Author/s:
van Berkel, N;Dennis, S;Zyphur, M;Li, J;Heathcote, A;Kostakos, V

Title:
Modeling interaction as a complex system

Date:
2021

Citation:
van Berkel, N., Dennis, S., Zyphur, M., Li, J., Heathcote, A. & Kostakos, V. (2021). Modeling interaction as a complex system. Human-Computer Interaction, 36 (4), pp.279-305. https://doi.org/10.1080/07370024.2020.1715221.

Persistent Link:
https://hdl.handle.net/11343/247884

Modelling Interaction as a Complex System

Abstract

Researchers in Human-Computer Interaction typically rely on experiments to assess the causal effects of experimental conditions on variables of interest. Although this classic approach can be very useful, it offers little help in tackling questions of causality in the kind of data that are increasingly common in HCI—capturing user behaviour 'in the wild'. To analyse such data, model-based regressions such as cross-lagged panel models or vector autoregressions can be used, but these require parametric assumptions about the structural form of effects among the variables. To overcome some of the limitations associated with experiments and model-based regressions, we adopt and extend 'empirical dynamic modelling' methods from ecology that lend themselves to conceptualizing users' behaviour as a complex nonlinear dynamical system. Extending a method known as 'convergent cross mapping', we show how to make causal inferences that do not rely on experimental manipulations or model-based regressions and, by virtue of being non-parametric, can accommodate data emanating from complex nonlinear dynamical systems. By using this extended approach, which we call 'multiple convergent cross mapping' or MCCM, researchers can achieve a better understanding of the interactions between users and technology – by distinguishing causality from correlation – in a wide variety of real-world settings.

1 INTRODUCTION

Human-Computer Interaction (HCI) research seeks to understand the causal interactions between users and technology, ultimately leading to the design of improved interactive technology. To study causal interactions, HCI researchers typically adopt one of two dominant approaches. The first is experimental in nature, wherein researchers introduce participants to two or more conditions and compare their effects on a variable of interest. The second relies on observational data and takes a model-based approach to estimating causal effects with regressions. For example, researchers may collect contextual data and model the effect of these variables on device usage. Although both of these approaches can be very useful, they have limitations. On the one hand, controlled experiments offer a simplified version of reality, which limits the generalisability of results in real-world settings. On the other hand, parametric modelling approaches rely on pre-determined equations that can be understood as hypotheses or assumptions about the structural relationships that define a system being modelled. Such parametric forms may not be suitable for modelling a variety of complex systems whose functioning is not known or cannot be known a priori [37].

To complement experimental and model-based regression approaches when studying Interaction, we propose an 'empirical dynamic modelling' (EDM) method drawn from ecology and applied physics (see [5,10,41,46]). EDM is a set of methods designed to characterise and test causality in complex dynamic systems, such as those associated with humans interacting with technology *over time*. By 'system' in the term 'complex system' we refer to the dynamic interaction among humans and technology rather than the more typical use of the term 'system' to refer to the technology itself. Our approach extends EDM techniques widely used in ecology, as in the following example:

> Consider an ecological system comprised of wolves and sheep. Over time, the number of wolves directly affects the number of sheep (since wolves eat sheep). At the same time, the number of sheep affects the number of wolves (fewer sheep means not enough food for wolves, and hence wolves die).

In such a scenario, we would say that the number of wolves and sheep affect each other simultaneously and over time, but it is not safe to assume that the system will tend towards any simple kind of equilibrium or even stable rates of change over time. Therefore, attempting to use typical correlational or experimental methods (*e.g.*, ANOVA) to understand this relationship is inappropriate because they are ill suited to modelling relationships that are bidirectional, simultaneous, and nonlinear. The techniques we draw on are designed to interrogate such complex systems, and untangle the effects that may be present.

