the observed series. Nonetheless, the time delay embedding approach is valuable because the features of the dynamics it reveals are useful for our primary purpose.

Three properties of the series are important in NLTS these being; (i) the average mutual information (AMI), $I(N_t; N_{t-\tau})$, of the time series data at successive lags, $\tau = 0, 1, 2, \ldots \tau_{max}$; (ii) the Theiler Window, *tw*, and (iii) the embedding dimension, *m*.

2.4.2.1. The AMI function The AMI function is calculated by binning observations and calculating the mutual information that the observation that $N_{t-\tau}$ is in the j^{th} bin provides about N_t being in the i^{th} bin. The results are averaged over all of the available data to produce the average mutual information. A graphical plot of $I(N_t; N_{t-\tau})$ against lag, $\tau = 0, 1, 2, \ldots \tau_{max}$ produces an information-theoretic analogue of the ACF plot, but one in which nonlinear, as opposed to linear, lagged dependence is visualized. The first minimum, or the first occurrence of a value below an empirical threshold, of the AMI function are taken to be an indication of the embedding time delay, *d*, of the series, since these values indicate a time lag at which observations have a low AMI and are, in a general sense, uncorrelated.

2.4.2.2. *The Theiler Window (tw)* The Theiler Window, *tw*, [20, 21] is used to define the minimum separation along the time series that two points must have in order to be included in procedures used to find the embedding dimension, *m* (see below). Theiler's review [21] gives a detailed and technical account of the issues and the various approaches suggested (up to that time).

For long time series both *TseriesChaos* and *nonlinearTseries* offer functions to generate a space-time plot [22] from which tw can be selected by choosing a value at which there is a low probability of points being close in the phase space for a given time lag separation. For short time series the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

space-time plot approach may not give usable results and other options may need to be tried; this was the case with our datasets which consist of 129 observations.

As an easily obtained first approximation, Huffaker et al. [12] suggest using the first minimum of the standard autocorrelation function (ACF). Since ACF is a linear function there are risks in using it to estimate correlation structure of nonlinear data, indeed this issue was one of the motivations for Theiler's review [23] of methods for identifying the dimensionality of nonlinear attractors. The problem, in general, appears to be that nonlinear correlation may occur at larger lag separation that would be suggested by the ACF.

In the current case, lacking a reasonable alternative, we opted for a trial-and-error approach. With both the AMI function and the ACF available we had estimates of both general and linear correlation with lag, while the original time series and the corresponding phase plots, in particular, also help to indicate suitable values of *tw*. For each series, we started with the value suggested by the first minimum of the ACF, noting also whether this lag separation was longer or shorter than the value suggested by the AMI. Where the AMI reached its first minimum at longer lag than the ACF we used a range of estimates for *tw* and examined the effect of changing *tw* on the estimated embedding dimension, *m*.

2.4.2.3. The embedding dimension, m. Options for estimating, *m*, are either the method of False Nearest Neighbors (FNN) offered in *TseriesChaos* (Huffaker et al. [12] pp 67-69) or Cao's [24] algorithm implemented in *nonlinearTseries*. Briefly, the motivation for the FNN approach comes from the idea that (in the current case) the observed time series of pathogen DNA copies represents only one dimension of a higher order dynamical system. We can think of the observed series as representing the whole higher order dynamical system projected onto a single dimension. With this perspective, points that appear close to one another may actually be widely separated in the full dimensional space of the dynamical system. The idea of the FNN computation is to select a subset of points within a given "radius" of each other, but separated by at least the value of *tw*, and to track whether they remain as neighbours as the dimensionality of the assumed attractor is incrementally increased. If the proportion of FNN is plotted against the number of dimensions, *m*, the first value of *m* at which the proportion of FNN is minimized provides an estimate of the embedding dimension.

In the approach suggested by Cao [24] the embedding dimension is identified by calculating a pair of functions, referred to as $E1(m)$ and $E2(m)$, of putative values for the embedding dimension. Note that Cao's original notation used d in place of *m*. Cao's method starts by calculating an overall Euclidean distance measure between pairs of points on time delay vectors for successively larger assumed values of *m*. Function $E1(m)$ calculates the ratio of the distance measure at successive pairs of values, $(m+1,m)$. Cao's insight was that this ratio stabilizes close to 1 if the data are generated by an attractor. The second function, $E2(m)$, focuses on the distance between only the ***nearest*** neighbours in the time delay vectors and operates on the distance measure based only on them. As with $E1(m)$, the function returns the ratio between successive pairs $(m+1,m)$. If the data are generated by a deterministic attractor $E2(m)$ has the property that at some value m^*, $E2(m^*)! = 1$, whereas if the data are generated by a process dominated by stochastic noise $E2(m) \cong 1, \forall m$. Thus, in addition to providing an estimate of the relevant embedding dimension, Cao's method offers the advantage over the FNN approach, of providing an indication of whether the data-generating process is characterized by deterministic or stochastic uncertainty.

2.5 Additional entropy measures. In addition to the characterization of the dynamics provided by the time-delay-embedding approach, we calculated two empirical entropy values to help in understanding the uncertainty of the data on airborne pathogen DNA. The first approach worked directly on the DNA copy number time series (following detrending if necessary, see above). The *entropy* function from the R package *entropy* was used to calculate empirical estimates of the entropy in the data at each time point by iteratively adding the datum for each time point to the data used for calculation and recalculating the entropy. Calculation using this approach starts by constructing a binning structure for the data and then estimates the entropy based on the frequencies of observation in each bin. We started the iterative process at the 10th time point, so that the first estimate of entropy was based on the first 10 observations of each series. The calculation then proceeded as outlined

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

above, with the second estimate being based on the first 11 data points and so on. The maximum likelihood option for the entropy function was used throughout.

As a second approach to characterize uncertainty in the time series data in relation to decision making, we first transformed each series into a binary string of length (t_{max}-1). First differences between successive pairs of values were calculated and if the resulting difference was greater than 0 (indicating $N_{t+1}>N_t$) then a 1 was entered for the value of the string; $N_{t+1}\leq N_t$ resulted in a 0. The calculation then proceeded along similar lines to those outlined for the entropy of the copy number, iteratively increasing the size of the dataset by one time point and calculating a new entropy value. In the current case at each time point we calculated the proportion of the data that were 1s and then used Shannon's equation for expected information to give an entropy value in bits for the string at each time point (including all data up to that time). The calculation was coded directly in R. We initiated the calculation with the first two observations and then iterated the calculation one time point at a time.

2.6. Linear auto-regressive models. In discussing the analysis of time series data for biological populations, Royama [25] noted that for auto-regressive (AR) models where instantaneous growth rate is modeled as a function of lagged population sizes, there is a qualitative difference in the types of behaviour that a second order lag model can display compared with a first order model. Further, given the capacity for second order models to generate quite complex oscillatory patterns, even when completely deterministic, Royama [25] suggested they could be expected to approximate the behaviour of simple nonlinear models. Since the main aim of our investigation is to look at the utility of non-linear methods, the linear AR models included here were fitted for the purpose of illustrating the extent to which a linear model can account for the observed behaviour of the data collected from air samplers.

