# Using Resampling Methods to Evaluate the Relative Performance of Regime Detection Measures Under Varying Conditions of Data Quality and Quantity

## Introduction

Ecological systems have many unpredictable and variably interacting components. Methods for analyzing these complex systems, e.g. Dynamic Bayesian Networks, network models, and food webs are designed to handle these complexities, yet require data- and knowledge-intensive models. Although ecological data collection and data management techniques are improving [@lasorte2018opportunities], the aforementioned approaches to modeling and understanding complex system are often infeasible in ecosystem research and management [@clements\_including\_2016].

A growing concern with anthropogenic impacts on the environment has increased the demand for mathematical and statistical techniques that capture these dynamics. These often undesirable changes in the structure or functioning of ecological systems are often referred to as *regime shifts*, *regime changes*, *state change*, *abrupt change*, etc. [@andersen\_ecological\_2009] . A yet-unattained goal of ecological research and management is to reach a point where these methods can predict impending regime shifts in real-time and with high confidence. Ideally, ecological regime shift detection methods (hereafter, regime detection measures) would require little knowledge of the intrinsic drivers of the system, and the users of the method would not be required to know if and where a regime shift occurred in the data.

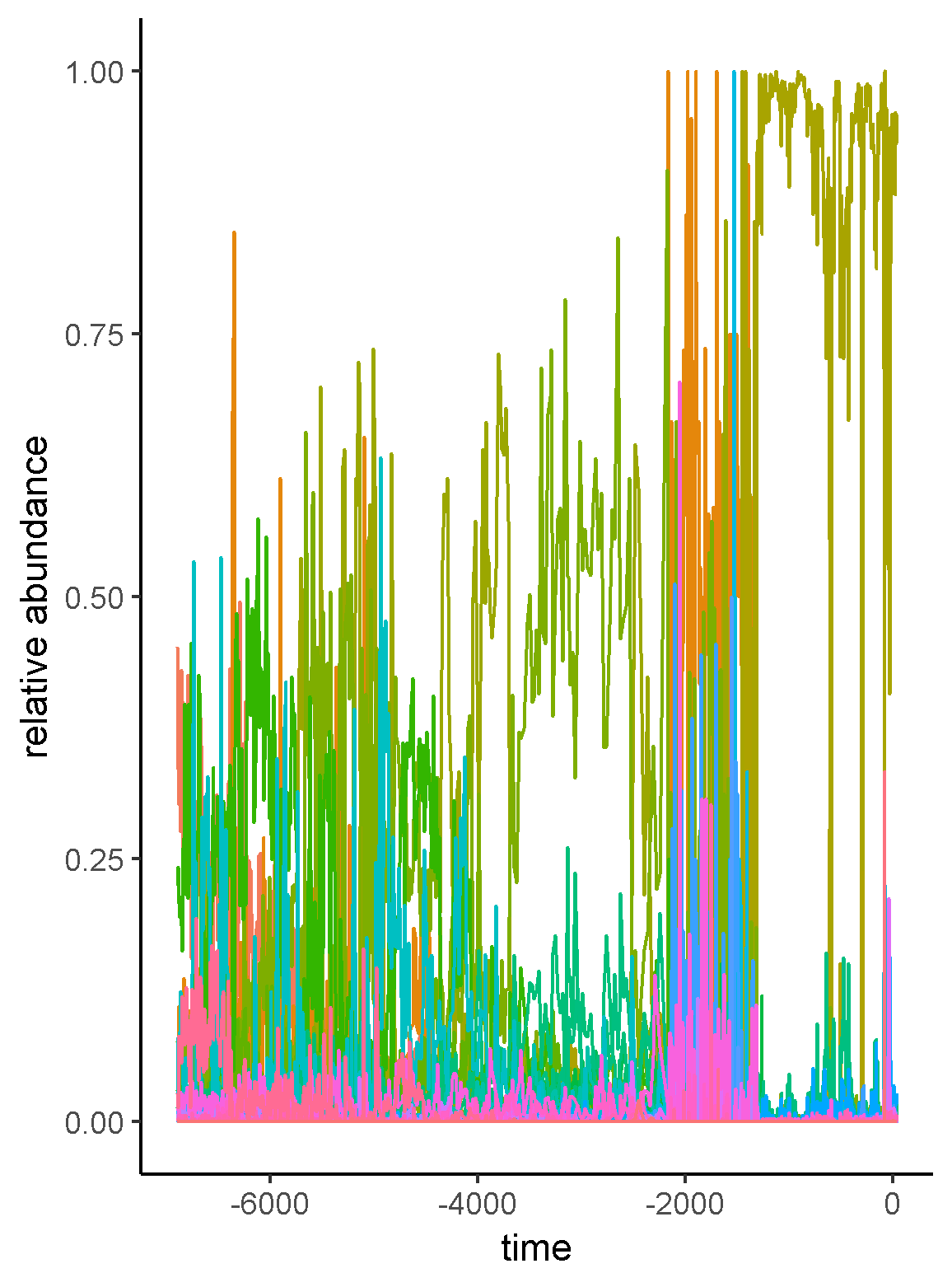
Despite the suite of regime detection measures in the environmental and ecological research literatures, they are not used in ecological management. We can describe the current state of regime detection measures as being either system specific (i.e., the method is not system agnostic) or not. Methods of the latter type are convenient in that they can be applied across various system and data types, but the results of these analyses require some degree of subjective interpretation [@clements2018indicators; \*c.f.\* @batt2013changes]. Efforts to develop and/or improve regime detection measures that do not require such subjectivity will aid the advance of regime detection measures research and application.