Why is the relationship between wolves and sheep relevant to HCI? Our discipline investigates the interactions among humans and technology, which we propose can be considered a complex dynamic system that involves nonlinear patterns of activity and potentially complex causal effects among users and software/hardware. Users have certain goals that they try to achieve by using an interface. Doing so changes the state of the interface, which in turn has an effect on the user, which in turn triggers changes to how

they use the interface, and so on. There is rich literature conceptually describing this relationship, such as Jack Carroll's "*Task-Artefact Cycle*" [3], with associated methods to understand the design requirements when creating new and evolving technology; and Don Norman's "*Gulfs of execution and evaluation"* [28], which highlights some of the cognitive challenges that the user-interface relationship imposes on users and designers. This line of thought is consistent with the long-standing recognition that understanding interaction among users and technology requires attending to the ways that technology impacts or 'conditions' users and vice versa (e.g., [15,16,29]). To describe this dynamic process, we refer to the ongoing interaction among humans and technology (e.g., hardware/software) as a 'system'.

Despite the literature recognizing, or at least suggesting, that interaction is a complex system of users and technology, most HCI research relies on the 'gold standard' of controlled experiments [22]. These experiments typically compare Artefact A vs. Artefact B in terms of human performance, error, or preference, and by strictly manipulating the different variables between A and B provide insights on how those variables affect users. This approach has great merit for making 'static' design decisions: comparing two feedback sounds in terms of human understanding; two input techniques in terms of human error; or two colour schemes in terms of human preference. By definition, such studies assume a unidirectional effect: the artefact affects the user. Because the manipulated (*i.e.*, control) variable is an artefact characteristic, the analysis assumes that changes to that variable must affect users. Even when control variables are user characteristics (*e.g.*, gender or age), the outcome variables are typically human behaviour or performance, therefore maintaining the unidirectional effect assumption. However, simply because manipulations and measured outcome variables are restricted to specific factors does *not* mean that causality is unidirectional. Instead, it implies a potential shortcoming of experimental methods wherein researchers assume unidirectional causality and incorporate this assumption into a study through its design, thus offering a partial picture of a potentially complex system.

Alternatively, 'dynamic' longitudinal studies (such as 'in-the-wild' studies) provide a more realistic setting for observations, often going beyond more simplistic and unidirectional approaches to understanding the complex relationship between people and technology. Researchers who try to make sense of longitudinal in-the-wild data typically approach analysis with two broad strategies. One relies on regression to estimate the effects of multiple variables on an outcome of interest. A second approach is a field experiment with efforts to randomize conditions and analyses that test differences between them. Unfortunately, both approaches may fail to characterise the complex dynamic relationship among users and technology—neither regression models nor field studies and their analyses can entirely capture the complexity of a complex system and the potentially bidirectional effects involved in real-world *interactions* among users and technology. On the one hand, regression models would need to be parameterized properly to estimate effects of interest, but if a system is complex it will be unreasonable to expect researchers to know these parameters a priori. On the other hand, field experiments often suffer from the same problems as typical lab experiments mentioned previously.

To overcome some of the limitations of experiments and typical model-based approaches, in what follows we present a novel method for analysing longitudinal human performance and artefact state. As the system defining their interaction changes over time

(*i.e.*, as the states of the system change) in potentially nonlinear and causal ways, our proposed approach is able to characterize the system and test causality in it—ideally suited for highly granular (*i.e.*, many observations over a period of time) datasets. Our method extends a technique known as convergent cross mapping (CCM) that distinguishes causation from correlation, which was published recently in *Science* [37]. This technique is a core component of the EDM approach that we describe and exemplify. Our paper makes a number of contributions:

- First, as a tutorial, we tease apart the nuances of modelling interactions in a complex system, and describe how complex dynamic system methods can be used within HCI. In what follows, we do so by initially elaborating on our points about more traditional HCI methods, experimental and observational studies, and then proceed to discuss EDM and its logic.
- Second, methodologically, we develop a novel way of combining, visualising, and evaluating the EDM results for multiple individuals in a sample. This has been a limitation of existing methods, which typically focus on a single entity (*e.g.*, a single ecosystem), or treat multiple entities as if they were homogenously defined by a single dynamic system (e.g., [6]). We call our alternative approach Multiple Convergent Cross Mapping (MCCM), which treats each entity as a potentially unique dynamic system.
- Third, as a case study, we apply our method in the context of HCI by analysing user interactions on mobile devices through multiple datasets, demonstrating how our analysis enables researchers to establish the causal direction and distance between two variables of interest. As part of exemplifying our method, we answer questions (amongst others) such as: do people tend to use their phones because they receive notifications, or do they receive notifications because they use their phones, or both (*i.e.*, bidirectional causality)?
- Fourth, we conclude with thoughts on wider applications of EDM for the HCI community, as well as advances to EDM that will aid HCI researcher's in their question to better understand and model causal effects in the complex dynamic systems that define humans and their interaction with technology.