We followed the conceptual approach that draws on the work of Royama [25] and Turchin [26] in fitting the AR linear models. The process starts with the log-transformed (log_e) time series, denoted n_t. The instantaneous log growth rate R_t, is defined as n_{t+1}-n_t and the estimated linear AR model is, then

$$R_t = a_0 + a_1 n_t + a_2 n_{t-\tau} + \varepsilon,\ \tau = 1, 2, \ldots \tau_{max} \quad (1)$$

in which a_0, a_1 *etc* are parameters to be estimated, ε is an error term, and τ is an index indicating lag dependence. Selection of the order of lag dependence (*i.e.* the value τ_{max}) to use in fitting the AR models in each case was guided by the estimates of ACF and AMI functions (see section 2.4.2, above). Parameter estimation was achieved by the standard least-squares approach implemented in the lm() function in the R base statistics package. For the selected model in each case we noted the percent variance accounted for by the model in form of the standard adjusted-R^2, and a coefficient of prediction similar to the one proposed by Turchin [26]. The coefficient was obtained as follows. We fitted a model consisting of only the mean value of the dependent variable and captured the residual sum of squares, (RSS_{mn}). Next we calculated 1-(RSS_{mod}/RSS_{mn}), in which RSSmod is the residual sum of squares from the selected model. When $RSS_{mod}>RSS_{mn}$ the coefficient has a negative value and indicates that the model is fitting noise. Values approaching 1 occur when the observed series has a pattern of oscillations that can be captured reasonably well in simple auto-regressive models. Finally, values in the region of 0 indicate that the series is dominated by noise and, possibly, too short and complex to be characterized well.

3. Results

3.1. Time series properties and nonlinearity. Time series graphs for the nine series of spore trap DNA copy number data are shown in Figure 1. Two of the nine series did not require detrending prior to analysis, these being King City South, 2014 and Gonzales, 2014. The results of testing for evidence of nonlinear dependence using Cao's method are shown in Table 1 along with other summary parameters of interest for the nine series. For four series – Salinas 2013, Soledad 2013, King City North

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

2013, and Gonzales 2014 – the surrogate (bootstrap) test led to rejection of the null hypothesis that the data were compatible with a stochastic linear (*i.e* time-reversible) process. The output from the bootstrap analysis for each series is shown in Figure 2.

Table 1. Summary statistics for the 9 time series of pathogen DNA copy number.

	Sal13	**Sal14**	**Gon13**	**Gon14**	**Sol13**	**Sol14**	**KcN13**	**KcS13**	**KcS14**[2]
ACF[1]	3	6	9	7	5	8	5	7	5
PACF	4	8	3	10	6	5	4	5	5
AMI	5	11	6	10	4	10	6	7	6
m	6	6	6	6	9	6	6	7	7
λ_1	0.05	0.16	0.09	0.06	0.13	0.18	0.04	0.04	0.05
Linear?	N	Y	Y	N	N	Y	N	Y	Y
Entropy, nats (copy no.)	1.00	0.98	1.58	0.88	1.70	0.88	0.63	2.04	1.14
Entropy, bits (binary)	0.99	0.99	1.00	1.00	0.98	0.98	1.00	1.00	1.00
%VAF	11.4	6.6	3.6	4.2	7.1	17.0	26.4	20.1	7.5
pred Coeff	0.14	0.14	0.10	0.06	0.08	0.18	0.30	0.22	0.09

[1]ACF, lag at which series autocorrelation function has first minimum; PACF, lag at which the partial autocorrelation function has first minimum; AMI, lag at which the series average mutual information function has its first minimum; m, estimated embedding dimension; λ_1, the maximum Lyapunov exponent; Linear?, outcome of surrogate test for compatibility of series with stochastic linearity; Entropy copy no., estimated entropy (nats) of the copy number time series; Entropy binary, entropy (bits) of the binary series indicating if the copy number increased between successive pairs of observations; %vaf, percent variance accounted for in the best auto-regressive linear model for the series of instantaneous rates of change in the log copy number data; pred Coeff, prediction coefficient for the auto-regressive linear model (see text for details)

[2]Location/year combination: Sal, Salinas; Gon, Gonzales; Sol, Soledad; KcN, King City, North; KcS, King City, South; 13, 2013, 14, 2014

The results of using Cao's [24] method to test for deterministic versus stochastic dynamics indicated that all 9 series had a stochastic nature; the value of function E2(m) stayed close to the value 1 for all values of m tested. Graphical output from the R function is given in the supplemental material In Appendix A in Figure A1. Note that the R function uses the symbol, d, in the place of, m.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