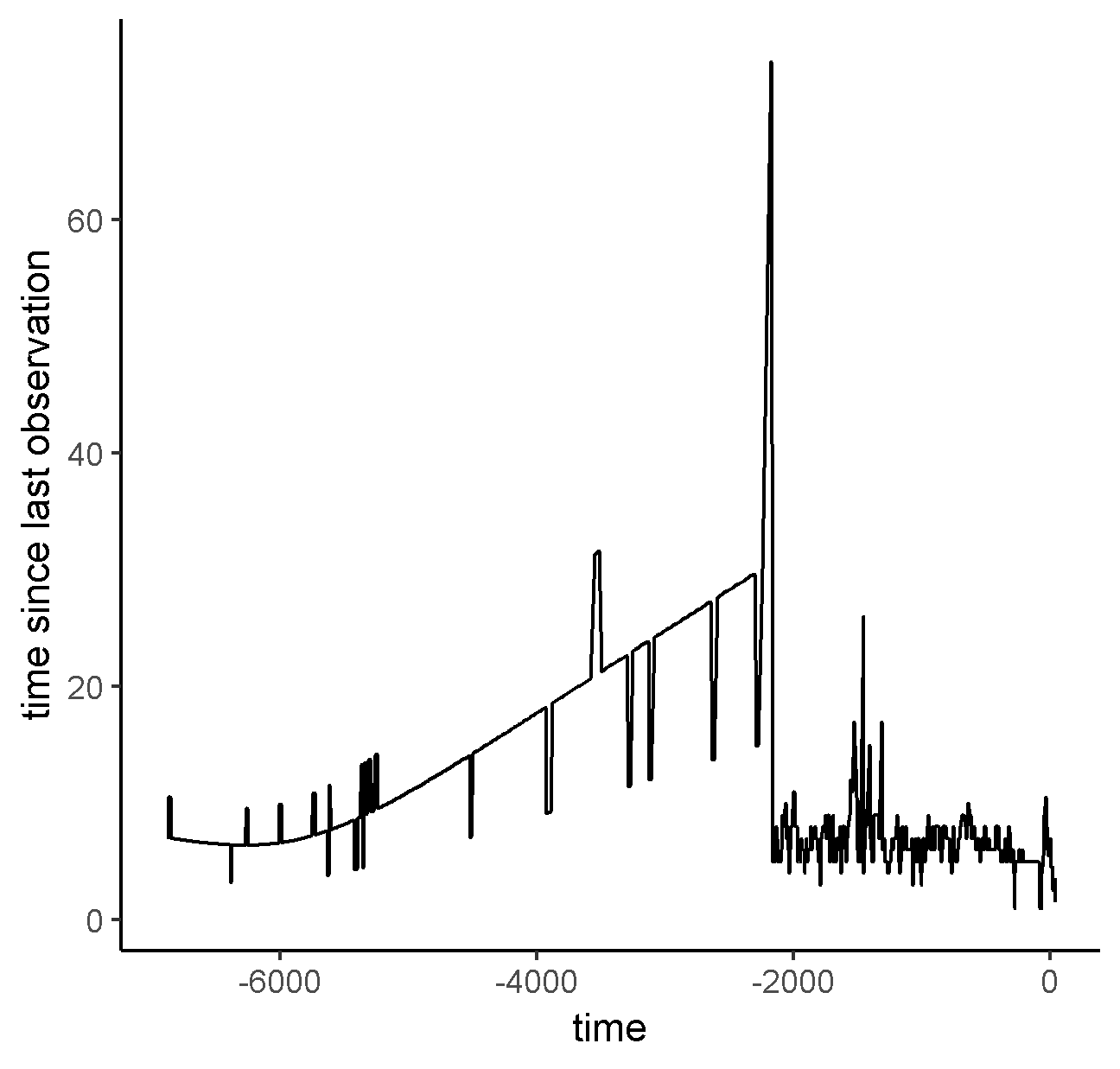
Current efforts to improve regime detection measures may be stunted by the lack of application beyond simple and/or theoretical (toy) systems data. Like most statistical and mathematical approaches, the evolution of many regime detection measures begins with application to theoretical data, followed by application to empirical data. Current applications of regime detection measures to empirical, ecological data are largely limited to data describing populations [@anderson\_community\_1999;@alheit\_synchronous\_2005; @deyoung\_regime\_2008], climatic, marine, and Paleolithic regime shifts [@spanbauer\_prolonged\_2014; @kong2017hydrological; @yang\_10\_2006], with few applications to terrestrial data [\*c.f.\* @bahlai2015shifts; @sundstrom2017detecting]. Although testing the performance and inference boundaries of theoretical and simple systems is important, they are of little use to ecosystem managers if they are not proven to be easily and reliably applicable to their system. Additionally, regime detection measures should be capable of handling empirical ecological data, which are often sparse, noisy, and irreguarly sampled..

Ecological systems data is expensive to capture, and has large process variation and observation errors. This variability reduces data quality and quantity, limiting the numerical tools for identifing trends and changes in the system [@thrush2009forecasting]. Some methods, new and old, proposed as regime detection measures are purported to handle the data limitation and quality issues inherent in ecological data, and minimize subjective decisions for choosing state variables and interpreting results. For example, variable reduction techniques, e.g. principal components analysis [@rodionov\_application\_2005; @andersen\_ecological\_2009; @reid\_global\_2016], clustering algorithms [@weijerman2005regime;@weissmann2016predicting], an index of variance [@brock\_variance\_2006], and Fisher Information [@cabezas\_towards\_2002; @fath\_exergy\_2004; @karunanithi\_detection\_2008] were introduced as methods which collapse the system into a single indicator of ecological regime shifts. Although these methods have been used on empirical ecological systems data, their robustness to empirical data quality and quantity have yet to be examined.

In this Chapter I examine the influence of observation and process errors on the inference obtained from select multivariable regime detection measures. There are three major objectives:

1. Identify the effects of data quality on regime detection measure inference.
2. Identify the effects of data quantity on regime detection measure inference.
3. Explore the relative performance of velocity (described in Chapter @ref(velocity)) to the above mentioned methods under multiple scenarios.

This Chapter provides baseline relative performance estimates of select, multivariable regime detection measures under various scenarios of data quality and quantity. The results from this Chapter inform the practical ecologist of the potential limitations to consider when applying these regime detection measures to their data, and has potential to inform the data collection process. Additionally, the software accompanying this Chapter allows the end user to implement these methods on this diatom system, a toy system, or their own data. 



The amount of time elapsed between observations.

## Data and Methodology

### Study system and data

I used paleodiatom time series from a freshwater system in North America (Foy Lake, present day Montana) that apparently underwent rapid shifts in algal community dynamics at multiple points in time. This data comes from a single soil core sample, from which the relative abundances of 109 diatom species were identified at 768 observations (time points) over years (Figure @ref(fig:origDat). Althouh the soil core was sampled at regular distances, the soil accumulation process is not necessarily linear over time, resulting in irregularly-sampled observations (i.e., time elapsed between sampling points differs varies; see Figure @ref(fig:timeElapsed)). The data were published in @spanbauer\_prolonged\_2014 and can be downloaded at the publisher’s website.

### Regime detection measures

Fewer model-free regime detection metrics exist than do model-based metrics (Chapter @ref(rdmReview)) and of these, only a few are suggested for multivariable data. Here, I compare the results for three regime detection metrics that are model-free and can handle multivariable data: velocity (Chapter @ref(velocity)), the Variance Index [@brock\_variance\_2006] and Fisher Information [@fath\_regime\_2003]. I chose the Variance Index, as this is one of the more widely applied multivariate, model-free regime detection measures, and has been shown to, in some empirical data, identify regime shifts *post hoc*. I introduced the velocity in Chapter @ref(velocity) as a new, potential regime detection metric. As this is the first time it has been used for such a purpose, including it in this approach allows us to further identify potential flaws with the method, but also to gain some baseline estimates of its relative performance. In Chapter @ref(fiGuide), I presented the Fisher Information metric as it is used in detecting ecological regime shfits, and discuss the situations under which it may or may not be a good metric.

#### Velocity () calculation

In Chapter @ref(velocity), I describe a new method, **velocity**, , as a potential dimension reduction and regime detection method. First introduced by @fath\_regime\_2003 as one of multiple steps in calculating their variant of Fisher Information, velocity calculates the cumulative sum of the square root of the sum of the squared change in all state variables over a period of time [Eq. @ref(eq:velocityEq)]. Steps for calculating this metric are described in detail in Chapters @ref(fiGuide) and @ref(velocity).

