Identifying Model Misspecification in VAR(1) Models: A Predictive Accuracy Analysis Approach

Ward B. Eiling
S3983919
July 13, 2023

Honours Bachelor Thesis BSc Programme of Psychology (PSB3E-BTHO)
Honours College and Excellence Program
Faculty of Behavioural and Social Sciences
University of Groningen

Supervised by: dr. Laura F. Bringmann
Second evaluator: prof. dr. Casper J. Albers
In collaboration with: MSc Yong Zhang

A thesis is an aptitude test for students. The approval of the thesis is proof that the student has sufficient research and reporting skills to graduate, but does not guarantee the quality of the research and the results of the research as such, and the thesis is therefore not necessarily suitable to be used as an academic source to refer to. If you would like to know more about the research discussed in this thesis and any publications based on it, to which you could refer, please contact the supervisor mentioned.

Identifying Model Misspecification in VAR(1) Models: A Predictive Accuracy Analysis Approach

Ward B. Eiling
University of Groningen

The rising popularity of the network model of psychopathology, which captures the dynamic interplay between symptoms over time, has led to the widespread use of the lag-1 vector autoregressive (VAR(1)) model. Networks based on the VAR(1) model have the potential to provide valuable insights for clinicians in understanding and treating mental disorders. However, to establish whether meaningful inferences can be drawn from these networks, the quality of the model must be ensured. To this end, predictive accuracy analysis (PAA) may be used to evaluate a model's effectiveness to capture dynamics and generalize to unseen data. While simulation-based PAA may identify overfitting, it may overlook violations of model assumptions. This study investigates this limitation by conducting empirical and simulation-based analyses to shed light on the potential consequences of employing large datasets in a psychopathological context. Specifically, it suggests that such datasets may be prone to violate the stationarity assumption, resulting in model misspecification. Therefore, when utilizing the VAR(1) model, researchers are advised to carefully balance sample size, ensuring power and preventing overfitting, while also avoiding model misspecification. In line with the hypothesis, the simulation-based PAA yielded overoptimistic results and failed to identify violations of the model's assumptions. To avoid misinterpretation of inaccurate networks, the study recommends the use of empirical-based cross-validation procedures to evaluate generalizability and predictive accuracy in real-world applications. Future research should address pressing questions regarding predictive accuracy metrics and explore the relationship between VAR assumption violations, model misspecification, and predictive accuracy outcomes.

Keywords: person-specific network, vector auto-regressive modeling, stationarity, model misspecification, predictive accuracy analysis

1 Introduction

Despite persistent efforts spanning well over a century, there has been limited progress in understanding the root causes or specific etiology of mental disorders (Cacioppo & Tassinary, 1990; Hayes et al., 1996; Kendler, 2016; Meehl, 1972; Schleim, 2022; Turkheimer, 1998). In response, trans-diagnostic approaches have emerged, adopting a functional classification approach that organizes behaviors based on the underlying functional processes that produce and maintain them (Hayes et al., 1996). The network approach to psychopathology, in line with this perspective, proposes that symptoms sustain and reinforce each other rather than being mere effects of a shared external source (Borsboom, 2008, 2017; Bringmann et al., 2016; Cramer et al., 2010). For example, in a patient with insomnia, the anticipation of a poor night's sleep can trigger worries about its consequences, further reinforcing sleep difficulties. The interplay between worry and tiredness may then impair emotional regulation, exacerbating the tendency to experience situations that provoke more worry (Wassing et al., 2019). Applying the network approach to person-specific data provides clinicians with insight into the temporal dynamics of a patient's symptoms, enabling tailored interventions that target their specific needs (Bringmann, 2021; von Klipstein et al., 2020).

To gather within-person longitudinal data, the experience sampling methodology (ESM) is commonly employed. This methodology involves collecting numerous in-the-moment assessments from a single subject—through mobile devices or electronic diaries—to record intraindividual variations in psychological processes over time and circumstance (Larson & Csikszentmihalyi, 2014). By obtaining data in the naturalistic settings of individuals' everyday lives, ESM offers a more accurate reflection of daily experiences than laboratory-based or retrospective self-reports. Within the field of psychology, the lag-1 vector autoregressive model, also known as the VAR(1) model, has emerged as the predominant choice for modeling the symptoms and momentary states observed in within-person time-series data (Bringmann, 2021; Bulteel, Mestdagh, et al., 2018). In the multivariate VAR(1) model, each variable is a linear function of all variables (including the variable itself) shifted back one time point (Haslbeck et al., 2021; Lafit et al., 2022). Accord-

ingly, the model enables the estimation of autoregressive effects, which refers to the relation between the current values of a symptom and its past values, as well as cross-regressive effects, which pertain to the association between the current values of a symptom and the past values of another symptom.

Recently, more attention has been drawn to the risk of highly parametrized models such as the VAR model to overfit the data (e.g., Bulteel, Mestdagh, et al., 2018; Bulteel, Tuerlinckx, et al., 2018; Lafit et al., 2022). That is, the model mistakenly treats sample-specific noise for the true underlying signals or patterns in the population (Yarkoni & Westfall, 2017). This results in inaccurate and random parameter estimates that inadequately represent population characteristics, making it difficult to draw meaningful conclusions about the population (Bulteel, Mestdagh, et al., 2018). In regression models, overfitting often arises when there are too many degrees of freedom relative to the sample size (Babyak, 2004). This disparity leads to fluctuations in weights across samples, undermining the reliability of the model. For regression models that employ within-person data, this implies that the primary cause of overfitting is a shortage in the number of time points (Bulteel, Mestdagh, et al., 2018). The extent to which models overfit the data can be assessed by evaluating a model's capacity to predict new or unseen data from the same population. When unseen data is not available for testing, cross-validation (CV) is employed to estimate the predictive accuracy of a model using a single sample. This technique involves dividing the dataset into two parts: one is used to fit the model, while the other contains the values that the model aims to predict.

Aside from overfitting, VAR models face the challenge of misspecification, leading to biased parameter estimates and poor model fit (Cragg, 1968; Kaplan, 1988). Misspecification occurs when the assumptions of a time-invariant VAR model—such as stationarity—are violated (Lütkepohl, 2005). The stationarity assumption (Chatfield, 1980) implies that the mean, variance, and autocovariance of the time-series remain constant over time (Hamaker & Dolan, 2009). However, time-series that involve psychopathological symptoms and a substantial number of timepoints are particularly susceptible to violations of the stationarity assumption (Bringmann et al., 2022). For instance, during the course of therapy for a patient with insomnia, taking more measurements for the symptom 'worry' may enhance the likelihood of observing fluctuations in the mean and variance over time. As the patient improves their sleep hygiene and experiences significant therapeutic benefits, there may be a noticeable reduction in average symptom levels compared to the early stages of therapy. To identify model misspecification arising from the violation of VAR assumptions, the cross-validation framework can be utilized.

Previous simulation studies conducted by Bulteel, Mestdagh, et al. (2018) and Lafit et al. (2022)—that employed (vector) autoregressive models—have consistently shown that larger time-series datasets yield higher predictive accuracy. However, it is important to recognize the inherent trade-off: while a larger number of time points reduces the risk of overfitting, it simultaneously increases the likelihood of violating model assumptions. Despite this trade-off, current simulation-based predictive accuracy analysis (PAA) methods ensure stationarity in extensive time-series (e.g., Bulteel, Mestdagh, et al., 2018; Lafit et al., 2022; Revol et al., 2023). As a result, this simulation-based approach may not accurately capture the issue of misspecification in extensive psychopathological time-series data, leading to an overly optimistic assessment of model performance that may not generalize well to empirical data.

To address this gap in the literature, we will investigate the impact of using empirical test data as opposed to simulated test data on predictive accuracy. This dissertation aims to accomplish two objectives. First, it aims to illustrate that VAR(1) models, estimated from a psychopathological dataset with many timepoints, can be susceptible to model misspecification. Second, it seeks to highlight that current simulation-based predictive accuracy procedures, which assume stationarity, may produce overly optimistic outcomes that neglect concerns of model misspecification. Based on these premises, it is hypothesized that a VAR(1) model estimated from a large time-series exhibits higher predictive performance when tested with simulated data compared to empirical data. The remainder of this paper proceeds as follows. First, we will shed light on the theoretical underpinnings and rationale behind the methods utilized. Next, we discuss the procedures used to conduct the simulation and empirical analyses. We then compare these findings and conclude with a summary discussion.

2 The Person-Specific VAR(1) Model

In the multivariate person-specific VAR(1) model, the data of each individual are treated separately, with each variable regressed on all variables p ($p = 1, 2, 3, \ldots, P$), including itself, at the previous measurement occasion. In the person-specific VAR(1) model, the $(P \times 1)$ vector $\mathbf{y}_t$, representing the variable scores at time t ($t = 1, 2, 3, \ldots, T$) is modeled using the equation

$$\mathbf{y}_t = \boldsymbol{\delta} + \boldsymbol{\Phi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \tag{1}$$

Here, $\mathbf{y}_{t-1}$ is the person's vector of variable scores at the previous measurement occasion. The vector $\boldsymbol{\delta}$ is a $(P \times 1)$ vector representing the intercepts. The matrix $\boldsymbol{\Phi}$ is a $(P \times P)$ matrix, which contains the regression coefficients that quantify the autoregressive effects (i.e., the diagonal elements) and cross-regressive effects (i.e., the off-diagonal elements) of the previous states on the current states. The vector $\boldsymbol{\varepsilon}_t$ is a $(P \times 1)$ vector of residuals at time t that captures the part of the variables' values that cannot be predicted based on their

values at the previous time point. The vector of residuals $\boldsymbol{\varepsilon}_t$ is commonly referred to as the "innovations" to emphasize their nature as new or unexpected information that cannot be attributed to past values.