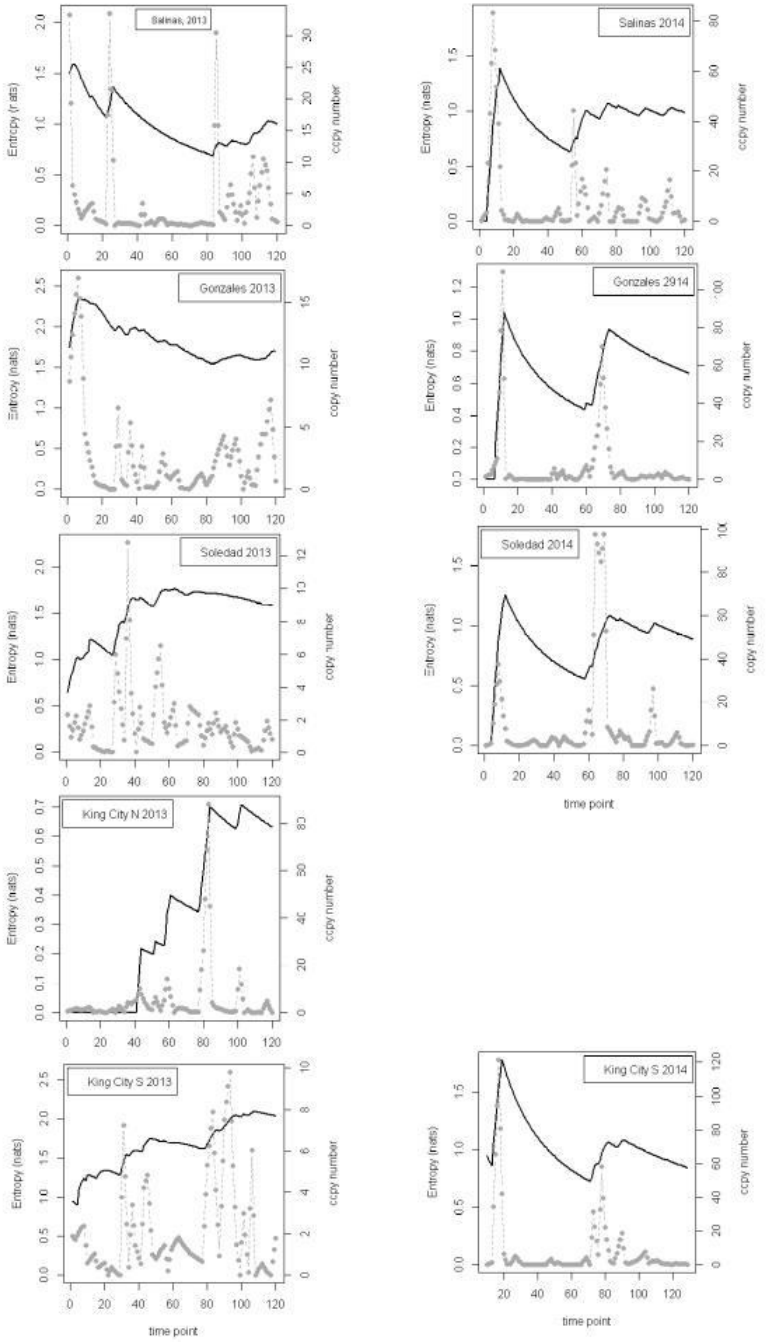

Figure 1. Detrended daily pathogen DNA copy number trapped (right axis scale) and cumulative entropy (nats, left axis scale) in the copy number series for 9 location/year combinations in which vortex air samplers were used to sample for the presence of DNA from the downy mildew pathogen of spinach, *P. effusa* in the Salinas Valley of California. Left column, 2013, right column, 2014. The King City, North location was sampled only in 2013.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

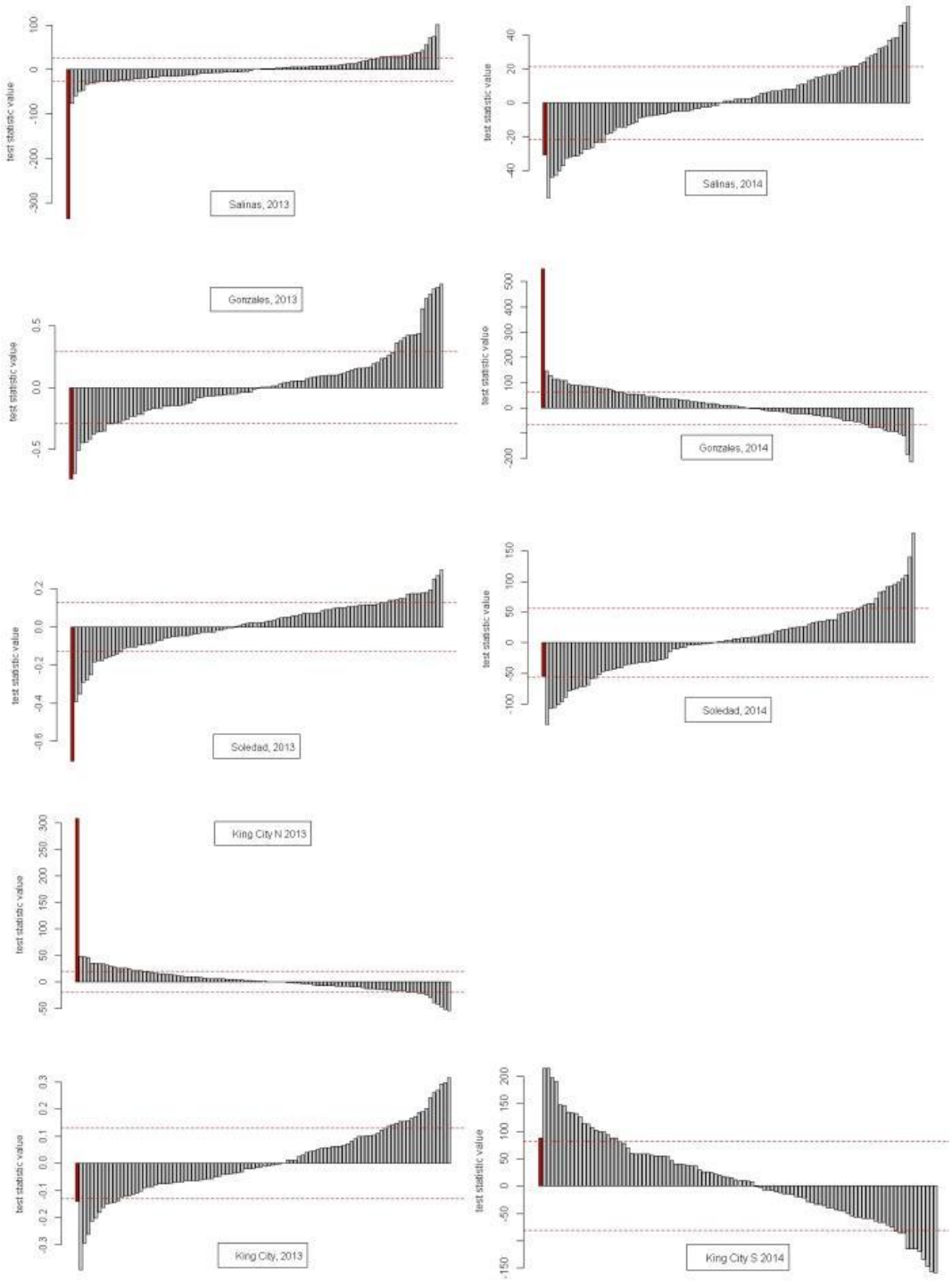

Figure 2. Results from surrogate tests (i.e. bootstrap data resampling) to assess the compatibility of spore trap data giving daily pathogen DNA copy numbers with time reversibility. The initial bar (in red) in each graph is the test statistic calculated for the original data. The remaining bars are the values calculated for bootstrap resamples of the data constructed in such a way as to break any temporal autocorrelation in the original data. The dashed horizontal lines show the standard deviation of the surrogates. Four of the nine series fail the two-sided hypothesis test for compatibility with time reversibility (*i.e.* stochastic linearity). Further details are given in the main text.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