#### Variance Index (VI) calculation

The Variance Index was first introduced by @brock\_variance\_2006, and can be simple defined as the maximum eigenvalue of the covariance matrix of the system within some period, or window, of time. The Variance Index (also called Variance Indicator) was originally applied to a modelled system [@brock\_variance\_2006] and has since been applied to empirical systems data [@spanbauer\_prolonged\_2014; @sundstrom2017detecting]. Although rising variance has been shown to manifest prior to abrupt shifts in some empirical systems data [@van2005implications;@brock\_variance\_2006], the Variance Index, which is intended for multivariate data, appears most useful when the system exhibits a discontinuous (non-linear) shift [@brock\_variance\_2006].

#### Fisher Information (FI) calculation

Fisher Information () is essentially the area under the curve of the acceleration to the fourth degree () divided by the squared velocity (; also referred to as in Chapter @ref(velocity)) of the distance travelled by the system, over some period of time (), and is given in Eq. @ref(eq:fiDerivs2):

I refer the reader to Chapter @ref(fiGuide) for a complete description and to @cabezas\_towards\_2002 for a complete derivation of Fisher Information.

#### Using moving window analysis to calculate Fisher Information and Variance Index

Unlike , the Variance Index and Fisher Information are calculated using moving window analysis. That is, over the entire time series, , these metrics are calculated within multiple windows of time, . In this approach, all state variables, , are used to inform the calculations (of Variance Index and Fisher Information) over a time interval, , where is the length in [time] units of the time interval and satisfies the following condition: . If , then only a single value of the metric will be calculated for entire time series, which does not allow for any estimate of change.

When using these metrics in the context of identifying abrupt changes in ecological systems data across , it is ideal the value of meets the following conditions: . The length of a time window dictates the number of calculations one can obtain over , such that the number of potential metric calulations increases as decreases. Previous applications of moving window analyses to calculate Fisher Information found that at least eight observations (time points) should be used [citation].

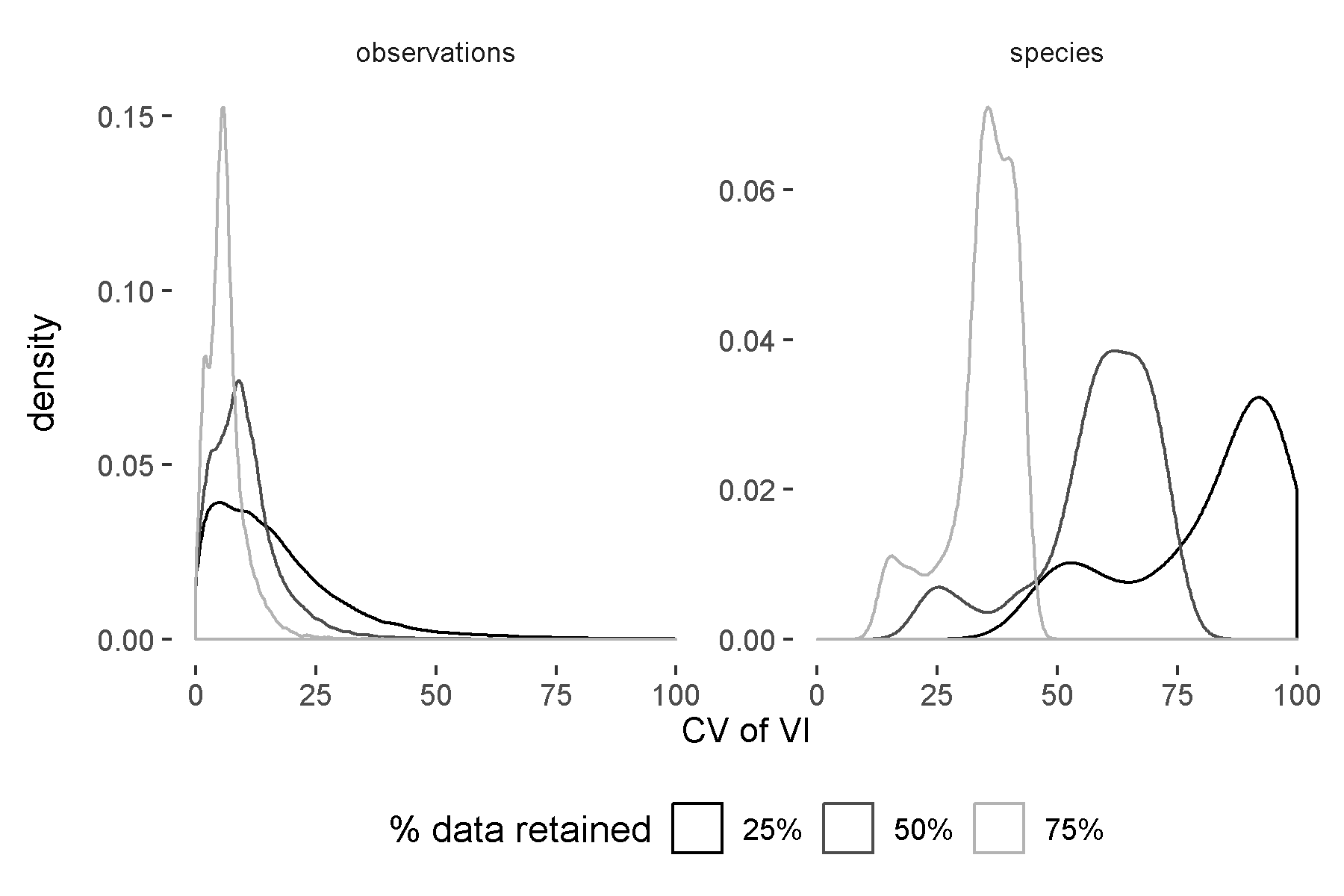
An additional parameter is required when conducting moving window analyses: the number of time points by which the window advances. In order to maximize the data, I advance the window at a rate of one time unit. However, it is important to note that because these data are not sampled annually and the because the window always advances by a single time unit, the number of observations included in each calculation will not be the same. If fewer than 5 observations are in a window, I did not calculate metrics, advancing the window forward. I assigned the calcuated values of Fisher Information and Variance Index within each moving window to the **end** (the last time unit) of the moving window. In temporal analyses, assigning the value[which value] to any other point in time (e.g., the beginning or the middle) muddles the interpretation of the metric over . Also note that this method has the potential to result in calculating a metric for all integers between and .