For instance, suppose we are investigating the temporal relationships between worry and sleep problems in a single patient. Let us denote their values at timepoint t as W_t and S_t, respectively. Estimating the person-specific bivariate VAR(1) model from these symptoms would yield:

$$W_t = \delta_w + \Phi_{ww} W_{t-1} + \Phi_{ws} S_{t-1} + \varepsilon_{w,t} \tag{2}$$

$$S_t = \delta_s + \Phi_{sw} \mathrm{W}_{t-1} + \Phi_{ss} \mathrm{S}_{t-1} + \varepsilon_{s,t} \tag{3}$$

Note that in the case of two variables, the VAR(1) model comprises 9 model parameters. Let us now assume that we fitted a bivariate VAR(1) to these symptoms and obtained the following parameter estimates:

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_w \\ \delta_s \end{bmatrix} = \begin{bmatrix} 1.0 \\ 1.2 \end{bmatrix} \tag{4}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} \Phi_{ww} & \Phi_{ws} \\ \Phi_{sw} & \Phi_{ss} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.5 \end{bmatrix} \tag{5}$$

$$\boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{w,t} \\ \varepsilon_{s,t} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} \right) \tag{6}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{ww} & \sigma_{ws} \\ \sigma_{sw} & \sigma_{ss} \end{bmatrix} = \begin{bmatrix} 10 & 4 \\ 4 & 10 \end{bmatrix} \tag{7}$$

In this VAR(1) model, $\boldsymbol{\delta}$ is a the (2×1) vector that represents the intercepts, containing δ_w and δ_s. The matrix $\boldsymbol{\Phi}$ is a (2×2) matrix containing the autoregressive and cross-regressive coefficients, where the diagonal elements Φ_{ww} and Φ_{ss} represent the autoregressive effects and the off-diagonal elements Φ_{ws} and Φ_{sw} resemble the cross-regressive effects. These parameter estimates indicate that the patient's worry and sleep problems at a previous timepoint $(t-1)$ are positively associated with their worry and sleep problems at the current timepoint (t). The innovation vector $\boldsymbol{\varepsilon}_t$ includes the (2×2) variance-covariance matrix $\boldsymbol{\Sigma}$. The diagonal elements of $\boldsymbol{\Sigma}$ represent the variances of the individual error terms ($\varepsilon_{w,t}$ and $\varepsilon_{s,t}$), while the off-diagonal elements represent their covariance.

By utilizing the parameter estimates derived from Equation 4 and 5, along with the lagged data points W_{t-1} and S_{t-1}, it is possible to make predictions on the present observation W_t. Let us assume that the reported values of W_{t-1} and S_{t-1} are 3 and 5 respectively where t is 8. Inserting these values in Equation 2 yields the prediction:

$$\hat{W}_{t=8} = 1.0 + 0.6 \times 3 + 0.4 \times 5 = 4.8 \tag{8}$$

The VAR(1) model has several assumptions. First, it is assumed that the innovations follow a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}$. Second, the assumption of global stationarity is imposed, indicating that the process $\mathbf{y}$ is time-invariant with respect to its mean, variance, and autocorrelation structure. The global stationarity assumption holds when the modulus of the eigenvalues of the matrix $\boldsymbol{\Phi}$ is less than one (Lütkepohl, 2005). In the event that this condition is not satisfied, the process outlined by the system of Equation 1 remains coherent; however, the state vector $\mathbf{y}_t$ will progressively diverge towards infinity (Loossens et al., 2021). This signifies system instability, rendering the model less reliable for accurate predictions. Third, the local stationarity assumption holds that the statistical properties of the time-series remain relatively constant within sub-samples of the data. It is important to note that this assumption can be violated even if the global stationarity assumption is fulfilled, as there may exist time-varying dynamics or structural breaks within certain sub-samples.

Estimation of the person-specific VAR(1) model is commonly performed using separate ordinary least squares (OLS) regressions for each variable or maximum likelihood estimation (MLE; Hamilton, 1994; Lütkepohl, 2005). Although MLE is more computationally intensive and complex than OLS, the asymptotic properties of the two approaches are identical (Lütkepohl, 2005). Another popular estimation method is provided by the Mplus software, which employs the dynamic structural equation modeling (DSEM) approach to directly estimate the full VAR(1) model with Bayesian methods (Asparouhov et al., 2018; Hamaker et al., 2018). While the implementation carries potential benefits, the closed-source license reduces the accessibility of the software, impeding the reproducibility of analyses. In contrast, the open-source software R (R Core Team, 2022) enables straightforward documentation of code in the implementation of OLS. Accordingly, in line with other simulation-based predictive accuracy studies by Lafit et al. (2022) and Revol et al. (2023), we adopt OLS for estimating a person-specific VAR(1) model. This approach involves performing a multiple linear regression analysis for each variable individually, where each variable serves as the dependent variable and the lagged values of all variables (including itself) act as independent variables.

3 Cross-Validation

To assess the predictive performance of a model, the cross-validation framework is commonly employed. This framework can evaluate how well a model can generalize to unseen data, detect overfitting, indicate model misspecification, select the best model among alternatives, assess model robustness, and optimize resource efficiency. In a cross-validation framework, the available data is partitioned into a training set and a test set. The models under investigation are fitted to the training set and the estimated parameters are used to predict observation(s) in the test set.

In a simulation study involving datasets with varying numbers of observations, Bulteel, Mestdagh, et al. (2018) compared different cross-validation procedures, including leave-one-out CV (LOOCV) and blocked CV, to evaluate the predictive accuracy of lag-1 (vector) autoregressive models. These procedures differ in their approaches to constructing the training and test sets. Namely, unlike blocked CV, LOOCV does not consider the serial dependency of a time-series. This discrepancy explains why LOOCV outperforms blocked CV when the simulated data lack time dependence but performs worse when time dependence is present.

In the condition with the most time points ($T = 500$), Bulteel, Mestdagh, et al. (2018) found that leave-one-out CV yielded the highest percentage of data sets for which the best predictive model was selected. Given the abundance of time points in the dataset used for this study, we anticipate that this procedure will yield sufficiently accurate results. An additional advantage of LOOCV over other CV methods is its ability to maximize the utilization of the available data. Accordingly, using the technique of LOOCV, we will "set aside one individual case, optimize for what is left, then test on the set-aside case" (Mosteller & Tukey, 1968, as cited in Stone, 1974).

Figure 1 illustrates how in each iteration of the LOOCV procedure, a distinct observation is systematically excluded from the dataset to form the test set (Hastie et al., 2009). The remaining data, known as the training set, is fitted to the model. Subsequently, the model parameters are employed to predict the value of the excluded observation (e.g., Equation 8). This iterative process is carried out for every timepoint t ($t = 1, 2, 3, \ldots, T$) in the dataset, ultimately generating a comprehensive collection of predictive values.

4 Predictive Accuracy Metrics

After the predicted values are computed for the test set, different predictive accuracy metrics may be employed in PAA. Previous studies examining the predictive accuracy of VAR(1) models have commonly utilized the Mean Squared Prediction Error (MSPE) as a performance metric (Bulteel, Mestdagh, et al., 2018; Bulteel, Tuerlinckx, et al., 2018; Lafit et al., 2022). These studies computed the MSPE for each variable p ($p = 1, \ldots, P$) by squaring and averaging the differences between the observed and predicted values in the test set ($t = 1, \ldots, T_{\text{Test}}$). In a LOOCV, the mean squared prediction error for variable p, MSPE_p, may be computed as follows:

$$\text{MSPE}_p = \frac{1}{T_{Test}} \sum_{t=1}^{T_{Test}} (\text{y}_{Test,p} - \hat{\text{y}}_{Test,p})^2 \tag{9}$$

In this equation, $\text{y}_{Test,p}$ denotes the observed value of variable p at timepoint t, and $\hat{\text{y}}_{Test,p}$ represents the predicted value of variable p at timepoint t based on the training data excluding that particular timepoint. The squared prediction error $(\text{y}_{Test,p} - \hat{\text{y}}_{Test,p})^2$ may be obtained by squaring the difference between the observed and predicted value. By summing the squared prediction errors across all T_{Test} timepoints and dividing by the total number of timepoints T_{Test}, we obtain the MSPE estimate for variable p.

An overall MSPE across all variables may then be computed by averaging the MSPE_p over the variables p ($p = 1, \ldots, P$)[1]:

$$\text{MSPE} = \frac{1}{P} \sum_{p=1}^{P} \text{MSPE}_p \tag{10}$$

MSPE is a reliable metric for univariate models like the lag-1 autoregressive (AR(1)) model, which focuses on a single variable. However, its suitability decreases when applied to multivariate models such as the VAR(1) model, which involves multiple variables. This is because MSPE assumes a constant innovation variance across variables, which does not hold in multivariate models, where each variable can have a unique pattern of innovation variance. Despite this, when calculating the overall MSPE for multivariate models, the prediction errors of each variable are averaged and squared without accounting for potential variations in innovation variances or the impact of covariance between prediction errors of different variables (Revol et al., 2023). In essence, MSPE treats all variables equally in terms of error, which can lead to misleading evaluations that disregard the intricate complexities and interdependencies inherent in multivariate models. Hence, recent studies (Lafit et al., 2022; Revol et al., 2023) have highlighted the need for caution when interpreting MSPE in VAR(1) models.