2 RELATED WORK

2.1 Correlation vs. Causation in HCI

In their book 'Research Methods in Human-Computer Interaction', Lazar *et al*. highlight that "*one of the most common objectives for HCI-related studies is to identify relationships between various factors*" [22]. To do this, HCI studies of human participants are either experimental or observational in nature. These approaches typically have the following characteristics:

- *Experimental study:* the researcher intervenes in the reality of participants (e.g., by introducing study conditions) and measures the effect of these interventions (Fig. 1A). Such studies are often, but not necessarily, conducted in a laboratory environment, which allows control over potential confounds. However, in such studies the lack of a real-world context can reduce the generalisability of study results [11]. Additionally,

one key element of control that underpins the ability of experimental approaches to determine causality is random and/or counterbalanced conditions for participants, which can bias or interact with the relationships among causal variables.

- *Observational study:* the researcher does not intervene in the reality of the participants, but instead attempts to understand the interplay between the artefact and user, typically by observing variables of interest over time (Fig. 1B). These studies are usually conducted in-the-wild rather than in a lab, and allow the researcher to observe the user, as well as their interaction with a potential artefact, in their natural environment. As the researcher does not intervene in the reality of the participants, no conditions (*i.e.*, manipulations) are introduced to participants. In such cases, it is important for the researcher to be able to distinguish between naturally occurring correlational versus causal relationships among variables.

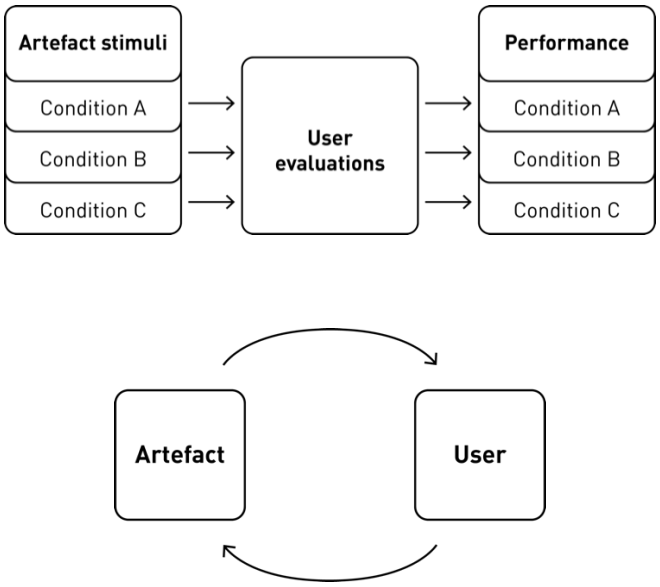

Fig. 1. Illustration of two commonly used research approaches in HCI.
A. Experimental study, in which the effect of different conditions is assessed.
B. Observational study, in which the interplay between artefact and user is more often assessed.

The primary purpose of experiments and observational studies is to help researchers estimate and infer causal effects. Consistent with the well-known dictum that '*correlation does not imply causation*', the problem is that many observed associations among variables cannot simply be understood as causal. In our example, researchers may find a relationship between mobile notifications and smartphone usage, but then not know if both variables have common causes, one causes the other, or bidirectional causality exists.

Untangling correlation and causation is a topic that has recently received increased attention across multiple disciplines, including HCI. To address it, researchers have taken various experimental and observational approaches, such as Tsapeli *et al*. [39] and Mehrotra *et al*. [26] in their analysis of the relationship between smartphone interaction and the emotional state of the user. Their findings indicate that the emotions of the smartphone user have a causal impact on different aspects of smartphone interaction. In

their work, Mehrotra *et al*. [26] first perform a correlation analysis to determine which variables have a significant relationship. For example, one of the tested correlations is the user's self-reported stress level and number of received notifications. Following this, the variables which are significantly correlated ($\alpha < 0.05$) are tested for causality using a matching design framework as introduced in [23]. In this approach, a pair of two variables is tested for causality and an 'average treatment effect' is calculated by taking into account a set of pre-selected confounding variables [40]. The average treatment effect indicates the direction of the causation and includes an indicator of significance.