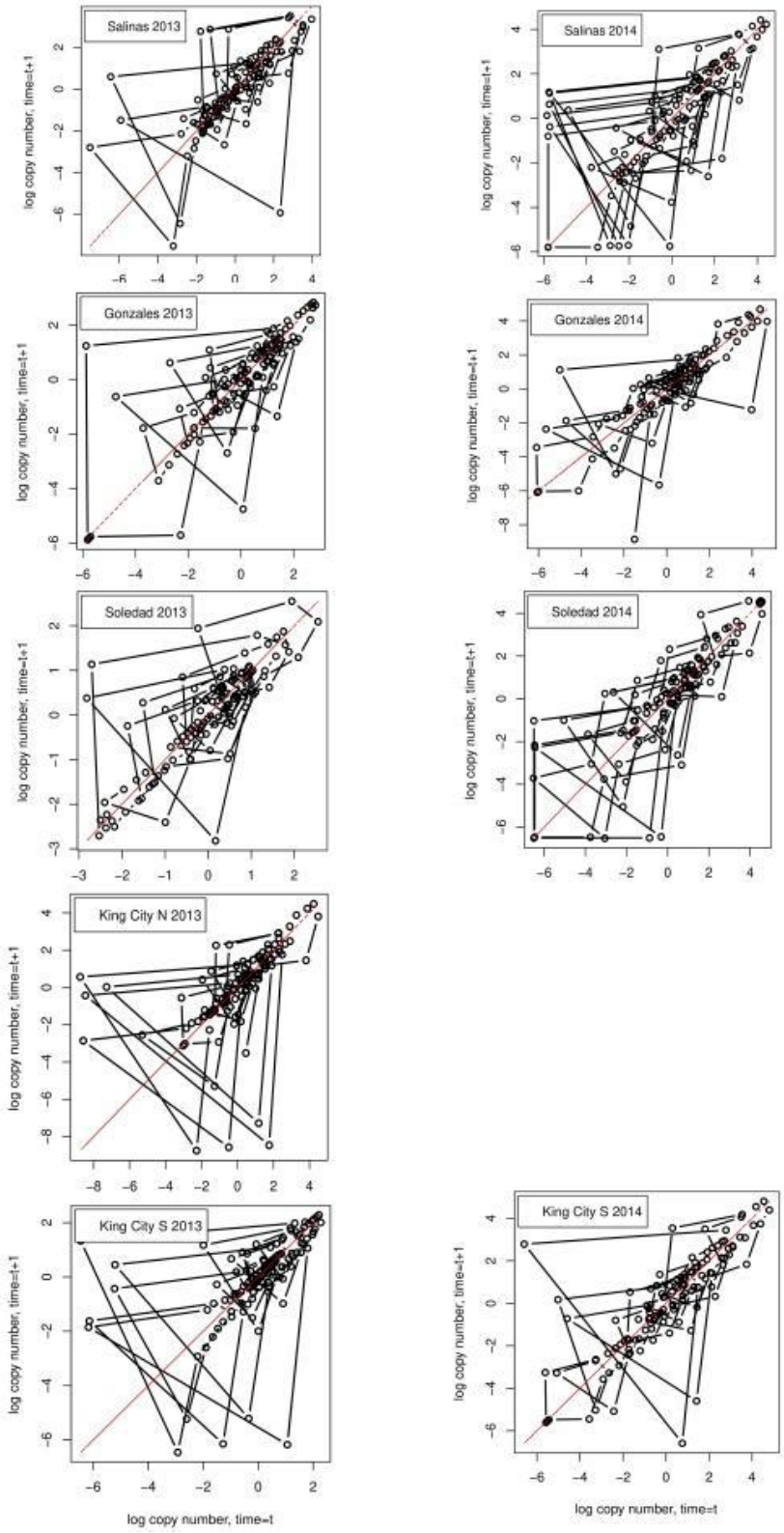

Figure 3. Phase space plots for the 9 series of detrended daily pathogen DNA copy numbers detected on vortex air samplers. The data are $\log_e$ values of the detrended data. Series orbiting a fixed attractor or a limit cycle show clockwise orbits. The obvious tendency for the phase portraits to lie along the diagonal for which $n_t = n_{t+1}$ is partly an artifact of detrending and partly a result of the fact that the series all contain sequences of observations that are very close to the mean value of the detrended series.

 doi:10.20944/preprints202010.0255.v1

The phase plots (Figure 3) for the detrended series show a strong tendency for the points to lie along the diagonal on which $n_t = n_{t+1}$, with short orbits away from this line, typically lasting no more than three to four time steps. These features are indicative of stochastic variation around a fixed value with a mixture of immediate and time-delayed feedback Turchin [26].

For all 9 series the value of the dominant Lyapunov exponent (λ_1) was greater than 0, indicating chaotic divergence would occur in independent realizations generated by the same data generating process in every case. Although positive, the values of λ_1 were small, ranging from 0.04 to 0.17 (Table 1). A correlation matrix plot for the numerical data in Table 1 is given as Figure A2 in Appendix A. Across the nine series the value of the Lyapunov coefficient was negatively correlated with the percent variance accounted for in fitting linear regression models to the series of instantaneous log growth rates and the coefficient of prediction for the linear fits; series with a higher Lyapunov exponent gave rise to poorer linear auto-regressive models.

The best fitting auto-regressive models for the time series of instantaneous rates generally captured a low proportion of the variance in the series (Table 1). Additional information about each model is given in the supplementary material. In general, the fitted values from the model showed less variability than the observed data, although in some cases the qualitative fit to the series, in tracking the direction of the oscillations was reasonably good (see Figure A3 and associated text in the supplemental material). The main result from these analyses was that while the data did not exhibit oscillations that could be easily attributed to a low-dimensional nonlinear attractor, nor were they easily described by auto-regressive linear models.

The AMI and ACF functions were correlated, but there was no consistent tendency for the AMI to reach its first minimum at higher lag than ACF. The AMI function minimized at higher lag than the ACF in five of nine cases, the functions minimized at the same lag in one case, and in the remaining four cases the AMI minimized at lower lag than the ACF.

In general, the estimated embedding dimension, m, was similar to the value suggested by the first minimum of the AMI and ACF functions; across the nine series m was negatively correlated with both AMI and ACF. The relatively large estimated values for m are indicative of complex dynamics in the observed data, but we note, again, that the data series are relatively short which may affect the accuracy of the estimated parameter.

Summarizing the results for the diagnosis of time series properties, a mixture of findings resulted. In some cases there were indications of deterministic chaos – *i.e.* positive estimated values for the Lyapunov coefficient, failure of time reversibility test in surrogates in some cases – while others were indicative of stochastic noise – *i.e.* in five out of nine cases the surrogate test failed to reject the hypothesis of time reversibility, and the first minimum value of the ACF and AMI functions were generally similar, indicating that the more general information-theoretic test of association based on average mutual information, did not routinely detect dependence in the series beyond the linear association measured by the ACF.

3.2 Entropy measures of time series uncertainty. We calculated entropy values along the time series for each location/year combination in two different ways. For the detrended copy number data the entropy was calculated (in nats) using an automated binning procedure. The resulting series of entropy values are shown together with the data in Figure 1.

In the second year of observations (2014) the detection of pathogen DNA on the traps was sporadic. All four series showed an early peak in copy numbers around day 10 and then a long period of low-to-no detection until around day 80, when all locations experienced another peak in detection. Apart from these two shared features, the time series of trap counts were superficially dissimilar across the

 doi:10.20944/preprints202010.0255.v1

four locations sampled in 2014, but the series of cumulative entropy values showed a similar pattern in all four cases, with an initial peak corresponding to the trap data at day 10 followed by a long reduction as successive, similar trap results resulted in a reduction in heterogeneity in the data. The peak in trap counts caused a further peak in entropy around day 80, followed by a second period of decline. In general the cumulative entropies in 2014 did not exceed 1.5nats, except in the case of King City, South for which the initial peak was 1.78 nats. The final values for the entropy of the four series in 2014 are given in Table 1 and range from 0.88 to 1.14 nats.