To address these issues, Revol et al. (2023) proposed the use of Mahalanobis distance—a statistic that takes into account the innovation covariances—which is squared for every timepoint to represent a standardized multivariate prediction error:

$$D^2 = (\text{y}_{\text{Test}} - \hat{\text{y}}_{\text{Test}})^T \boldsymbol{\Sigma}^{-1} (\text{y}_{\text{Test}} - \hat{\text{y}}_{\text{Test}}) \tag{11}$$

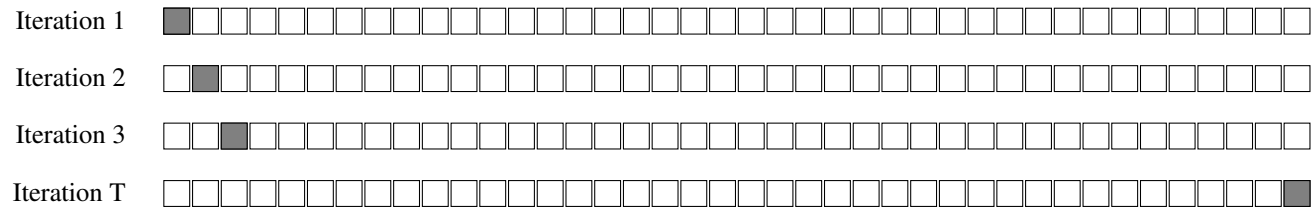
Here, y_{Test} is a $(P \times 1)$ vector that represents the observed values of the test sample and $\hat{\text{y}}_{\text{Test}}$ is a $(P \times 1)$ vector that represents the estimated or predicted values of the test sample. The matrix $\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix of the training sample. By employing $\boldsymbol{\Sigma}^{-1}$, the resulting distance measure D^2 can account for the underlying multivariate distribution of the innovation terms.

If the true model parameters are used to predict the test set, the squared Mahalanobis distances D^2 conform to a univariate χ^2 distribution with degrees of freedom equal to the number of variables (P). This implies that if a training set

[1] Note that averaging the MSPEp across variables does not make it a multivariate statistic.

Figure 1

Leave-One-Out Cross-Validation Procedure



Note. For a dataset of $T = 40$ observations, the figure represents the elements of the test set as dark-colored squares and the elements of the training set as blank squares. The number of timepoints in each iteration, as well as the total number of iterations, is equal to T.

is sufficiently large, the distribution of D^2 should approximate the χ^2 distribution. Conversely, if the training set is too small, larger D^2 will be observed, leading to a distribution with a heavier right tail. This behavior was employed to quantify the predictive accuracy of the VAR(1) model estimates. Specifically, Revol et al. (2023) computed the proportion of D^2 among the associated prediction errors that fall below the 95th percentile of the $\chi^2(P)$ distribution for a given training set n ($n = 1, \ldots, N_{\text{Training}}$). This proportion was denoted as $p_{\text{Mal},.95}$. To assess the predictive accuracy, Revol et al. (2023) proposed a threshold of .94, whereby $p_{\text{Mal},.95}$ values surpassing this threshold indicate satisfactory predictive performance for a specific training set. To provide an overall assessment of how $p_{\text{Mal},.95}$ performed across all training sets, the predictive accuracy probability (PAP) may be computed. The PAP represents the proportion of generated training sets in which $p_{\text{Mal},.95}$ exceeded the predefined threshold of .94. Subsequently, Revol et al. (2023) established a criterion for the sufficient PAP of 0.8, serving as a benchmark to determine the minimum sample size required for obtaining generalizable results.

5 Methods

This study carried out two main analyses to explore the potential overoptimistic predictive performance of the simulated test data: 1) a leave-one-out cross-validation (LOOCV) using the empirical test set and 2) a predictive accuracy analysis for the stimulated test set. Given the extensive number of timepoints ($t = 1, \ldots, 966$) involved in both analyses, a high predictive accuracy should be expected with minimal overfitting. However, on the one hand, it is anticipated that the empirical-based analysis will detect misspecification, resulting in a lower predictive accuracy. On the other hand, the simulation-based analysis is expected to overlook the misspecification and maintain a high predictive accuracy. To check for the robustness of the results, two different predictive accuracy metrics, namely MSPE and Mahalanobis distance, were compared. All analyses were performed in R version 4.2.0 (R Core Team, 2022). The scripts used to conduct the analyses described in this paper will be accessible on the project's OSF page (https://osf.io/7pm89/) two years after the project's completion. The following section provides an overview of the data and outlines the procedures employed to evaluate stationarity and predictive accuracy.

5.1 Data Description

To illustrate the detection of model-misspecification through empirical- and simulation-based PAA, we utilized the Peter Groot (2010) dataset, which was previously examined by Wichers and Groot (2016), Cabrieto et al. (2018), and Albers and Bringmann (2020). These studies specifically identified departures from stationarity in the dataset, manifested as changes in the autocorrelation of momentary symptoms over time. This phenomenon, commonly referred to as 'critical slowing down,' suggests that when the underlying dynamics of symptoms deviate from stationarity, it may indicate a 'critical state' where treatment interventions are most effective (van de Leemput et al., 2014).

The dataset was collected through ESM and consisted of 2390 prospective momentary observations of daily life experiences from a mental health care user (Kossakowski et al., 2017; Wichers & Groot, 2016). The individual in question has a medical history characterized by recurring episodes of Major Depression (MD) and has been prescribed antidepressant medication for the past 8.5 years[2]. Wichers and Groot (2016) examined the data according to the complex dynamical system theory and estimated a VAR(1) model from five symptoms. Among these symptoms, the item 'worry' was used as a measure of cognition, and the item 'suspicious' was used to assess psychotic experiences. The remaining three symptoms, namely negative affect, positive affect, and mental unrest, were derived from principal components analysis conducted on several affect items. For the purposes of maintaining consistency with the original study Wichers and

[2]For more information on the items used and the publicly accessible OSF page, please refer to Kossakowski et al. (2017).

Groot (2016)—while avoiding the complexity of retrieving the principal components[3]—the current study replaced the three components with single items. Specifically, 'satisfied' replaced the component 'positive affect', 'down' replaced 'negative affect', and 'restless' replaced 'mental unrest'.

5.2 Preliminary Analysis

Prior to conducting the main predictive accuracy analyses, the data underwent pre-processing, and a VAR(1) model was fitted. For simplicity, listwise deletion was employed to handle the missing data, selecting cases that have no missing values across all variables and discarding cases with missing values (Schafer & Graham, 2002). This procedure assumes that missing data are missing completely at random (MCAR)[4]. Once the missing observations of these symptoms were handled through listwise deletion, 966 observations remained, and 1424 were discarded. Although the reduction in power resulting from discarding this many cases may be deemed tolerable, it raises doubts about the assumption that the missing data are MCAR. Subsequently, the VAR(1) was fitted on the remaining observations, and the parameter estimates were extracted. This resulted in the (5×1) vector $\boldsymbol{\delta}$, the (5×5) matrix $\boldsymbol{\Phi}$, and the (5×5) variance-covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_{pa} \\ \delta_{mu} \\ \delta_{na} \\ \delta_{wo} \\ \delta_{su} \end{bmatrix} = \begin{bmatrix} 3.24 \\ 1.24 \\ 0.62 \\ 1.09 \\ 0.77 \end{bmatrix} \tag{12}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} 0.27 & -0.07 & -0.18 & 0.01 & -0.02 \\ -0.03 & 0.40 & 0.06 & 0.00 & 0.06 \\ -0.16 & -0.03 & 0.17 & 0.19 & -0.00 \\ -0.08 & 0.00 & 0.06 & 0.41 & 0.03 \\ -0.01 & 0.04 & 0.05 & 0.07 & 0.25 \end{bmatrix} \tag{13}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.79 & -0.34 & -0.30 & -0.25 & -0.15 \\ -0.34 & 0.69 & 0.12 & 0.09 & 0.12 \\ -0.30 & 0.12 & 0.41 & 0.30 & 0.14 \\ -0.25 & 0.09 & 0.30 & 0.51 & 0.19 \\ -0.15 & 0.12 & 0.14 & 0.19 & 0.24 \end{bmatrix} \tag{14}$$

Note. The parameter estimates in each column and row consistently align with the associated variables in the following order: positive affect, mental unrest, negative affect, worry, and suspicious.

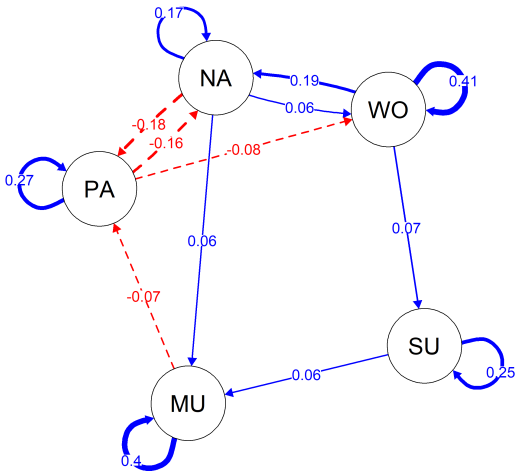
All eigenvalues of the matrix $\boldsymbol{\Phi}$ had a modulus of less than one, indicating that the global stationarity assumption of the VAR(1) model was not violated.

5.3 Network

To provide intuitive insight into the VAR(1) model, a network visualization was created (see Figure 2). The network depicted variables as nodes and represented edges based on the VAR(1) coefficients. Positive affect was represented by the node 'PA', mental unrest by 'MU', negative affect by 'NA', worry by 'WO', and suspicious by 'SU'. The network diagram displayed the significant edges of the matrix $\boldsymbol{\Phi}$ (Equation 13) with autoregressive effects as self-loops and cross-regressive effects as directed arrows connecting distinct nodes. To generate the network diagram, we utilized the qgraph package (Epskamp et al., 2012). Positive VAR(1) weights were depicted as blue solid arrows, while negative weights were represented as red dashed arrows. The thickness of the edges was relative to the magnitude of the corresponding VAR(1) coefficient, with thicker edges indicating larger coefficients for lagged effects.

Figure 2

Network Depicting the VAR(1) Coefficients



Note. Only edges that are significantly different from 0 ($p < .05$) are visualized in the network, where each edge represents the point estimate of a lagged effect.

5.4 Stationarity Assessment

Given that the studies by Wichers and Groot (2016), Cabrieto et al. (2018), and Albers and Bringmann (2020) utilized different variables, we performed additional assessments of stationarity for each variable (see Appendix A).