Such approaches rely on some form of regression to attempt to mimic an experiment. For example, instances where users exhibit behaviour A can be matched to instances where users exhibited behaviour B under the same circumstances. Although the goal of establishment causality in this way is important, such methods are problematic when dealing with complex systems for at least two reasons: 1) when variables are deterministically related in a system, then controlling for any one variable can eliminate important aspects of overall system dynamics; and 2) assuming how variables are related to each other is required *a priori* to construct a parametric model that will have the assumptions embedded in it, such as when researchers first check for significant bivariate correlations, which may be zero even when causal relationships exist that are nonlinear in nature [37].

Of course, researchers are well aware that two variables may be correlated but lack a causal relationship. However, researchers are often less aware of the inverse fact: two variables may be *uncorrelated* but can still be causally related because they are a function of complex system dynamics that are nonlinear. This fact undermines both experimental and observational approaches, and thus even observational methods that are designed to mimic an experiment are not necessarily well suited for studying complex dynamic systems, which we now discuss.

2.2 Complex dynamic systems and EDM

Complex dynamic systems consist of multiple interacting components that produce inherently unstable and nonlinear behaviour as a system evolves over time, such as the interaction between users and technology evolving dynamically over time. This evolution occurs in state-dependent ways, meaning the way a system functions depends on its current and historical states, such as a user's next actions or a technology's next alerts depending on the recent past. Complex dynamic systems can be found all around us, and as such the idea of dynamic systems has been applied to a wide variety of disciplines and application areas which can be described using a small number of variables/dimensions [21], including the spread of diseases [14], ecological diversity [2], financial markets [24], and human development [12]. Dynamic systems evolve based on the interaction of the components in the system, with the goal of the researcher often being to predict or forecast the next state of the system based on recent states. Given the complexity and non-linearity of dynamic systems, the use of linear statistical methods (common in experimental and observational designs) is not suitable: "*Linear approaches are fundamentally based on correlation. Thus, they are ill-posed for dynamical systems, where correlation can occur without causation, and causation may also occur in the absence of correlation*" [5].

Empirical dynamic modelling or EDM is a non-linear approach to studying dynamic systems based on Takens's theorem [38], which describes how the behaviour of a multi-

variable complex system can be reconstructed based on a time series of a single variable, as can be seen in the video[1] included in [37]. The crucial insight in Takens's theorem is that all the richness and diverse behaviour of a complex system can be reconstructed by analysing any **single** variable that is associated with the system. This theorem is important for HCI studies, because even though there may be confounding variables that a study has not captured, Takens's theorem suggests that those confounding variables nevertheless leave an imprint on the variables that are measured. As such, we can reconstruct the behaviour of an entire system by capturing just a single variable. Thus, traditional concerns about confounds are fundamentally altered and potentially alleviated because the traditional approach of attempting to control for relevant system variables can make it impossible to reconstruct the dynamical behaviour of a system—controlling for any variables relevant to a system can eliminate important system dynamics—thus making the results from typical regression models potentially suspect.

Based on Takens's theorem, EDM utilises a time series to reconstruct the behaviour of a dynamic system and operates with minimal assumptions about the exact nature of a dynamic system, making it "*suitable for studying systems that exhibit non-equilibrium dynamics and nonlinear state-dependent behavior*" [47]. This is done in a three-step process:

1) using a method known as simplex projection, the dimensionality E of a dynamical system is assessed, and we continue to the next step if E is sufficiently low (<15 in our case);

2) using a method known as S-mapping, we assess whether the system evolves nonlinearly (i.e., in state-dependent ways); and then

3) based on results from the previous analyses, convergent cross mapping or CCM is used to assess causal relationships among variables that define a dynamical system (see [37]). For example, variables in dynamic systems can display a positive correlation at some times, while displaying no correlation or even a negative correlation at other times [37] – a phenomenon known as 'mirage correlation'. Thus, linear analysis methods may misconstrue and fail to uncover a large number of nonlinear behaviour and causal effects due to the nonlinearity of variables in dynamic systems. CCM overcomes this by assessing causal effects without linear assumptions.