In contrast to the more or less consistent pattern revealed by the 2014 data, the cumulative entropy values for the 2013 data sets were more variable. The final values for the five series tended to be higher than those in 2014 ranging from 1.00 to 2.04 nats, with the exception of the King City North location which had a final entropy value of 0.63 nats. In Salinas and Gonzales the entropy value peaked early at over 2 nats and declined somewhat over the course of the season, although still finished at or above 1.00 nats. In contrast to this early peak and decline pattern, at the remaining three sites in 2013 uncertainty increased through much of the season, in association with repeated oscillations in the trap copy number data.

In addition to characterizing uncertainty in the daily trap data directly we also assessed the uncertainty in the simpler issue of whether the observed series increased or not between each successive pair of days. Figure 4 shows the time series for the entropy of the cumulative binary series together with the corresponding series of R_t, the instantaneous change in the log copy numbers between pairs of observations. The analysis showed that in all nine series the entropy remained close to its theoretical maximum value (*i.e.* 1 bit) over much of the season, following an initial transient period lasting approximately 30 days. In three of the series (King City, North 2013, Soledad, 2013 and Salinas, 2014)) the entropy did settle close to its maximum until later in the season, but even in these cases the final entropy value was close to the theoretical maximum of 1 bit. Note that in Figure 4 the entropy values are shown on a log scale to allow detail of the changes over time to be visible.

4. Discussion

The quotation from the late Sir Robert May's introduction to the Landmark edition [1] of his monograph Stability and Complexity in Model Ecosystems was chosen deliberately, and for more than one reason. First, May's point that the dynamics of real systems are likely to be a mixture of stochastic and deterministic processes applies directly to our observations on the time series of spore trap DNA copy numbers for *P. effusa* in the Salinas Valley. Secondly, May was an advocate of the idea that models can and should be used in biology in a strategic way to try to understand broad types of behaviour, without necessarily considering immediate questions of application or numerical accuracy in any specific case. While our analyses are predominantly statistical in nature, they are nonetheless carried out from the same strategic perspective. Our aim in this study was not so much to produce accurate predictive models of any of the series, as it was to use the tools of nonlinear time series analysis, together with some linear methods, to investigate the broad properties of pathogen DNA copy data collected from vortex air samplers.

Our analyses of the copy number data obtained over two seasons in the Salinas Valley indicate that the time series exhibit a mixture of stochastic and deterministic properties. For example, all of the series had positive Lyapunov exponents, indicating a tendency to deterministic sensitivity to initial conditions. On the other hand application of Cao's [24] approach indicated that the series were stochastic. The surrogate (bootstrap) test of time reversibility indicated that five series were compatible with the hypothesis that they were generated by a stochastic linear process while four were not. Relatively low values for the coefficient of prediction calculated from linear autoregressive models also suggests that the series were strongly influenced by stochastic noise. Taken together these results indicate that the series lie in the transition between stochastic and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

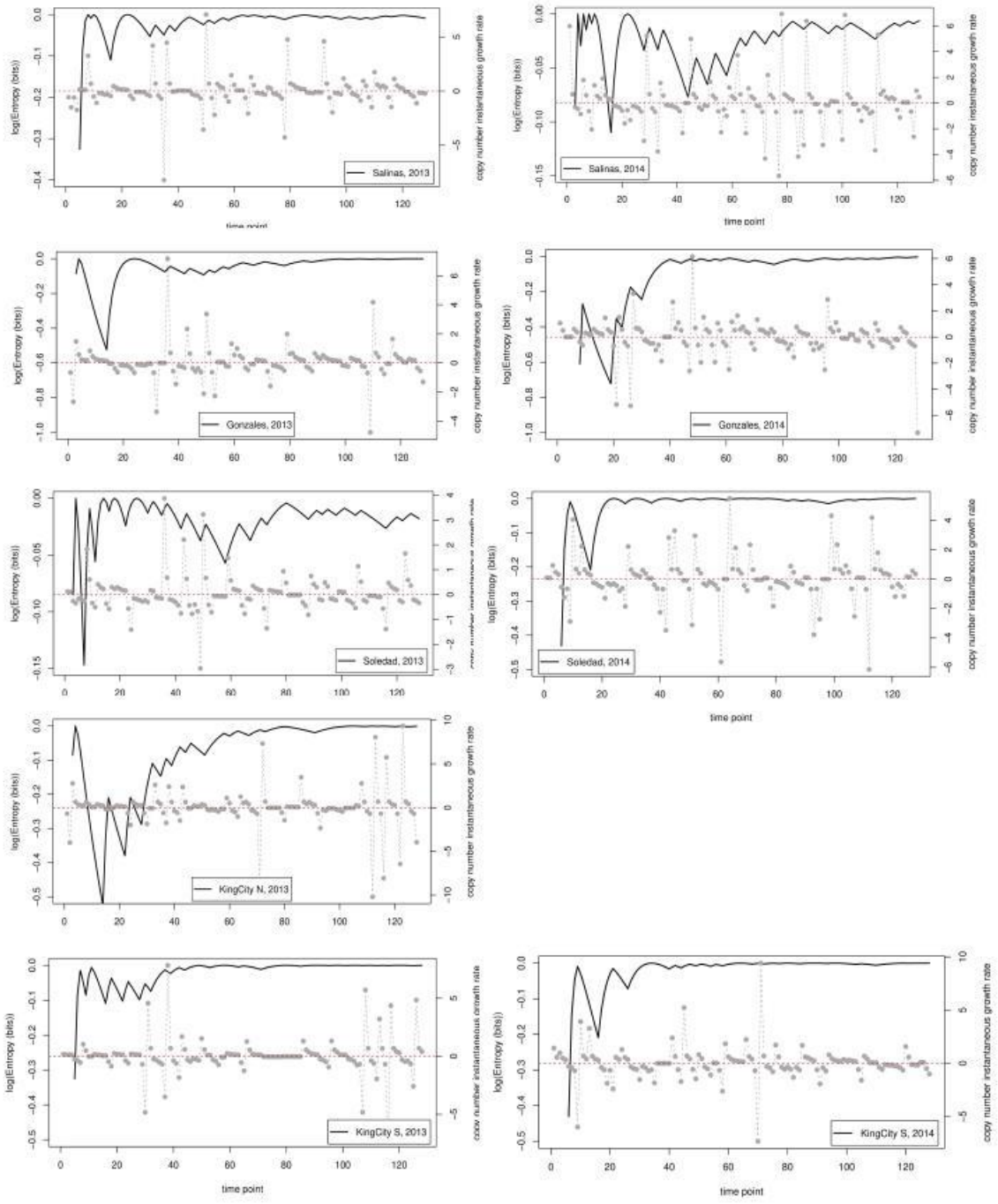

Figure 4. Graphs for 9 location years showing the series of instantaneous growth rates between successive time points (right axis scale) and the cumulative entropy (bits) of the series of binary values indicating whether the growth rate series is positive or not (left axis scale). The left axis is shown on a $\log_e$ scale to allow the variation in the entropy values to be visible. Note that on this scale the theoretical maximum value is 0.

deterministic uncertainty, in what Turchin [26] refers to as *quasi-chaotic* territory at the boundary between the two types of dynamics.