[3]Wichers and Groot (2016) provides no documentation on the methods used to retrieve the principal components.

[4]Note that listwise deletion may bias parameter estimates if missing data are not MCAR and can still be inefficient in multivariate analyses if MCAR holds Schafer and Graham (2002).

Specifically, we conducted univariate time-series analyses at lag 1 using the Augmented Dickey-Fuller (ADF) test to detect unit roots (Dickey & Fuller, 1979), and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test to assess trend stationarity (Kwiatkowski et al., 1992). For the ADF test, the null hypothesis was rejected for all variables, indicating stationary behavior without a unit root (see Table A1). However, the KPSS test for trend stationarity yielded mixed results (see Table A2): the null hypothesis was rejected for four of the five variables, suggesting the presence of a trend around a constant mean. The contrasting results from these tests imply ambiguity in the stationarity characteristics of the time-series.

The local stationarity of the variables was evaluated by employing the ADF and KPSS test statistics as rolling statistics over time with a window size of 50 timepoints. First, the ADF test statistic was consistently less negative than the critical value for four of the five time-series, indicating that there may be structural breaks within the time-series (see Figure A1). Second, the KPSS test statistic exceeded the critical value at several points in all five variables, illustrating temporary departures from stationarity in the time-series (see Figure A2). Taken together, the results imply that the data exhibit non-stationary behavior in sub-samples of the data.

5.5 Empirical-Based Cross-Validation Analysis

To evaluate the predictive performance of the VAR(1) model estimated from empirical data, we utilized the LOOCV technique (as shown in Figure 1). This technique allowed us to calculate both the MSPE and the squared Mahalanobis distance. Since the analyses share many similarities, the latter will focus solely on highlighting the differences.

5.5.1 MSPE Evaluation Using LOOCV

In this subsection, we utilized the LOOCV technique discussed in Section 3 to calculate the MSPE_p for each of the five variables, as well as the average MSPE across all variables. For each iteration, we followed these steps:

Step 1: Create a Training and Test Set. The training set ($T_{\text{Training}} = T - 1 = 965$) was constructed by excluding a single unique observation from the dataset. The excluded observation formed the test set ($T_{\text{Test}} = 1$).

Step 2: Fit the Model. A person-specific VAR(1) model was fitted on the training set as discussed in Section 2.

Step 3: Make Predictions. Based on model parameters of the fitted VAR(1) model, predictions were made for the test set as described in Section 2 (e.g., Equation 4).

Step 4: Compute Prediction Errors. The prediction errors were computed by subtracting the predicted values $\hat{\text{y}}_{\text{Test}}$ from the current observations y_{Test} (see Equation 9).

This iterative process was carried out for every timepoint t ($t = 1, \ldots, 966$) in the dataset, generating a comprehensive collection of predictive errors. To quantify the prediction accuracy, the MSPE_p was calculated for each criterion variable (as discussed in Section 4; see Equation 9). Next, the MSPE_p scores were averaged over all variables, yielding an average MSPE (see Equation 10; see R code heading "Mean Squared Prediction Error").

5.5.2 Mahalanobis Distance Evaluation Using LOOCV

In this subsection, we employed the LOOCV technique to compute the squared Mahalanobis distance and the estimated predictive accuracy (%). Again, Steps 1 through 4 (outlined in Section 5.5.1) were followed to obtain a collection of prediction errors. Subsequently, the squared Mahalanobis distances were computed (as explained in Section 4; see Equation 11). To assess the estimated predictive accuracy of the model, the proportion of squared Mahalanobis distances that were lower than the 95th percentile of the $\chi^2(df = 5)$ distribution (see R code heading "Estimated Predictive Accuracy").

5.6 Simulation-Based Predictive Accuracy Analysis

Besides the cross-validation analysis with empirical test data, we undertook a simulation study to create a benchmark scenario that upholds VAR assumptions and avoids any model misspecification. Adding onto the empirical-based analysis, the simulation-based analysis performed another round of computations for the MSPE and the squared Mahalanobis distance. As there were similarities between the analyses, this section specifically aims to emphasize the differences.

5.6.1 MSPE Evaluation in the Simulation Study

This subsection will provide an overview of the steps taken to perform a simulation-based PAA to compute the MSPE_p for each variable, as well as the average MSPE.

Step 1: Select Parameters. Prior to simulating the data, we need to obtain the parameters from the VAR(1) model estimated using the empirical dataset. These model parameters were obtained in Section 5.2 (see Equation 12, 13 and 14).

Step 2: Generate Training Data. To simulate the training data, several steps were undertaken. First, the specifications were defined, including the number of simulated time points ($T_{\text{Training}} = 966$) and the number of sets to simulate ($N_{\text{Training}} = 1000$). In addition, a seed was set to ensure reproducibility[5] and the number of burning observations were established ($T_{\text{Burning}} = 1000$). By eventually discarding the burning observations, bias can be reduced[6] and models can

[5] Setting a seed ensures that the same sequence of random numbers will be generated each time the code is run.

[6] Discarding the burning observations can reduce bias introduced during the initial phase of the simulation, as the initial observations may not accurately represent the true system behavior.

reach a steady-state behavior. Next, the innovations were simulated according to a multivariate normal distribution. The starting value of each time-series is then set by combining the expected value and the innovation at the first timepoint. For each training set n ($n = 1, \ldots, 1000$), the entire time-series was then simulated using the VAR model parameters, incorporating the intercepts $\boldsymbol{\delta}$, the matrix $\boldsymbol{\Phi}$, and the simulated innovations $\boldsymbol{\varepsilon}$. Finally, the burning observations were excluded from the data and lagged variables were created.

Step 3: Fit the Model. To obtain the model parameters, a VAR(1) model was fitted on each training set n ($n = 1, \ldots, 1000$). For each variable, the intercept δ, the autoregressive coefficient, and the four cross-regressive coefficients were collected and organized for further analysis.

Step 4: Generate Test data. To evaluate whether the VAR(1) model of a given training set could make accurate predictions, a test set was generated. This test set serves as a representation of unseen values, allowing for a direct comparison with the predicted values. The complete time-series for the test set was simulated using the same procedure outlined in Step 2. However, in this case, the simulation of test data encompassed a much larger number of timepoints ($T_{\text{Test}} = 100,000$). This ensured that predictive accuracy was not affected by sampling variability resulting from a small testing set.

Step 5a: Compute MSPE. The final step involved the computation of the MSPE_p for each variable estimated from each training set n ($n = 1, \ldots, 1000$), as well as the average MSPE. The computation followed the same procedure outlined in Section 5.5.1, with the only difference being that the test and training sets used in this analysis were larger in size than in the empirical analysis (see R code headings "Step 5a: Compute MSPE").

5.6.2 Mahalanobis Distance Evaluation in the Simulation Study

In this subsection, we describe the steps taken to compute the estimated predictive accuracy and predictive accuracy probability in a simulation-based analysis by employing the squared Mahalanobis distance. This evaluation consists of five steps, with Steps 1 through Step 4 being identical to the steps presented in Section 5.6.1.

Step 5b: Compute Estimated Predictive Accuracy and PAP. Compute Estimated Predictive Accuracy and PAP. In the alternative fifth step, the predictive accuracy of the estimated models and the PAP were computed. The estimation of predictive accuracy followed the same procedure outlined in Section 5.5.2 (see R code headings "Step 5b: Compute Estimated Predictive Accuracy"), with the exception that the test and training sets employed in this analysis were of a larger magnitude than in the empirical analysis. Subsequently, the PAP was determined by computing the proportion of generated training sets in which the predictive accuracy, represented as $p_{\text{Mal},.95}$, exceeded the threshold of 0.94 (see R code heading "Step 5b: Compute PAP").

6 Results

6.1 MSPE Evaluation

Table 1 presents a summary of the findings of the MSPE evaluation for the empirical- and simulated-based analyses. The findings indicated that MSPE_p was greater across every variable for the empirical-based analysis compared to the simulated-based analysis. In order to assess the significance of these differences, we analyzed the 95th percentile of the simulated training sets, as shown in Table 1. These revealed that the empirical-based MSPE_p exceeded the 95th percentile for the variables negative affect, worry, and suspicious; while falling below the 95th percentile for positive affect and mental unrest. In addition, the empirical-based MSPE taken across all variables was significantly greater than the simulation-based outcome. All things considered, these findings provide support for the hypothesis that the simulation-based analysis will yield overoptimistic results in the MSPE evaluation.