In contrast to predictions based on a predefined set of equations as in typical regression models, "*EDM [...] relies on time series data to reveal the dynamic relationships among variables as they occur*" [46]. This dynamic relationship among variables, wherein correlations depend on the state of a system, is a typical aspect of complex nonlinear systems. EDM was originally applied to ecosystem forecasting [46], where it outperformed traditional modelling methods. EDM is now being applied in a wide range of disciplines, including finance, neuroscience, and genetics [31]. These fields typically produce a large amount of longitudinal data, and their data entail causal effects between variables. As such, we argue that EDM and CCM can be helpful in analysing human-technology interaction data in HCI studies, especially under conditions that share these characteristics of dynamic systems.

Specifically, inspired by Jack Carroll's [3], Don Norman's [28], and William Gaver's [15] work, we argue that human-technology interaction bears the hallmarks of an evolving complex dynamic system. Technology use is often non-linear and episodic, as shown by a

[1] https://www.youtube.com/watch?v=6i57udsPKms

wide range of studies on technology interaction (*e.g.*, learning effect, technological adoption). Furthermore, our relationship with technology is bidirectional (e.g., a person's interest in social media causally drives smartphone use, and increased smartphone use can lead to increased time spent on social media). Finally, our interaction with technology is driven by many factors (*e.g.*, friends, weather, trends), which can dynamically change both by themselves and in relation to one another in nonlinear ways. By definition, it is impossible to account for all of these (confounding) factors in typical linear models. As such, we propose to model the interaction between humans and technology as a dynamic system in order to gain further insights into how such systems function, including causal effects among the variables that define them.

Although a limited number of researchers have attempted to separate correlation and causality in the domain of HCI, their approaches typically concern linear systems which are unable to account for the complex patterns of technology use. Although EDM has great potential for fields such as HCI that involve the study of complex dynamic systems, to our knowledge no paper has described the use of EDM in HCI literature.

2.3 Our methodological contribution

Previous work has been highly successful in applying the concept of complex systems to study singular entities (*e.g.*, an individual rainforest or a single stock market). Although these analyses have traditionally focused on ecosystems [37,46], recent work has also begun to investigate individual human behaviour. All aforementioned studies, however, are limited to the analysis of an individual system. This works well in the analysis of ecosystems, where a researcher tries to understand the dynamics of a single ecosystem, or a single financial market. But as we show here, this approach may not transfer well to the domain of HCI, where we argue that each user needs to be treated as an independent ecosystem.

For example, while different groups of shoaling fish can ultimately be considered as functioning with common rules that define a common dynamic system, the same may not apply to a group of study participants, each of which may function according to unique rules that define a unique dynamic system. Participants in a typical HCI study are independent from one another, they utilise personal devices in potentially unique ways rather than sharing one interactive artefact in a shared way, and they spend most of their time in different contexts. Furthermore, consistent with the notion of a complex dynamic system, we expect that even with very similar starting conditions (e.g., a new smart-phone with a single set of default settings), differences will emerge over time between participants in the ways that they interact with technology. Therefore, analysing the data from multiple participants as if they originate from a single ecosystem may mask the uniqueness of interactions that define each individual and their technology. In sum, EDM can be used to analyse data from one entity or multiple entities (see *e.g.* [6]) but heretofore this has been done by treating the behaviour of the entities as being part of a single dynamic system. In what follows, we extend EDM to the case of multiple users who may not share a common dynamic system, and we develop a meaningful way to summarise and validate the multiple independent analyses associated with this case. We call this approach multiple convergent cross mapping or MCCM, which estimates unique causal effects for each system/participant sampled over time. In what follows, we describe the logic of EDM, CCM, and then MCCM through a real-world illustration.

3 METHOD AND RESULTS

3.1 Datasets

We exemplify our method by applying it to multiple independent datasets—each with multiple users and, thus, multiple unique dynamic systems. Our purpose is to identify and characterise relationships for a range of variables associated with mobile device use, specifically by: 1) characterising the dynamical system associated with each user's data using simplex projection and S-mapping as noted previously, and then; 2) with results from this step use CCM to assess causal effects among variables of interest.