It seems reasonable, based on the dependence of oomycete pathogens such as *P. effusa* on suitable weather for spore production and release, that the copy number on air sampler traps would show

 doi:10.20944/preprints202010.0255.v1

appreciable stochasticity. Not only is the number of DNA copies detected dependent on the response of the pathogen to uncertain weather conditions, the physical processes of dispersal and transport in the air, together with the vortex sampling process mean that there are multiple sources of stochasticity between the release of spores and subsequent trapping events. At the same time, crop management practices associated with planting and harvesting salad spinach happen on cycles of between 30 and 45 days and may be a source of deterministic forcing in the data which complicates the dynamics. If the data are predominantly stochastic in nature, then traditional statistical models should be able to describe the pattern and characterize the uncertainty. Similarly, Turchin [26] argues that dynamic patterns generated by low-dimensional attractors can also successfully be described by relatively simple models. Our analyses indicate that, at least in the case of *P. effusa* in the Salinas Valley in California, the observed dynamics may fall between these two preferred making characterization of the dynamics difficult. The estimated embedding dimension values ranged from 6 to 9, indicating that the detrended time series did not have dynamics compatible with a low-dimensional attractor. On the other hand, when the time series were regressed on their own lagged history the adjusted percentage variance accounted for was small, ranging from a minimum of 3.6% (Gonzales, 2013) to a maximum of 26.4% (King City, N, 2013).

If we consider the wider implications of our results some general comments are warranted. First, if we consider the data in relation to variability in time and space there are clear implications for making robust inferences about the quantity of pathogen inoculum in the air. For example, at three of the four locations where samplers were deployed in two successive years, the dynamics were classed as linear in one year and not linear in the other. The four locations span a linear distance of approximately 80km for Salinas in the north to King City in the south. In 2014, a year with relatively little pathogen activity, peaks in trap counts and corresponding time series of entropy values showed relatively good agreement. In contrast, in 2013, when inoculum pressure was higher generally, there was much less agreement between locations, and extrapolation from one location to another would not necessarily have yielded robust conclusions about the dynamics of the pathogen at another location. The most striking example is the contrast between Salinas and King City S. In Salinas between day 20 and day 80, trap catches were relatively low and the cumulative value of the entropy showed a steady decline from approximately 1.5 nats to under 1 nat. In contrast, over the same period in King City multiple peaks in trap catches were noted and entropy in the catch data rose from approximately 1 nat to approximately 2 nats.

Our analyses suggest that there may be quite severe practical limitations to being able to characterize pathogen dynamics using the combination of vortex air sampling and DNA target amplification. There are few, if any, other comparable published datasets to compare with these data. Even with data series extending over to nearly 130 data points, the fact that the coefficient of prediction for autoregressive models was close to zero is an indication that the time series may be so noisy that extracting a useful, succinct, model of the dynamics may be difficult. This view is reinforced by the results of analyzing the simplified data series in which the quantitative variation in the copy number was replaced by a binary string indicating whether the differences between successive pairs of points were positive or not. Analysis of that binary string showed that in all 9 cases it was at, or close to its theoretical maximum value of one bit. As Grünwald [27] points out, model selection and fitting can be considered as analogous to data compression and when a string of bits is essentially random it is difficult to achieve an accurate description of the data that is more concise than simply writing the data out. The results obtained here indicate that the binary strings derived from the time series data are close to being simple sequences of independent Bernoulli trials with a probability of 0.5 determining the outcome. While these results point to restrictions in the utility of dynamical analysis for helping with practical problems in disease risk forecasting, at the same time they suggest a great deal of interesting investigative research on dynamics and their positions on the continuum form pure deterministic complexity to pure stochastic noise.

 doi:10.20944/preprints202010.0255.v1

In spite of the potential limitations that our work has indicated in relation to being able to describe dynamics in a concise, robust way, there are existing examples of the successful use of spore samplers combined with DNA quantification in disease management [10,11] and the limitations that may exist in being able to classify dynamics, do not prevent the use of the technology in providing empirical evidence of abrupt peaks in the abundance of aerial inoculum that may indicate the onset of disease.

Supplementary Materials: The following are available online at https://github.com/robchoudhury/spore_trap_information_theory, R code to reproduce results for Salinas, 2103:Sal_13.R, Data required to run the analysis for Salinas 2013: spore_fill.csv, R Project file:spore_trap_information_theory.Rproj.

Author Contributions: Conceptualization, N.M. and R.C.; methodology, N.M. and R.C..; software, N.M. and R.C.; formal analysis, N.M.; resources, N.M. and R.C.; data curation, R.C.; writing—original draft preparation, N.M.; writing—review and editing, R.C.; visualization, N.M. and R.C.; funding acquisition, N.M. and R.C..

Funding: This research received no external funding.

Acknowledgments: Work by N.M. on this paper is aligned with USDA-NIFA Hatch project CA-D-PPA-2131-H. R.C. was supported by funds from the University of Texas, Rio Grande Valley.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