Table 1

MSPE Evaluation Results

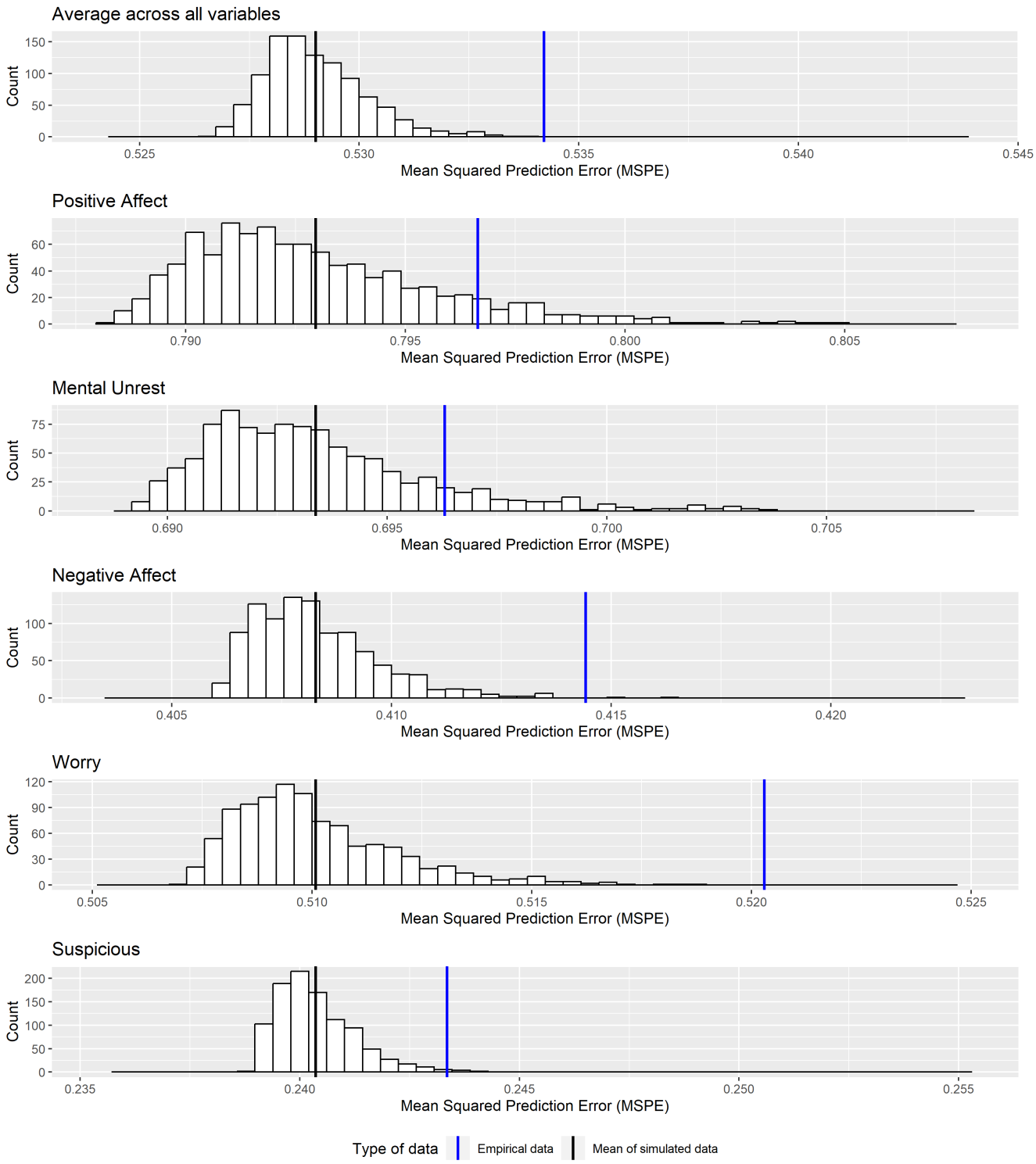
		Simulated Test data	
Variable	Empirical test data	Mean (SD)	95th Percentile
Average	.534	.529 (.001)	.531
Positive Affect	.797	.793 (.003)	.798
Mental Unrest	.696	.693 (.003)	.698
Negative Affect	.414	.408 (.001)	.411
Worry	.520	.510 (.002)	.514
Suspicious	.243	.240 (.001)	.242

Note. The simulated test data MSPE was averaged across all 1000 training sets. The first row is the average MSPE across variables. The parentheses refer to the standard deviation across the training sets.

While the absolute differences shown in Table 1 may appear small, Figure 3 provides a visual representation of how these differences can be substantial when viewed within the context of the overall distribution. As depicted in Figure 3, the average MSPE and the MSPE_p of negative affect, worry, and suspicious exhibit a relatively narrow distribution. In contrast, the distribution of positive affect and mental unrest occupy a significantly wider range.

Figure 3

Histograms of the Distribution of MSPE Across Training Sets



Note. Simulations were conducted using 1000 sets for $N_{Training}$, 100,000 observations for T_{Test}, and 966 observations for $T_{Training}$. The blue line refers to the empirical predictive accuracy obtained in Section 5.5.1.

6.2 Mahalanobis Distance Evaluation

Table 2 presents the summarized results of the evaluation of Mahalanobis distance for the empirical- and simulated-based analyses. The findings suggest that the average squared Mahalanobis distance was greater for the empirical-based analysis compared to the simulation-based analysis ($D^2_\Delta = 0.044$). Furthermore, the average estimated predictive accuracy of the simulation-based analysis was greater than the outcome for the empirical-based analysis ($p_{\text{Mal},.95\Delta} = 3.587$). Notably, this difference surpasses the standard deviation calculated across all thousand training sets by more than 8-fold. It is worth noting that all 1000 estimated predictive accuracy models $p_{\text{Mal},.95}$ surpassed the threshold of 0.94, resulting in a perfect PAP score of 100%. Accordingly, the findings provide support for the hypothesis that the simulated-based analysis will yield overoptimistic results in the Mahalanobis evaluation.

Table 2

Mahalanobis Distance Evaluation Results

Type of Data	Average D^2	Estimated Predictive Accuracy (%)	Predictive Accuracy Probability (%)
Empirical	5.069	91.304	–
Simulated	5.025 (.008)	94.891 (.041)	100

Note. For the simulated data, the D^2 and estimated predictive accuracy are represented as an average across training sets n ($n = 1, \ldots, 1000$). The parentheses refer to the standard deviation across the training sets.

Figure 4 depicts the distribution of the predictive accuracy estimates of the thousand training sets. The distribution of the predictive accuracy estimates of the simulated scores was characterized by a narrow range that exceeded the threshold of 0.94. Reducing the number of timepoints in the VAR(1) model led to a flattened distribution, characterized by lower values (see Appendix B). More specifically, the lowest sample size needed in order for the mean to exceed the threshold amounted to 150 observations. Additionally, in order to meet the sufficient PAP criterion of 0.8, approximately 200 observations were necessary.

7 Discussion

There has been a growing interest in psychological research on quantifying the dynamics of complex processes over time within individuals using VAR(1) models. To evaluate whether VAR(1) models effectively capture the dynamics of and generalize to unseen data, predictive accuracy analyses may be performed. Although predictive accuracy analyses based on simulation promise to give indications of overfitting, they are unable to detect violations of model assumptions. The present study was designed to examine the effect of model misspecification on predictive accuracy measures. In line with the hypothesis, it was found that employing simulated data in a predictive accuracy analysis of a VAR(1) model estimated from an extensive time-series dataset results in higher predictive performance compared to using empirical data. This finding was corroborated by two distinct predictive accuracy metrics. First, the MSPE_p from the empirical analysis was greater than the simulation analysis for all variables, exceeding the 95th percentile for three of the five variables. Second, the estimated predictive accuracy for the simulation far exceeded the PAP threshold, which, in turn, greatly surpassed the empirical outcome.

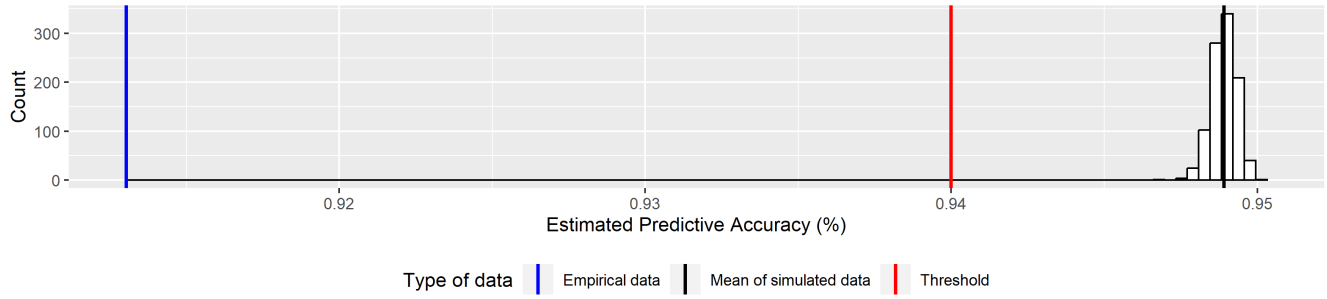
7.1 Main Results

The findings from this study highlight two important points. Firstly, intensive longitudinal psychopathological datasets have the potential to violate the stationarity assumption, leading to misspecified parameters in VAR(1) models. Although the assessment of overall stationarity in the univariate time-series yielded mixed findings, temporary departures from stationarity were found across all variables. These patterns of non-stationarity are consistent with previous studies (e.g., Albers & Bringmann, 2020; Cabrieto et al., 2018; Wichers & Groot, 2016) that identified temporal fluctuations in the dynamics of symptoms in the Peter Groot (2010) dataset. Indeed, change lies at the core of networks that capture momentary psychopathological symptoms (Bringmann et al., 2022). The considerable size of the dataset further amplifies the potential for non-stationarity, providing an extended time frame for these changes to manifest. This extended timeframe enables the identification of various shifts, such as changes in treatment effectiveness, the influence of life events, and the development of symptom expression. This finding challenges the conventional recommendation of employing a large time-series to prevent overfitting. While a larger sample size is generally desirable, our study suggests that excessively large sample sizes may negate the advantages by elevating the risk of obtaining a misspecified model due to non-stationarity. Thus, when utilizing the time-invariant VAR(1) model, it is important to strike a balance and carefully consider the optimal sample size to ensure sufficient power and prevent overfitting, while also taking precautions to avoid model misspecification[7]. Additionally, researchers should consider the type of target functions underlying the data, as simpler models like AR(1) may be more

[7]The app developed by Revol et al. (2023) may be a valuable tool for estimating the minimum number of measurement occasions necessary to achieve generalizable and robust results in VAR models with varying characteristics.

Figure 4

Histogram of the Distribution of Estimated Predictive Accuracy Across Training Sets



Note. The red line refers to the .94 threshold proposed for $p_{\text{Mal},.95}$ by Revol et al. (2023) and the blue line refers to the empirical predictive accuracy obtained in Section 5.5.2.

appropriate if the target functions are relatively simple (Bulteel, Mestdagh, et al., 2018; Lafit et al., 2022).