Dataset 1 consists of smartphone use traces from 20 participants collected during a 3-week in-the-wild study [43]. Participants were recruited from a university campus using mailing lists and had a diverse educational background. A mobile application was installed on participants' phones, and ran continuously in the devices' backgrounds. Participants used their personal phones in order to ensure realistic usage behaviour. During the study, participants were asked to complete up to six questionnaires per day using the Experience Sampling Method (ESM). The application collected, *inter alia*, device ID, phone usage, battery level, and application usage data. The data were cleaned by removing applications initiated by the operating system (*e.g.*, application launcher, keyboard). Following this, the dataset contained over 78,500 application usage events, over 137,000 notification events, and close to 3 million battery events (*i.e.*, changes in battery level or charging status).

Dataset 2 is an extension of the dataset of the study reported in [44]. It consists of smartwatch use traces of 589 smartwatch users, collected between January 2016 and February 2017. 67.9% of the users (N = 400) had the application installed and logging for less than 30 days (M = 7.49, median = 5), 17.7% of the users (N = 104) for a timespan between 30 and 90 days (M = 54.66, median = 52), and 14.4% of the users (N = 85) for more than 90 days (M = 178.36, median = 148). Data collected by the application which are of relevance here are device ID, smartwatch screen events (turned on, turned off), and notification information (time and application). The total dataset consists of 6.1 million notifications and 2.0 million screen usage events.

Dataset 3 contains data from a laboratory experiment [33], whereby participants' finger temperature was recorded while using a smartphone. The sample contains 24 participants, each of whom spent approximately 90 minutes completing tasks on a smartphone. Two of the experimental conditions took place in a cold chamber with a temperature of -10 °C, whereas the remaining two conditions took place in room temperature. During the study, the thumb and index finger temperature of the participants was recorded continuously. In addition, the temperature of the smartphone battery was collected continuously.

For datasets 1 and 2 we calculate an hourly metric per measurement variable for each participant, and for battery data we calculate the average battery percentage per hour. Phone usage is calculated as the number of times the phone was turned on per hour. For the remaining variables (application usage and incoming notifications) we count the total number of events per hour. For Dataset 3, we consider each experimental task as the unit of analysis. For each task (which lasted a few seconds) we calculate both the average temperature of the participant's active finger (thumb or index depending on how the phone was held) and the average battery temperature during that period. The participants from all three datasets are unique.

3.2 Method

To conduct analyses we use the R package 'rEMD' by Ye *et al*. [47]. We now describe the six steps of our method, adapted from [47], including the process of data wrangling, MCCM, and a final robustness check. We develop this process in order to highlight the differences/similarities between participants in a study We apply this method to the aforementioned datasets in the subsequent section.

1. **Data treatment**. EDM requires data in a typical time-series or panel data format (i.e., a 'long' format where each occasion of measurement is a row and variables are columns; see Table 1). For Datasets 1 and 2 we formed a time series consisting of 24 hourly entries per day. For Dataset 3 the time series consists of the experimental tasks, each of which lasted a few seconds. If the variable of interest did not occur in a specific time period (*e.g.*, a participant did not receive a notification during a given hour), we assigned a value of 0 for that time period. If a participant has insufficient data available for analysis (*i.e.*, limited number of rows), the participant was completely discarded. In our case we discard participants with less than 10 data points. Using this cut-off point, we discard zero participants from Dataset 1, 30 participants from Dataset 2, and zero participants from Dataset 3. The minimum number of data points is dependent on the study design and research questions.

Table 1. Example slice from collected data

Row #	Participant ID	Date	Hour	Notifications
171	4939097448	01-01-2018	14	12
172	4939097448	01-01-2018	15	0
173	4939097448	01-01-2018	16	8