### Simulating data quality and quantity issues using resampling techniques

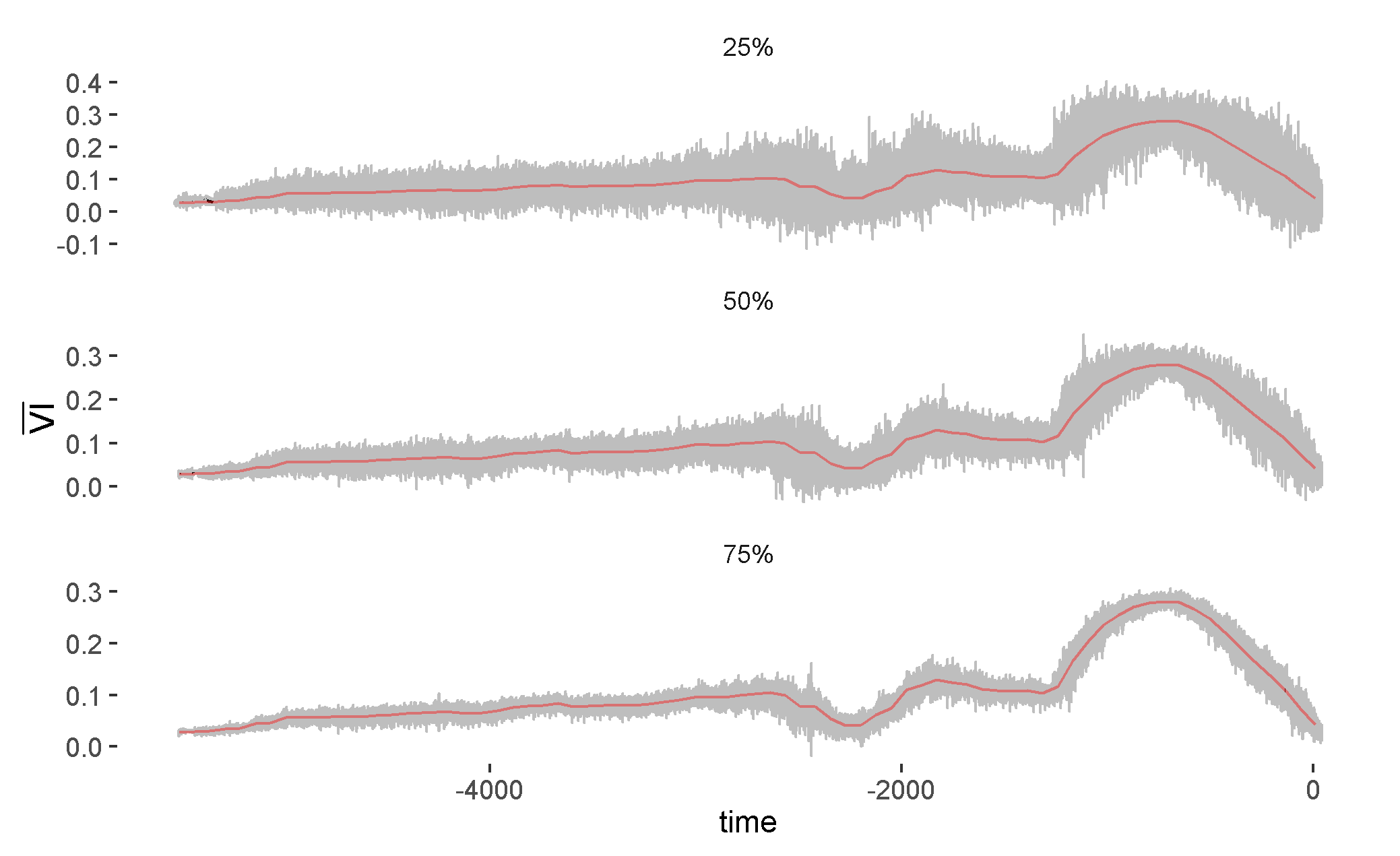
Using a resampling approach I calculated the regime detection measures over different scenarios simulating data quality and data quantity issues common to ecological data anlaysis. The scenarios are categorized as *observations* and *species*. The observations scenario simulates a loss of temporal observations (decreasing the number of times the system was observed), and the species scenario simulates a loss of information about the system by removing some proportion of the species. The loss of temporal observations and the loss of species were examined at three proportions: , where is the proportion of species and time points **retained** for analysis. For example, when , a random selection of of the species are retained for analysis in the species scenario. I resampled the data over iterations () for each scenario and combination. Note that because when , all data are retained. Therefore, no resampling was conducted at this level because only a single metric (e.g. Velocity) value is possible.

### Comparing regime detection measures

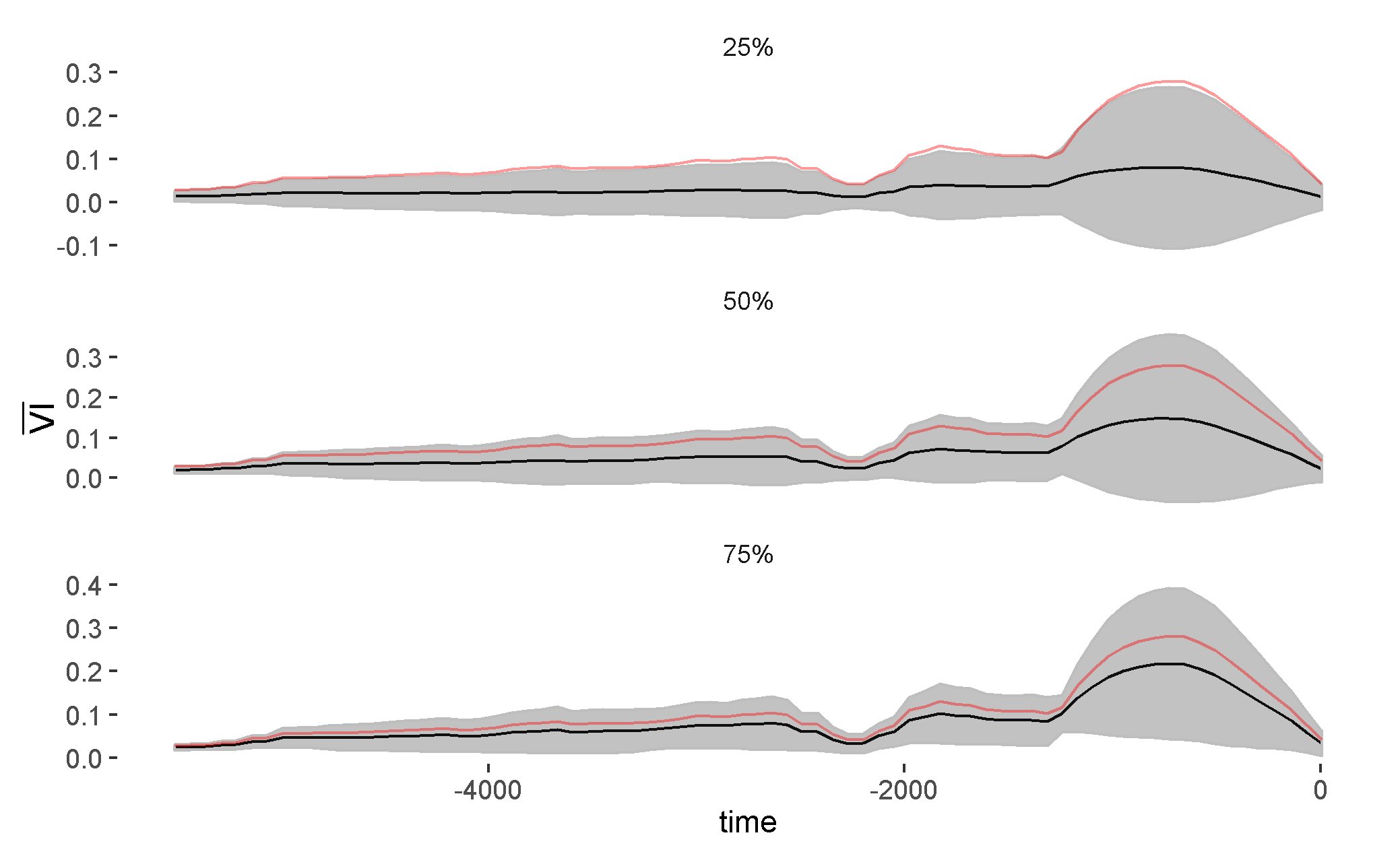
Interpretation of the regime detection measures used in this analysis are currently limited to visual inspection. Therefore, I limit inference in this study largely to the impact of data loss on the variability with a regime detection measure (i.e. how robust is the measure to data loss). It is important to not only identify the influence of data quality and quantity on the performance of individual regime detection metrics, but also to somehow relate these qualities. I visually inspect the relatve performance of these metrics by comparing the coefficient of variation of the resampled samples for the results of resampling method (; species, observations) and sampling percentage (; 25%, 50%, 75%) combination for each metric (FI, VI, ). The coefficient of variation measures provides a relative measure of the variability in the estimated metric across resampled samples as , where is the standard deviation and is the mean value. I observed the distributions of the CV to identify potential flaws in the metrics should data quality or quantity (, ) decrease. First, within a value of a low error to mean ratio (CV) indicates that the metric value is similar across the resampled samples (). The efficacy of the metric should be questioned as CV, and perahps even abandoned as CV. Next, we can examine how the distribution of CV changes within and across . As we increase , we are increasing the volume of data we are feeding the metric. Intuitively, we can assume that as we add more data (volume), we are supplying the metric with more *information*, theoretically increasing the signal-to-noise ratio. Following this logic, we should expect the distribution of CV to generally decrease in mean CV value and also become less variable (less dispersion around the mean CV). A visual examination of the distribution of CV across and within was suffcient to achieve inference regarding the quality of these metrics upon data loss and lessened quality.