Secondly, the present study indicates that VAR model misspecification is not accurately captured by a simulation-based PAA. These findings are consistent with the cautious approach advocated by Revol et al. (2023), who emphasized that drawing conclusions about the predictive accuracy of empirical data solely based on simulated PAA is insufficient. The present study adds to this argument, highlighting the misleading nature of simulation-based assessments in terms of generalizability and predictive performance. In line with Revol et al. (2023), we recommend that researchers interested in evaluating the generalizability and predictive accuracy of VAR(1) models turn to cross-validation procedures[8]. The constraints of simulation-based PAA in detecting VAR assumption violations highlight the challenge of accurately reproducing the intricate nature of real-world data through current data-generating procedures. Particularly, the assumption that the multivariate normal distribution accurately represents the true distribution of innovations in empirical data may not hold. When the assumed distribution significantly deviates from the distribution of innovations in the empirical data, the simulated data may fall short in capturing the complexities and heterogeneity observed in real-world data. Accordingly, further investigation and development of simulation techniques that better emulate the characteristics of empirical data are warranted to address these concerns.

7.2 Comparison of Predictive Accuracy Metrics

While the empirical analysis yielded lower predictive accuracy performance than the simulated analysis on both metrics, the difference was more pronounced when considering the multivariate squared Mahalanobis distance. Considering the simulation-based predictive accuracy analysis served as a benchmark, where stationarity was ensured, these findings imply that the empirical-based squared Mahalanobis distance metric outperformed MSPE in identifying violations of VAR assumptions. This discrepancy could be attributed to a core theoretical difference between the two metrics alluded to in Section 4. Namely, in contrast to MSPE, the squared Mahalanobis distances are sensitive to disparities in innovations variances across variables and the influence of innovation covariance. As a result, the squared Mahalanobis distances are better able to consider the complexities and interdependencies inherent in the VAR(1) model. By considering these factors, this metric promises to be a robust alternative to MSPE in the realm of multivariate models. Therefore, we highly recommend utilizing squared Mahalanobis distances as a metric for conducting predictive accuracy analyses.

However, this conclusion may not hold for datasets with fewer timepoints. As discussed in Section 4, the distribution of the squared Mahalanobis distance values should approximate the χ^2 distribution when the training set is sufficiently large. In the present study, there were 966 timepoints available for the analyses, suggesting that the assumption of an adequately sized training set was likely satisfied. However, it is crucial to acknowledge that with a smaller dataset, this distributional assumption may be violated, potentially introducing additional bias into the estimated VAR model parameters. This suggests that the ability of the squared Mahalanobis distance to effectively identify assumption violations over MSPE may be contingent on the availability of a sufficiently large training set. It would be beneficial for future research to investigate the performance of these metrics on datasets with varying sample sizes to gain a more comprehensive understanding of their generalizability and robustness across different data contexts.

[8]The application of the LOOCV is available on the OSF page of this project.

7.3 Limitations and Future Directions

This current study is limited by several factors. The first limitation pertains to the failure to replicate the nodes used in Wichers and Groot (2016). Namely, for the affect items, the three components from Wichers and Groot (2016) were replaced by single variables, which may have affected the autoregressive and cross-lagged edges within the VAR(1) network. Collapsing variables into a smaller number of principal components or common factors results in a sparser VAR(1) model that has a reduced dimensional representation of the contemporaneous correlation structure (Ariens et al., 2020; Bulteel, Tuerlinckx, et al., 2018). In turn, VAR(1) networks based on principal components offer the potential to lower measurement error and enhance generalization to unseen data (Bulteel, Tuerlinckx, et al., 2018). This implies that the replacement of components with individual variables might have had a detrimental impact on the predictive accuracy results.

Another limitation arises from the use of listwise deletion to handle missing data. Dealing with missing data is a common challenge when working with intensive longitudinal data, and the approach we take to address this issue can significantly impact the quality of our results (Hamaker et al., 2018; Shin et al., 2009; Yuan et al., 2020). Numerous studies have demonstrated that estimation methods relying on listwise deletion cannot be generally recommended when the missing mechanism deviates from MCAR (Duncan et al., 1998; Muthén et al., 1987; Newman, 2003; Schafer & Graham, 2002; Shin et al., 2009). Such approaches can lead to a loss in efficiency, as well as biases and distortions in the estimation of variable interrelations. Even under MCAR conditions, listwise deletion may result in suboptimal efficiency (Newman, 2003; Schafer & Graham, 2002). The abundance of missing observations in the dataset suggests that the data may not conform to the assumption of MCAR, biasing and invalidating the parameter estimates. Alternative estimation methods, such as MLE and DSEM (described in Section 2) handle missing data directly, thereby eliminating the need for dedicated missing data procedures such as listwise deletion. In addition, these methods yield more accurate parameter estimates and model fit statistics (Shin et al., 2009). In particular, the Mplus implementation of DSEM holds promise due to its Bayesian estimation and Markov Chain Monte Carlo (MCMC) sampling, which capture individual-specific parameters and account for the autocorrelation structure of the data. This approach allows for simultaneous consideration of random and nonrandom parameters, potentially enhancing the estimation of VAR(1) model parameters and the predictive accuracy outcomes.

The final limitation of this study lies in the assumption that the observed predictive accuracy outcomes directly reflect the extent of model misspecification and violations of model assumptions. While this conceptual assumption holds merit, it is important to acknowledge that the methods employed in this investigation only indirectly assessed the relationship between predictive accuracy and model misspecification. Consequently, the precise connection between the violation of specific assumptions, model misspecification, and the resulting predictive accuracy results remains somewhat ambiguous. To gain a clearer understanding of this relationship, further research is warranted to explore how the violation of specific VAR assumptions leads to model misspecification and subsequently impacts predictive accuracy. For instance, it would be valuable to investigate the magnitude of bias induced in the VAR parameters by the violation of the stationarity assumption. And how the biased caused by this violation alone contributes to the predictive accuracy. Investigating the other assumptions in a similar manner and examining the interaction effects would contribute to a more comprehensive understanding of the impact of violations of VAR assumptions on predictive accuracy outcomes. The findings from such research endeavors could help in the development of more robust sample size planning approaches.

7.4. Conclusion

In conclusion, our study highlights the potential presence of non-stationarity in intensive longitudinal psychopathological datasets, resulting in biased parameters that do not accurately represent the target function. This finding holds important implications for the practical application of networks in clinical practice. Specifically, when VAR(1) models estimated from psychopathological data are misspecified, it is not meaningful to draw inferences from the model parameters and network diagrams. To mitigate the interpretation of misspecified models, the implementation of empirical-based cross-validation procedures becomes imperative in assessing the generalizability and predictive accuracy of future applications.

Acknowledgments

Conflict of Interest

The Author declares that they have no conflict of interest.

Data Availability Statement

To ensure reproducibility, the files necessary to conduct the empirical and simulation studies will be shared on the OSF platform. These files will be made accessible on the project's OSF page (https://osf.io/7pm89/) two years after the project's completion.

References

Albers, C. J., & Bringmann, L. F. (2020). Inspecting Gradual and Abrupt Changes in Emotion Dynamics With the Time-Varying Change Point Autoregressive Model. *European Journal of Psychological Assessment*, *36*(3), 492–499. https://doi.org/10.1027/1015-5759/a000589

Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psychosomatic Research*, *137*, Article 110191. https://doi.org/10.1016/j.jpsychores.2020.110191

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling*, *25*, 359–388. https://doi.org/10.1080/10705511.2017.1406803

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine*, *66*(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, *64*(9), 1089–1108. https://doi.org/10.1002/jclp.20503

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Bringmann, L. F. (2021). Person-specific networks in psychopathology: Past, present, and future. *Current Opinion in Psychology*, *41*, 59–64. https://doi.org/10.1016/J.COPSYC.2021.03.004

Bringmann, L. F., Albers, C., Bockting, C., Borsboom, D., Ceulemans, E., Cramer, A., Epskamp, S., Eronen, M. I., Hamaker, E., Kuppens, P., Lutz, W., McNally, R. J., Molenaar, P., Tio, P., Voelkle, M. C., & Wichers, M. (2022). Psychopathological networks: Theory, methods and practice. *Behaviour Research and Therapy*, *149*, Article 104011. https://doi.org/10.1016/j.brat.2021.104011

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., Tuerlinckx, F., & Kuppens, P. (2016). Assessing Temporal Emotion Dynamics Using Networks. *Assessment*, *23*(4), 425–435. https://doi.org/10.1177/1073191116645909

Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). VAR(1) based models do not always outpredict AR(1) models in typical psychological applications. *Psychological Methods*, *23*(4), 740–756. https://doi.org/10.1037/met0000178

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2018). Improved Insight into and Prediction of Network Dynamics by Combining VAR and Dimension Reduction. *Multivariate Behavioral Research*, *53*(6), 853–875. https://doi.org/10.1080/00273171.2018.1516540

Cabrieto, J., Tuerlinckx, F., Kuppens, P., Hunyadi, B., & Ceulemans, E. (2018). Testing for the Presence of Correlation Changes in a Multivariate Time Series: A Permutation Based Approach. *Scientific Reports*, *8*(1), Article 769. https://doi.org/10.1038/s41598-017-19067-2

Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist*, *45*(1), 16–28. https://doi.org/10.1037/0003-066X.45.1.16

Chatfield, C. (1980). *The Analysis of Time Series: An Introduction*. Chapman and Hall. http://journal.um-surabaya.ac.id/index.php/JKM/article/view/2203

Cragg, J. G. (1968). Some Effects of Incorrect Specification on the Small-Sample Properties of Several Simultaneous-Equation Estimators. *International Economic Review*, *9*(1), 63–86. https://doi.org/10.2307/2525614

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, *33*(2-3), 137–150. https://doi.org/10.1017/S0140525X09991567

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, *74*, 427–431. https://doi.org/10.1080/01621459.1979.10482531

Duncan, T. E., Duncan, S. C., & Li, F. (1998). A comparison of model- and multiple imputation-based approaches to longitudinal analyses with partial missingness. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 1–21. https://doi.org/10.1080/10705519809540086

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18.