Appendix A

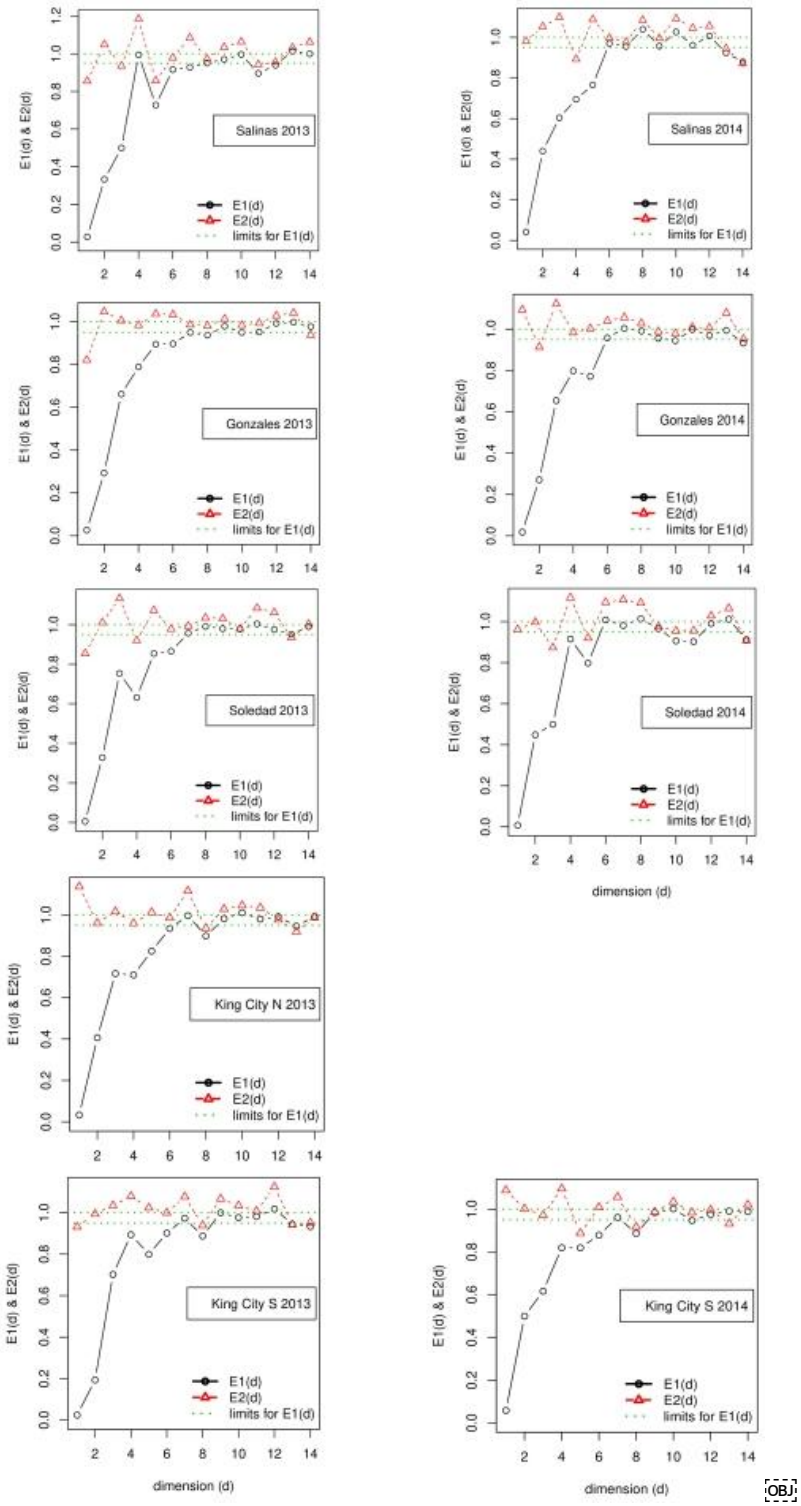

Figure A1. Output from the R function *estimateEmbeddingDim* for each series. The dimension, d, at which function E1 first falls within the critical region around 1 is taken to be the embedding dimension for the series. The behaviour of function E2 is provides an indication of the nature of the series. In this case E2 is approximately equal to 1 for all values of d for all series, indicating the the series are stochastic in nature. Note that the embedding dimension is denoted with the letter *m* in the main text of the paper.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

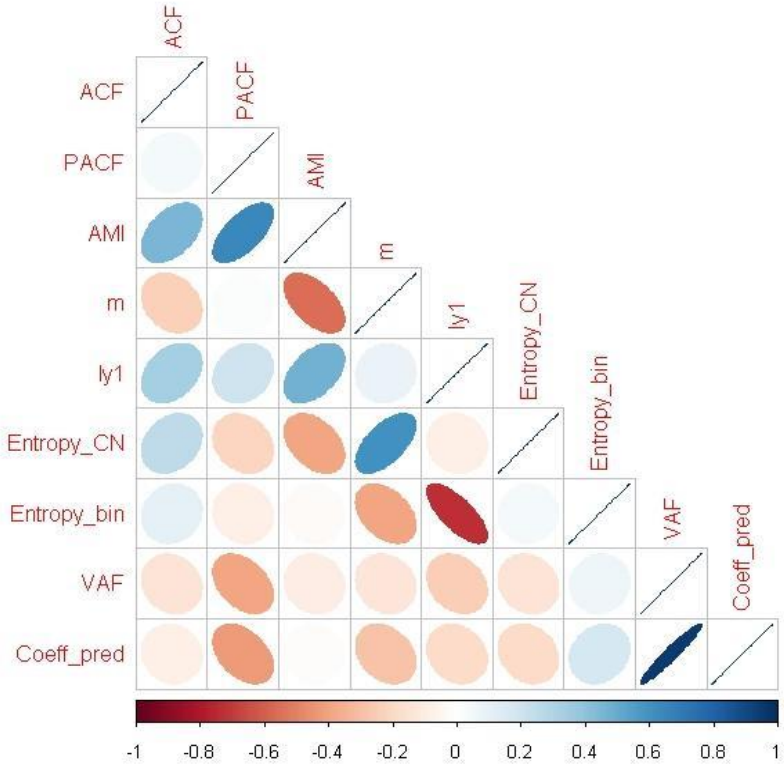

Figure A2. Correlation matrix plot indicating the numerical value and direction of correlations among the summary variables for of time series properties across the 9 example times series. Abbreviations for the series are explained in footnote 1 of Table 1 in the main text.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2020 doi:10.20944/preprints202010.0255.v1