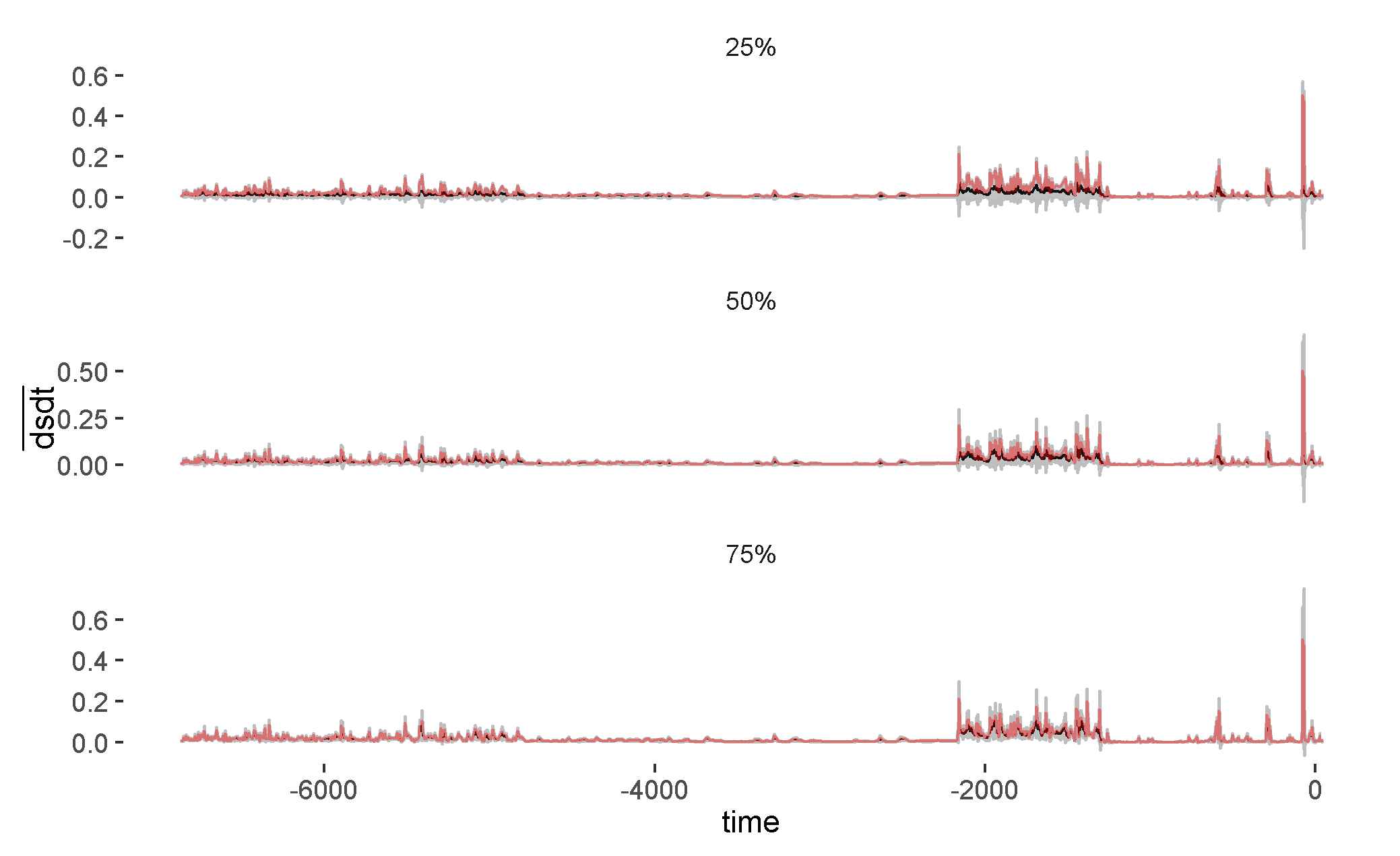
Density plot of the coefficient of variation (CV) as a percentage (%) of the Variance Index resampled values over 10,000 iterations. Densities are drawn based on all values of CV but values >100% are not printed.