Groot, P. C. (2010). Patients can diagnose too: How continuous self-assessment aids diagnosis of, and recovery from, depression. *Journal of Mental Health (Abingdon, England)*, *19*(4), 352–362. https://doi.org/10.3109/09638237.2010.494188

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the Frontiers of Modeling Intensive Longitudinal Data: Dynamic Structural Equation Models for the Affective Measure-

ments from the COGITO Study. *Multivariate Behavioral Research*, *53*(6), 820–841. https://doi.org/10.1080/00273171.2018.1446819

Hamaker, E. L., & Dolan, C. V. (2009). Idiographic Data Analysis: Quantitative Methods—From Simple to Advanced. In J. Valsiner, P. C. M. Molenaar, M. C. Lyra, & N. Chaudhary (Eds.), *Dynamic Process Methodology in the Social and Developmental Sciences* (pp. 191–216). Springer US. https://doi.org/10.1007/978-0-387-95922-1_9

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press. Retrieved May 26, 2023, from https://press.princeton.edu/books/ebook/9780691218632/time-series-analysis

Haslbeck, J. M. B., Bringmann, L. F., & Waldorp, L. J. (2021). A Tutorial on Estimating Time-Varying Vector Autoregressive Models. *Multivariate Behavioral Research*, *56*(1), 120–149. https://doi.org/10.1080/00273171.2020.1743630

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Elements of Statistical Learning: Data mining, inference, and prediction* (Second). Springer. Retrieved June 15, 2023, from https://hastie.su.domains/ElemStatLearn/

Hayes, S. C., Wilson, K. G., Gifford, E. V., & Follette, V. M. (1996). Experiential Avoidance and Behavioral Disorders: A Functional Dimensional Approach to Diagnosis and Treatment. *Journal of Consulting and Clinical Psychology*, *64*(6), 1152–1168.

Kaplan, D. (1988). The Impact of Specification Error on the Estimation, Testing, and Improvement of Structural Equation Models. *Multivariate Behavioral Research*, *23*(1), 69–86. https://doi.org/10.1207/s15327906mbr2301_4

Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, *15*(1), 5–12. https://doi.org/10.1002/wps.20292

Kossakowski, J. J., Groot, P. C., Haslbeck, J. M. B., Borsboom, D., & Wichers, M. (2017). Data from 'Critical Slowing Down as a Personalized Early Warning Signal for Depression'. *Journal of Open Psychology Data*, *5*, 1–3. https://doi.org/10.5334/jopd.29

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Lafit, G., Meers, K., & Ceulemans, E. (2022). A Systematic Study into the Factors that Affect the Predictive Accuracy of Multilevel VAR(1) Models. *Psychometrika*, *87*(2), 432–476. https://doi.org/10.1007/S11336-021-09803-Z/FIGURES/12

Larson, R., & Csikszentmihalyi, M. (2014). The Experience Sampling Method. In M. Csikszentmihalyi (Ed.), *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi* (pp. 21–34). Springer Netherlands. https://doi.org/10.1007/978-94-017-9088-8_2

Loossens, T., Dejonckheere, E., Tuerlinckx, F., & Verdonck, S. (2021). Informing VAR(1) with qualitative dynamical features improves predictive accuracy. *Psychological Methods*, *26*, 635–659. https://doi.org/10.1037/met0000401

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer. https://doi.org/10.1007/978-3-540-27752-1

Meehl, P. E. (1972). Specific genetic etiology, psychodynamics, and therapeutic nihilism. *International Journal of Mental Health*, *1*(1-2), 10–27. https://doi.org/10.1080/00207411.1972.11448562

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*(3), 431–462. https://doi.org/10.1007/BF02294365

Newman, D. A. (2003). Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. *Organizational Research Methods*, *6*(3), 328–362. https://doi.org/10.1177/1094428103254673

Pfaff, B. (2008). *Analysis of integrated and cointegrated time series with R* (2nd ed.). Springer. https://www.pfaffikus.de

R Core Team. (2022). *R: A language and environment for statistical computing*. manual. Vienna, Austria, R Foundation for Statistical Computing. https://www.R-project.org/

Revol, J., Lafit, G., & Ceulemans, E. (2023). A new sample size planning approach for the (V)AR(1) model: Predictive Accuracy Analysis. https://doi.org/10.31234/osf.io/2geh4

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schleim, S. (2022). Why mental disorders are brain disorders. And why they are not: ADHD and the challenges of heterogeneity and reification. *Frontiers in Psychiatry*, *13*. https://doi.org/10.3389/fpsyt.2022.943049

Shin, T., Davison, M. L., & Long, J. D. (2009). Effects of Missing Data Methods in Structural Equation Modeling With Nonnormal Longitudinal Data. *Structural Equation Modeling: A Multidisciplinary Jour-*

nal, *16*(1), 70–98. https://doi.org/10.1080/10705510802569918

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Turkheimer, E. (1998). Heritability and biological explanation. *Psychological Review*, *105*(4), 782–791. https://doi.org/10.1037/0033-295X.105.4.782-791

van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E. H., Viechtbauer, W., Giltay, E. J., Aggen, S. H., Derom, C., Jacobs, N., Kendler, K. S., van der Maas, H. L. J., Neale, M. C., Peeters, F., Thiery, E., Zachar, P., & Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(1), 87–92. https://doi.org/10.1073/pnas.1312114110

von Klipstein, L., Riese, H., van der Veen, D. C., Servaas, M. N., & Schoevers, R. A. (2020). Using person-specific networks in psychotherapy: Challenges, limitations, and how we could use them anyway. *BMC Medicine*, *18*(1), Article 345. https://doi.org/10.1186/s12916-020-01818-0

Wassing, R., Benjamins, J. S., Talamini, L. M., Schalkwijk, F., & Van Someren, E. J. W. (2019). Overnight worsening of emotional distress indicates maladaptive sleep in insomnia. *Sleep*, *42*(4), 1–8. https://doi.org/10.1093/sleep/zsy268

Wichers, M., & Groot, P. C. (2016). Critical Slowing Down as a Personalized Early Warning Signal for Depression. *Psychotherapy and Psychosomatics*, *85*(2), 114–116. https://doi.org/10.1159/000441458

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yuan, C., Hedeker, D., Mermelstein, R., & Xie, H. (2020). A tractable method to account for high-dimensional nonignorable missing data in intensive longitudinal data. *Statistics in Medicine*, *39*(20), 2589–2605. https://doi.org/10.1002/sim.8560

Appendix A
Stationarity Testing

Table A1

Augmented Dickey-Fuller Test Unit Root Test Results

Variable	ADF Test-statistic
Positive affect	-2.652*
Mental unrest	-4.910*
Negative affect	-13.864*
Worry	-6.346*
Suspicious	-4.477*

Note. The test was conducted with the ur.df function of the urca library (Pfaff, 2008) with the type set to 'none' and the number of lags set to 1. At a significance level of 0.05, the critical value was determined to be -1.95.
*$p < .05$.

Table A2

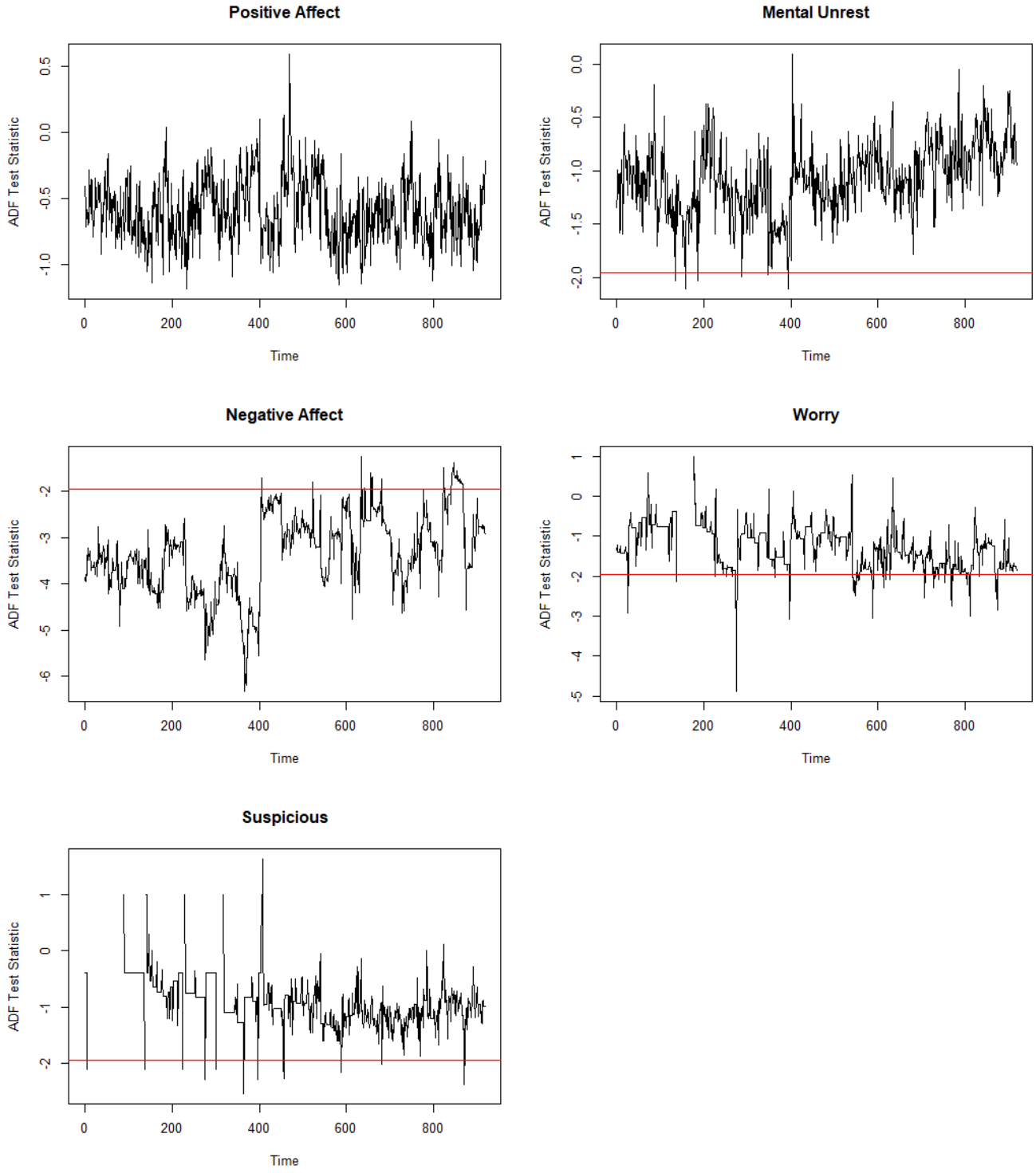
Kwiatkowski-Phillips-Schmidt-Shin Test Results

Variable	KPSS Test-statistic
Positive affect	.255
Mental unrest	5.686*
Negative affect	3.437*
Worry	4.369*
Suspicious	10.388*

Note. The test was conducted with the ur.kpss function of the urca library (Pfaff, 2008) with the type set to 'mu' and the number of lags set to 1. At a significance level of 0.05, the critical value was determined to be 0.574.
*$p < .05$.