2. **Identify optimum value for E (Embedding Dimension)**. In this step, we identify the optimum embedding dimension (E) using simplex projection, as recommended for rEDM [47]. In this step, the method uses time delay embedding on a single variable to generate a complex system reconstruction, and then applies the simplex projection algorithm to make forecasts. In brief, consistent with Takens's theorem the idea is to use a set of E lagged values of a variable to reconstruct the behaviour of a dynamic system in E-space. Each point in E-space is formed using E lags of a variable, and these points form an 'attractor' or an 'attractor manifold' that defines system evolution (e.g., a classic example of the Lorenz or 'butterfly' attractor). Then, for a given point on the manifold, the quality of the reconstruction is evaluated by finding the E+1 'nearest neighbour' points on the manifold, and then projecting these neighbours into the future to make predictions. The optimum value of E provides the best out-of-sample predictions of the future, implying that an underlying dynamical system has been optimally reconstructed. This forecast ability is calculated as the correlation between the observed and predicted values – we annotate this value as 'rho' ρ. In our case, this E value is used to further analyse each variable for each participant. For different values of E we plot the forecast skill as the correlation ρ among predicted and observed future values in a hold-out subsample (Fig. 2A & 2B). We select the value of E that maximises this correlation. Furthermore, the functional form of the E-ρ relationship is useful for diagnosing the nature of a system. Low-dimensional deterministic systems with low noise will have ρ maximized at a large value (i.e., close 1) when E is small (in our case less than 15). Alternatively, high-dimensional deterministic systems or stochastic systems with autocorrelation will typically show ρ increasing with E, and potentially stabilizing rather than falling at very large E [36].
3. **Test for nonlinearity**. CCM is a nonlinear analysis technique, and it is therefore useful to check whether a system evolves in a nonlinear fashion rather than merely being defined by linear autocorrelated noise. rEMD uses S-maps [36] to distinguish between Brownian noise (also

known as 'red noise') and nonlinear deterministic behaviour [47]. In brief, this is done by using the E chosen from the previous step of simplex projection, and then estimating a linear map that uses the E-dimensional points on a manifold's surface to predict the future. As done for rEDM, we define 'theta' as the S-map tuning parameter which adjusts the sensitivity to nearby versus distant points for the mapping. When theta = 0, all points on the manifold are equally weighted and therefore the map reduces to a kind of autoregression, but when theta > 0 the map is more sensitive to nearby points and thus the mapping is more local and, therefore, state-dependent. As can been seen from Fig. 2C & 2D, if in the produced graph the forecast ability is greatest when theta = 0, then this means the data can be modelled by an autoregressive model. If the prediction is greatest when theta > 0, then more local information is more useful for prediction of the future, implying state-dependent system evolution and therefore a nonlinear process. We explicitly label as "invalid" participants whose data is auto-correlated, and assign theta=0. In our case this tends to happen due to a small dataset, or a dataset with non-rich data. This is critical, as even a purely stochastic (*i.e.*, random) time series may show predictability as the result of linear autocorrelation. Using the aforementioned approach, we are able to distinguish between autocorrelation and nonlinearity.

4. **Convergent cross mapping for each user**. The next step is to apply CCM to identify a potential causal link between the variables for each user. CCM is specifically developed for analysing causality in time series variables [37]. In brief, the method works by mapping two variables to each other using the nearest neighbours of each point on the E-dimensional manifolds. When the number of points on the manifold or the 'library' size *L* increases, the nearest neighbours tend to become nearer, which improves predictions if the variables are causally linked (*i.e.*, more local information improves prediction if the variables are causally linked). This improvement is called convergence. The results of CCM are displayed in Fig. 2E. We must apply a number of heuristics when interpreting these results for each user. First, we look for a clear convergence of the CCM value, *i.e.*, verifying that the blue and red solid lines are initially increasing and then eventually level off. We also decided to apply an asymptote function to identify a single point on the y-axis where we assume each series converges. Next, we compare the two asymptotes and determine which one is the largest (*i.e.*, which one is on top). Finally, we check to see if the asymptote that is on top is also above the bivariate correlation among the variables (indicated with a black dashed line). This correlation value is calculated as a straightforward Pearson correlation between the two variables of interest.
5. **Combine the results from multiple analyses**. This step represents our extension to the CCM technique, which we developed to summarise the results from multiple CCM analyses, hence the acronym MCCM. We developed this step with the objective of obtaining a rich summary of the similarities and differences between large numbers of participants. Our objective was to generate a single graph to summarise a large number of analyses. After trying a number of visual approaches, we settled on the following approach which highlights differences between participants: For each analysis (*i.e.*, participant) we plot the difference in asymptotes (the relative difference indicates the direction of the effect) versus the difference of the largest asymptote and the correlation value (indicated the effect size). We finally calculate the mean value for the direction of the effect across all study participants in order to summarise effects among the variables for the entire sample (the standard deviation can also be used to assess dispersion if the effect distributions are approximately normal).
6. **Robustness check**. The final step is to determine the robustness of the findings. We do so by adopting a "proxy data" approach. Here, we compare our findings against a null model obtained by random permutations of the raw data—also known as a 'surrogate data' method in the EDM literature.