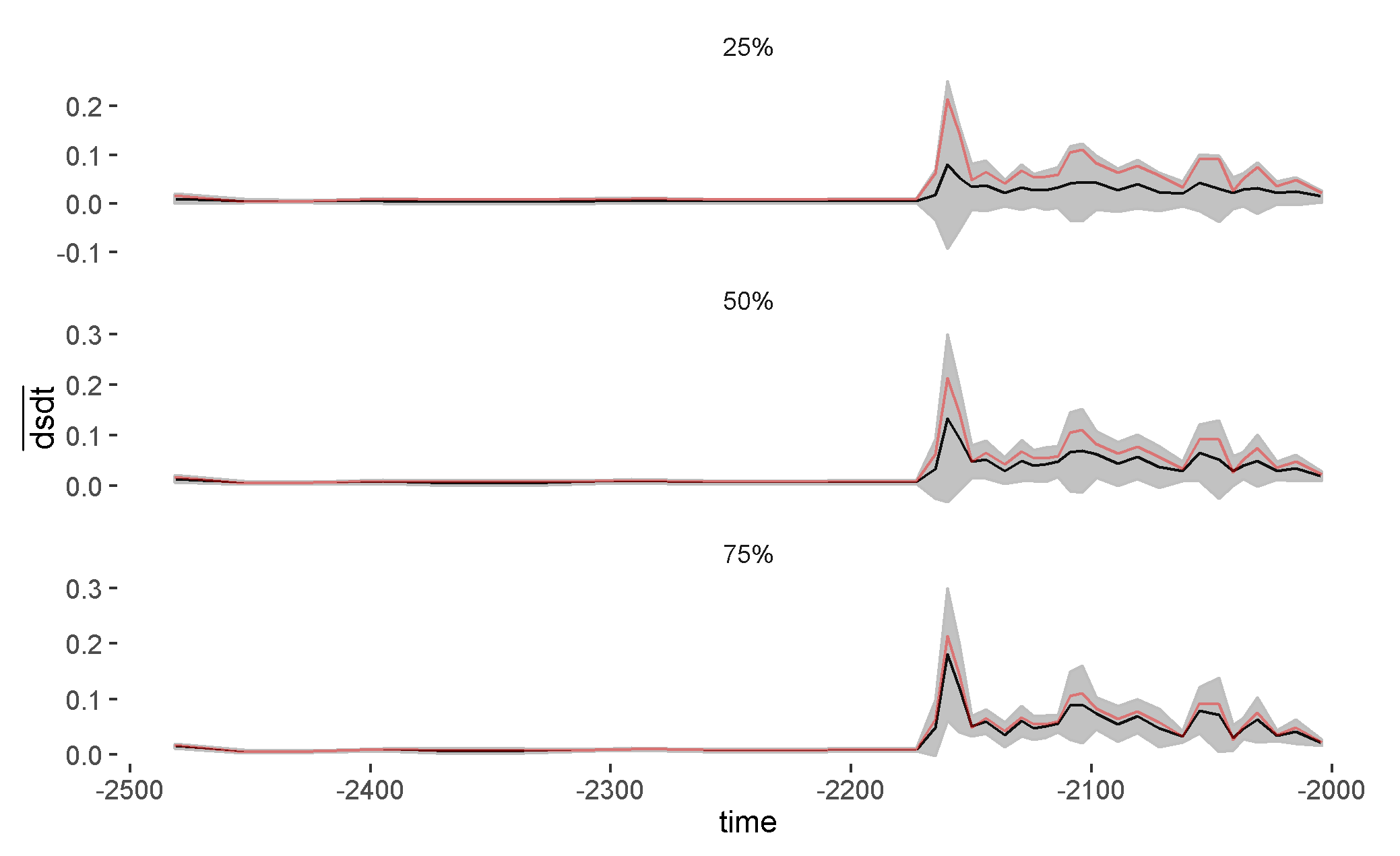
Mean Variance Index (VI) and associated 95% confidence intervals over 10,000 iterations using the **observations** resampling method. Red line indicates the value of VI when **M** and **P** = 100%.



Mean Variance Index (VI) and associated 95% confidence intervals over 10,000 iterations using the **species** resampling method. Red line indicates the value of VI when **M** and **P** = 100%.



Mean velocity and associated 95% confidence intervals over 10,000 iterations using the observations resampling method. Red line indicates the value of velocity when **M** and **P** = 100%.



Mean velocity and associated 95% confidence intervals over 10,000 iterations using the observations resampling method for a subset of the time series (the second ‘regime’ identified). Red line indicates the value of velocity when **M** and **P** = 100%.

## Results

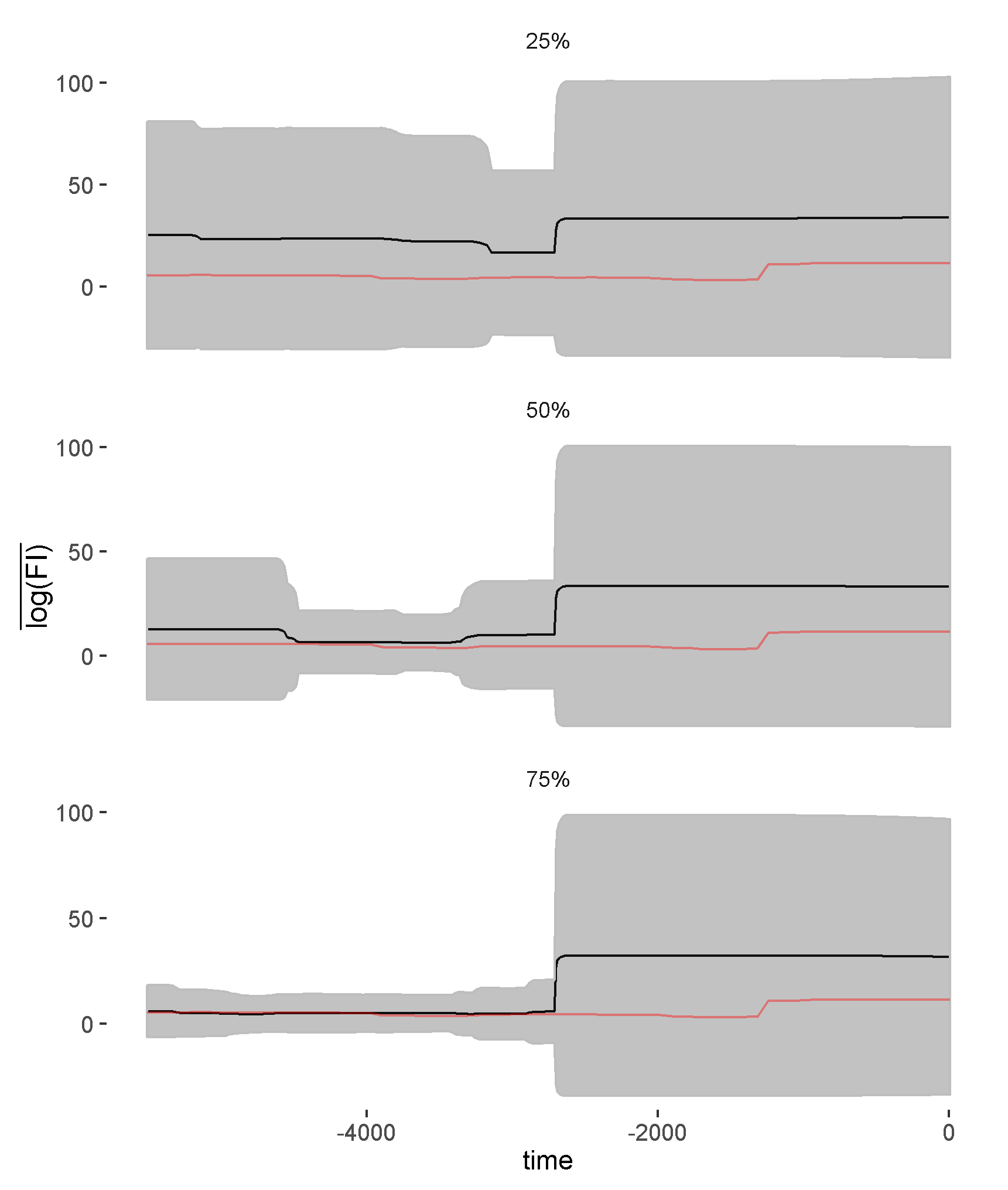
### Velocity of the distance travelled ()