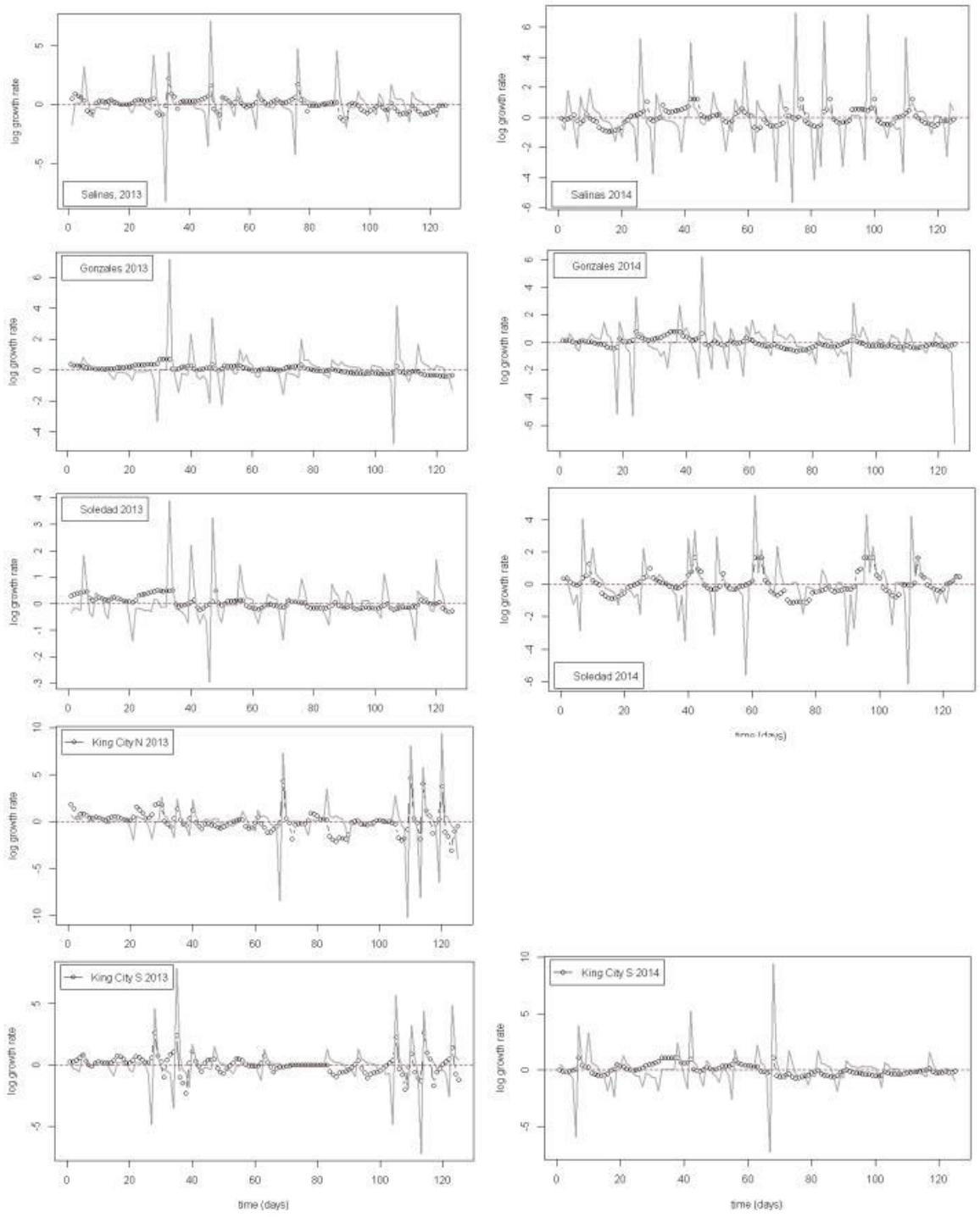

Figure A3. Observed data and fitted values for linear auto-regressive models fitted to equation 1 in the main text in the case of each of the nine series. Observed data: gray dashed line with open symbols.; fitted values black open symbols. The simple autoregressive models are capable of capturing some of the dynamic behaviour in the series, but generally lack the amplitude of the observed data and are poor at representing abrupt changes from large positive to large negative growth rates.

 doi:10.20944/preprints202010.0255.v1

Appendix B

All appendix sections must be cited in the main text. In the appendixes, Figures, Tables, etc. should be labeled starting with 'A', e.g., Figure A1, Figure A2, etc.

References

1. May, R. M., *Stability and complexity in model ecosystems*. Princeton University Press: 2019; Vol. 1.
2. Madden, L. V.; Hughes, G.; Van Den Bosch, F., The study of plant disease epidemics. **2007**.
3. Zwankhuizen, M.; Zadoks, J., *Phytophthora infestans*'s 10-year truce with Holland: a long-term analysis of potato late-blight epidemics in the Netherlands. *Plant Pathology* **2002,** 51, (4), 413-423.
4. Kriss, A.; Paul, P.; Madden, L., Relationship between yearly fluctuations in Fusarium head blight intensity and environmental variables: a window-pane analysis. *Phytopathology* **2010,** 100, (8), 784-797.
5. Carisse, O.; McRoberts, N.; Brodeur, L., Comparison of monitoring-and weather-based risk indicators of botrytis leaf blight of onion and determination of action thresholds. *Canadian Journal of Plant Pathology* **2008,** 30, (3), 442-456.
6. Choudhury, R.; Koike, S.; Fox, A.; Anchieta, A.; Subbarao, K.; Klosterman, S.; McRoberts, N., Season-long dynamics of spinach downy mildew determined by spore trapping and disease incidence. *Phytopathology* **2016,** 106, (11), 1311-1318.
7. Kasprzyk, I.; Worek, M., Airborne fungal spores in urban and rural environments in Poland. *Aerobiologia* **2006,** 22, (3), 169.
8. Klosterman, S. J.; Anchieta, A.; McRoberts, N.; Koike, S. T.; Subbarao, K. V.; Voglmayr, H.; Choi, Y.-J.; Thines, M.; Martin, F. N., Coupling spore traps and quantitative PCR assays for detection of the downy mildew pathogens of spinach (*Peronospora effusa*) and beet (*P. schachtii*). *Phytopathology* **2014,** 104, (12), 1349-1359.
9. Carisse, O.; Tremblay, D.; Lévesque, C.; Gindro, K.; Ward, P.; Houde, A., Development of a TaqMan real-time PCR assay for quantification of airborne conidia of Botrytis squamosa and management of Botrytis leaf blight of onion. *Phytopathology* **2009,** 99, (11), 1273-1280.
10. Carisse, O.; Tremblay, D.; McDonald, M.; Brodeur, L.; McRoberts, N., Management of Botrytis leaf blight of onion: the Québec experience of 20 years of continual improvement. *Plant disease* **2011,** 95, (5), 504-514.
11. Falacy, J.; Grove, G.; Mahaffee, W.; Galloway, H.; Glawe, D.; Larsen, R.; Vandemark, G., Detection of *Erysiphe necator* in air samples using the polymerase chain reaction and species-specific primers. *Phytopathology* **2007,** 97, (10), 1290-1297.
12. Huffaker, R.; Bittelli, M.; Rosa, R., *Nonlinear time series analysis with R*. Oxford University Press: 2017.
13. Kantz, H.; Schreiber, T., *Nonlinear time series analysis*. Cambridge University Press: 2004; Vol. 7.
14. Garcia, C.; Sawitzki, G., nonlinearTseries: nonlinear time series analysis. *R package version 0.2* **2020,** 3.
15. Di Narzo, A. F.; Di Narzo, F., TseriesChaos: Analysis of Nonlinear Time Series. R package version 0.1-13. In 2019.

 doi:10.20944/preprints202010.0255.v1

16. Giannerini, S., tseriesEntropy: entropy based analysis and tests for time series. *R package version 0.5-13* **2017**.
17. Hausser, J.; Strimmer, K., Entropy: estimation of entropy, mutual information and related quantities. *R package version* **2015,** 1, (1).
18. Granger, C. W.; Maasoumi, E.; Racine, J., A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis* **2004,** 25, (5), 649-669.
19. Takens, F., Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, Springer: Warwick 1981; pp 366-381.
20. Theiler, J., Spurious dimension from correlation algorithms applied to limited time-series data. *Physical review A* **1986,** 34, (3), 2427.
21. Theiler, J., Estimating fractal dimension. *JOSA A* **1990,** 7, (6), 1055-1073.
22. Provenzale, A.; Smith, L. A.; Vio, R.; Murante, G., Distinguishing between low-dimensional dynamics and randomness in measured time series. *Physica D: Nonlinear Phenomena* **1992,** 58, (1-4), 31-49.
23. Casdagli, M.; Eubank, S.; Farmer, J.; Gibson, J.; Desjardins, D.; Hunter, N.; Theiler, J., Nonlinear modeling of chaotic time series: theory and applications. *EPRI* **1990**.
24. Cao, L., Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena* **1997,** 110, (1-2), 43-50.
25. Royama, T., *Analytical population dynamics*. Springer Science & Business Media: 2012; Vol. 10.
26. Turchin, P., *Complex population dynamics: a theoretical/empirical synthesis*. Princeton University Press: 2003; Vol. 35.
27. Grünwald, P. G., *The Minimum Description Length Principle*. The MIT Press, Cambridge, MA: 2007.