Figure A1

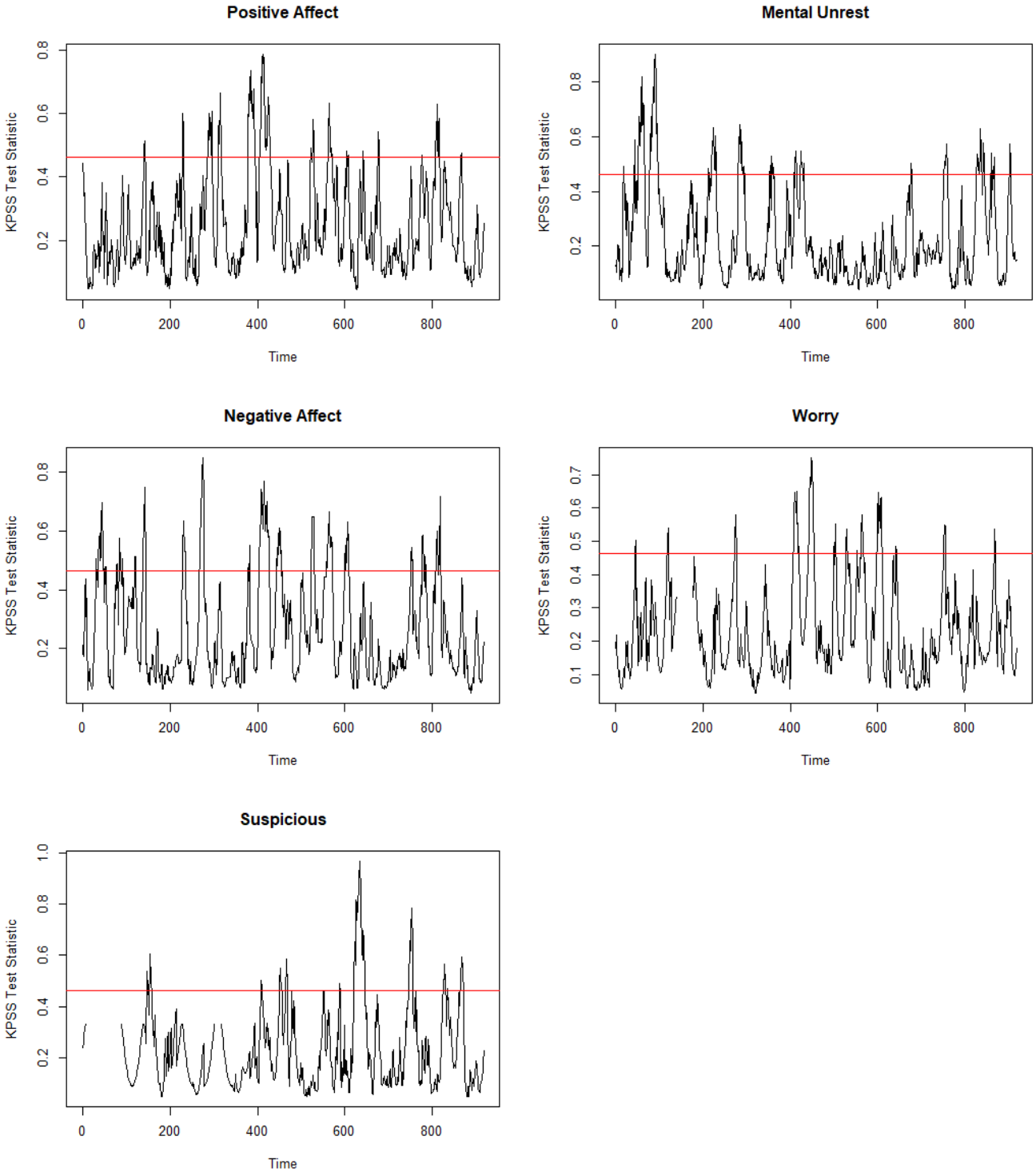
The ADF Unit Root Test-Statistics Over Time With Rolling Windows of 50 Timepoints



Note. The ADF test was performed for each variable with a rolling window of 50 timepoints. The critical value at a .05 significance level is represented by the red horizontal line.

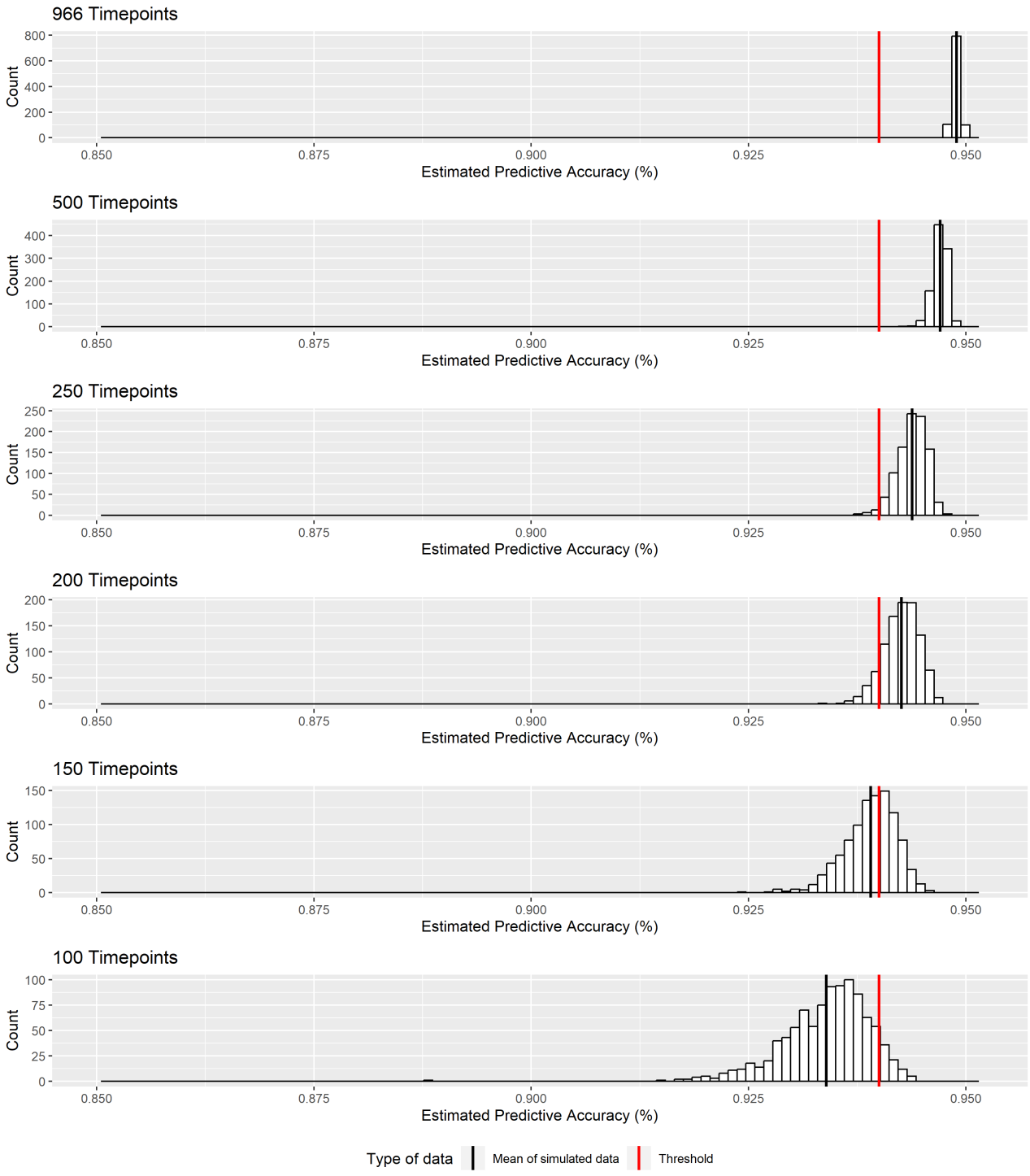
Figure A2

The KPSS Unit Root Test-Statistics Over Time With Rolling Windows of 50 Timepoints



Note. The KPSS test was performed for each variable with a rolling window of 50 timepoints. The critical value at a .05 significance level is represented by the red horizontal line.

Appendix B
Relationship between Training Set Size and Estimated Predictive Accuracy



Note. The red line refers to the threshold of .94 that Revol et al. (2023) proposed for the $p_{\text{Mal},.95}$.