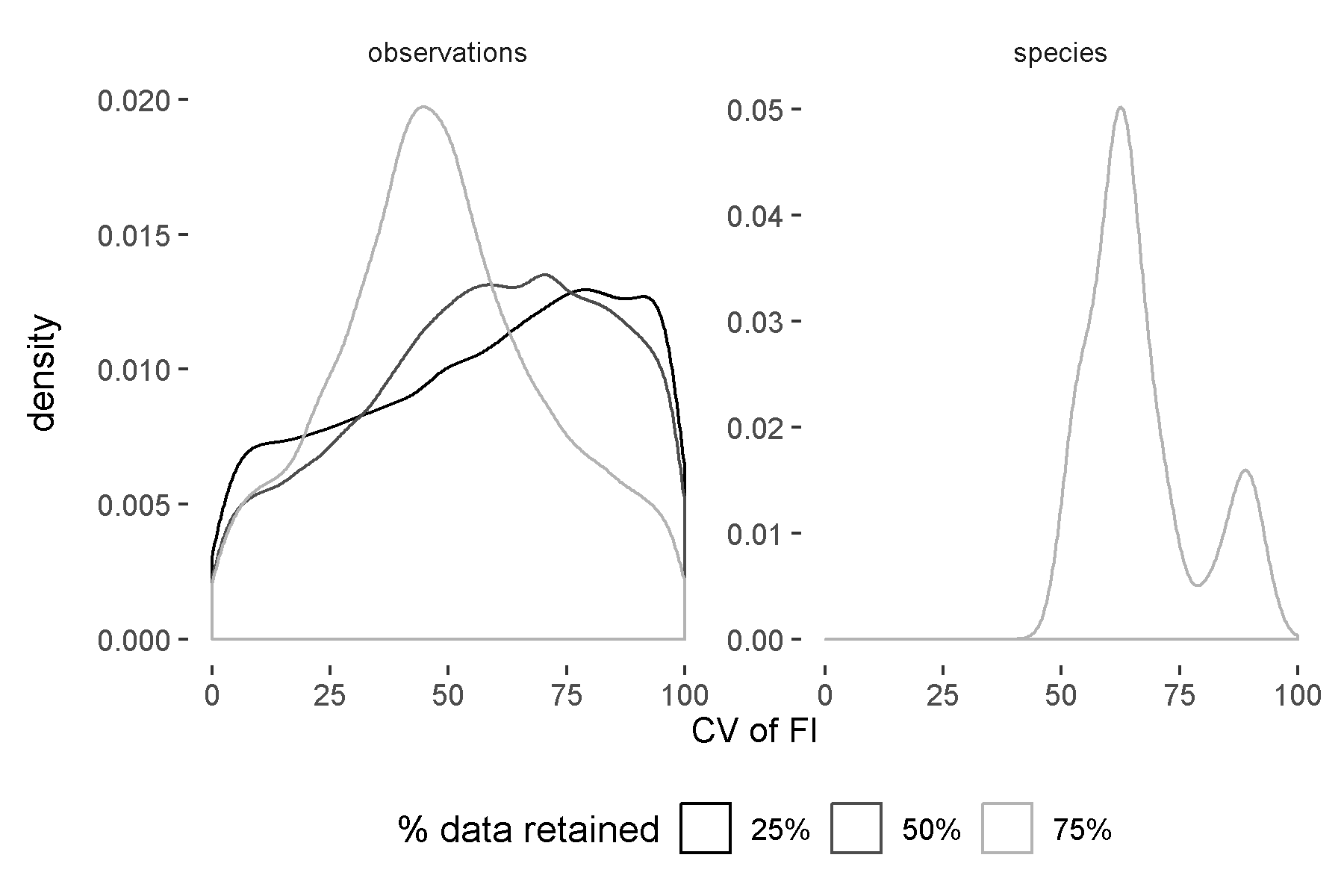
The velocity of the distance travelled, or , exhibited dispersion across the values of , however, yielded consistent results (i.e., high overlap in the densities of the CV across values of and across methodologies; see Figure @ref(fig:dsdtCV)). Further, it should be noted that because is calculated using first differences, it will be sensitive to large changes in the state variables. By examining the density plot of the CV of the dsitance travelled, , we notice that this measure is highly *insensitive* to data loss (Figure @ref(fig:sCV)), suggesting that a finite differencing appraoch (e.g., using total variation regularized differentiation; see Chapter ) which can yield a much smoother derivatie than the approach used here, may decrease the sensitivity of to data loss. This hypothesis is further supported when examining the effect of species (Figure @ref(fig:dsdtResamp)) and temporal observation loss (Figure @ref(fig:dsdtResampRegime2)) on the velocity metric. These conditions are representative of the other combinations.

### Variance Index

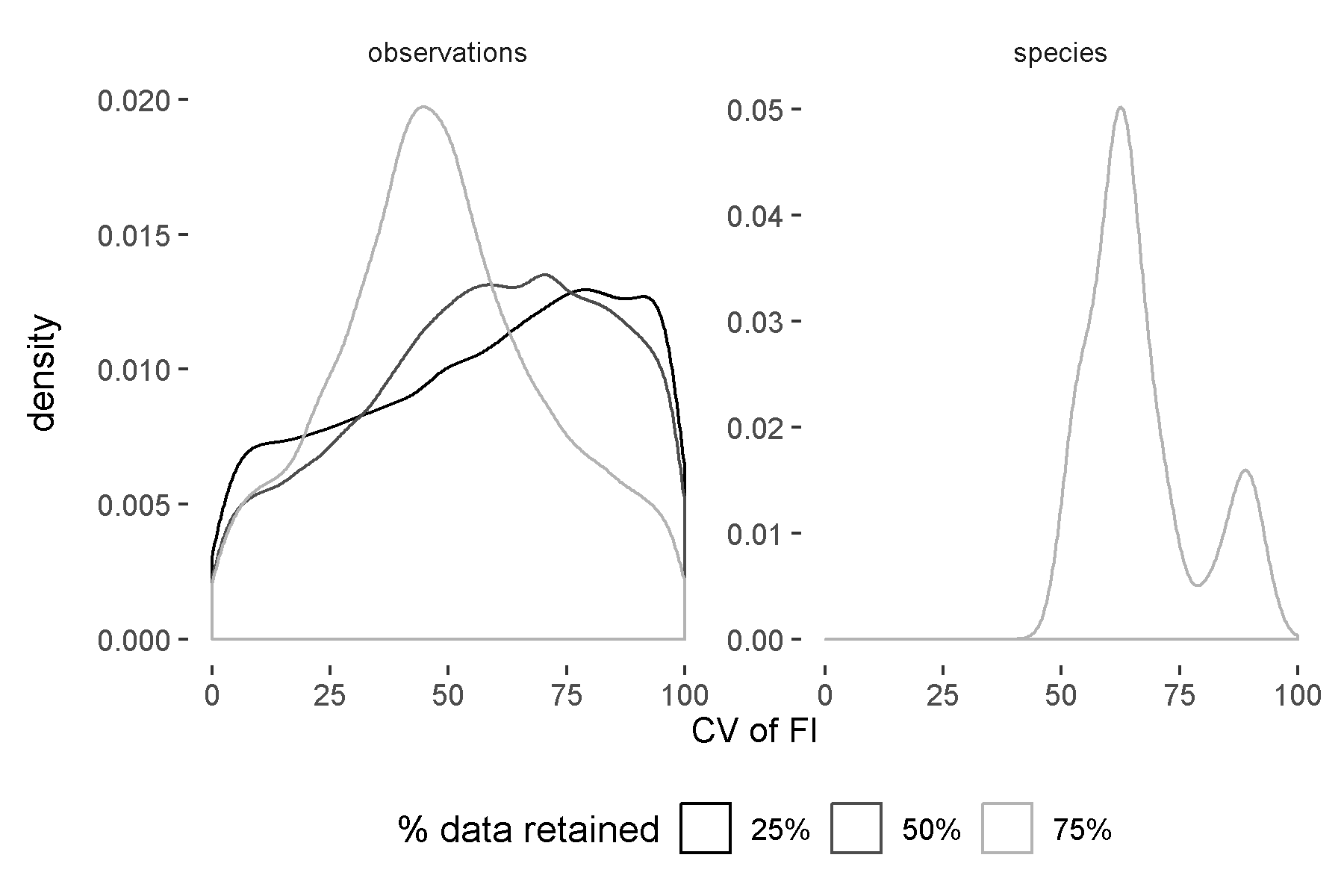
The Variance Index (VI) performed best under the the observations resampling method, exhibiting low values for and low dispersion in the CV density (Figure @ref(fig:viCV)) across iterations. However, the VI appears sensitive to high losses of species information, where the density of the CV still exhibits low dispersion but with higher overall mean values (Figure @ref(fig:viCV)). Surprisingly, the Variance Index was insensitive to temporal observation loss (Figure @ref(fig:viResamp)), exhibiting a similar amount of noise across various degrees of data loss (). Although the signal was dampened under the species method, the signals for the shifts in community composition were not lost across levels of (Figure @ref(fig:viResamp2)). This is likey due to the high probably that the dominant species were rarely always excluded from the resampled observations.

### Fisher Information is highly sensitive to information loss

The Fisher Information method did not yield conclusive results regarding the abrupt shifts in the paleodiatom community composition. Further, this method appears highly sensitive to varying quality and quantitaties of data (Figures @ref(fig:fiResamp), @ref(fig:fiCV)). Although the Fisher Information identifies the shift in community composition at years before present, it fails to identify shifts outside this period. Further, it is difficult to visually analyze any value of the Fisher Information on the original scale as the values range from to (Figure @ref(fig:fiResamp)). In addition to failing to identify the shifts in community composition, the standard deviation of Fisher Information far exceeded the mean value of Fisher Information under all scenarios (Figure @ref(fig:fiCV)). When I resampled the data using 25% and 50% of the species the ratio of mean Fisher Information to standard deviation (CV) of Fisher Information is always (i.e, not pictured in Figure @ref(fig:fiCV)). The high variation in FI values across resampled iterations coupled with the high dispersion within each combination (Figure @ref(fig:fiCV)) suggests Fisher Information will not produce similar trends when we lose or distort the data collected. This is also sugggested by the high confidence intervals surrounding each combination (Figure @ref(fig:fiResamp)).  




## Detrending the Data Prior to Calculations

If and how to manipulate the original data prior to calculating various regime detection methods is an important consideration, and a line of research that has not yet been fully explored. Although most of the multivariate methods identified in the literature review do not require data conforms to a specific distribution, how th results of these methods can vary as we change the quality and characteristics of the original data [@michener2012ecoinformatics]. In fact, since many of the methods for regime shift detection are specifically looking for changes in variance structure and autocorrelation, standardizing variances is not counterintuitive. \begin{landscape}  \end{landscape}

Some studies detrend the original time series prior to data aggregation and calculation of regime detection metrics. I did not detrend the original data for two reasons. First, the authors of the original paper analysing this dataset [@spanbauer\_prolonged\_2014] did not detrend species time series. Like @spanbauer\_prolonged\_2014 I only scaled the original data, rather than detrending. Second, detrending a time series requires yet another subjective decision by the data analyst. For example, a “spanning” parameter must be chosen when detrending (smoothing) non-linear time series using local regression (Loess) regression (see Figure @ref(fig:loessEx)). Other smoothing methods are being explored for both detrending [e.g., PcR; @beck\_variance\_2018] and regime shift identification [e.g., generalized additive modelling; @beck\_variance\_2018]. Finally, this data exhibits rapid and drastic shifts in community composition *and* contains a disproprtionate amount of dominant versus non-dominant species. Consequently, most species contain more zero than non-zero observations, which makes loess smoothing difficult. Although this chapter concerns impacts of data quality and quantity based on hypothetical data collection and analytical decisions, adding yet another parameter necessitates another layer of compartive analysis. Future work studying the impact of detrending, data scaling, outlier removal, and other related decisions would be of value in understanding the efficacy of these and other regime detection measures in real-world situations.

## Conclusion

In this chapter I provide additional evidence for the sensitivity of select regime detection measures to information (data) quality and quantity loss. The loss of data quantity was simulated by randomly sampling subsets of both the species and the temporal observations, and the reduction in data quality manifests as a function of removing whole species from the community profile. Previous studies of the robustness of univariate regime detection metrics have found similar results, suggesting the measures fail in numerous real-world ecological conditions [@andersen\_ecological\_2009;@contamin\_indicators\_2009]. This chapter also highlights the relative insensitivity of the new velocity metric (see Chapters @ref(fiGuide), @ref(velocity)) to data and information quality and quantity (e.g., Figure @ref(fig:dsdtResamp)) loss.

## Ackowledgements

This study was conceptualized at the International Institute for Applied Systems Analysis (IIASA) as part of the Young Scholars Summer Program in 2018. I thank my IIASA program supervisors, Drs. Brian Fath and Elena Rovenskaya, for their advisement of this project and IIASA scientists Drs. Matthias Jonas, Chai Molina, Piotr Zebrowski for feedback on study design.