In the figure below, we summarise our analysis for a single participant in the smartphone dataset. Here we seek to understand the effect of two variables on each other: the number of times the participant turned on the phone by pressing the unlock button ("screen_on"), and the number of notifications that the phone received. The variables were coded as we described earlier, and therefore are counted on an hourly basis. In Fig. 2A &

2B we identify that the optimum E for this participant's data is 11 and 30 respectively. In Fig. 2C & 2D we verify that the data is nonlinear (since the maximum forecasts skill is at ~2.0 and ~1.8, which are greater than 0). Finally, in Fig. 2E we observe that, for this participant, the number of times they unlock the screen is more likely to drive the number of notifications they receive (rho = 0.5) rather than the other way around (rho = 0.17). This is because the red line reaches a higher convergence point than the blue line. Furthermore, we observe that this is a substantial finding since the red line converges at 0.5 which is much larger than the raw bivariate correlation between these two variables ($r = 0.21$, shown in black dashed line).

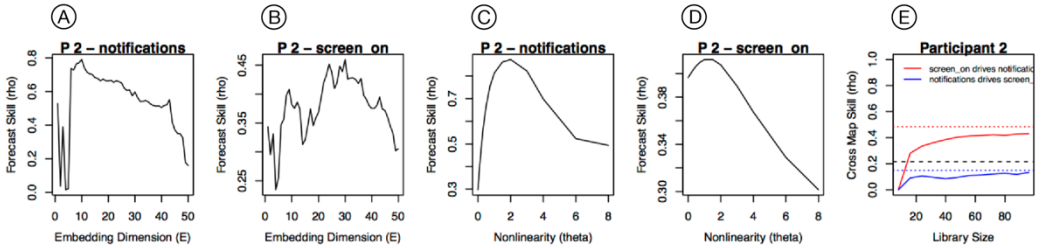

Fig. 2. Three steps in calculating an Empirical Dynamic Model.
A&B: Identify optimum values for E for both variables.
C&D: Verify non-linearity for both variables.
E: Convergent cross mapping.

In Fig. 3 we provide examples where the data fail our heuristics and we decide to discard the participant from the analysis. In Fig. 3A we show data from a participant whose data does not appear to be non-linear. In Fig. 3B we show data from a participant where our analysis does not provide significant results since the highest asymptote (red dashed line in this case) is below the correlation line (black dashed line). Finally, in Fig. 3C we show the results from a user where the CCM results do not converge (*i.e.*, the top red line appears to be flat) and thus causal effects cannot be inferred.

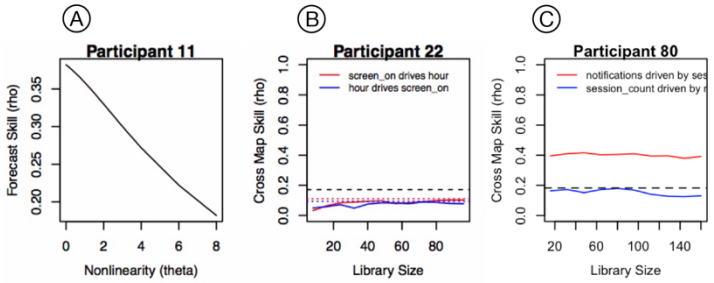

Fig. 3. Examples where a participant's data must be discarded as it fails the heuristics of the analysis.
3A: The data fails the non-linearity test.
3B: The CCM results are not significant: they are below the correlation value.
3C: the CCM results do not converge.

The process we have described in Fig. 2 is what would be followed to analyse a single complex dynamic system, e.g., studying the relationship between the number of toads and snakes in the amazon rainforest. We apply our method to each participant independently,

and therefore our analysis produces one graph per participant (as shown in Fig. 2E). Therefore, we need to extend this method and develop a meaningful way to summarise all results from all participants, and draw conclusions about the variables we are analysing for an entire sample of people. For example, this would be equivalent to studying the relationship between toads and snakes across *N* different rainforests. In our case, we may get conflicting results from different participants, and different effect sizes, and therefore it is necessary to arrive at a conclusion that moves beyond simply eyeballing the hundreds of graphs we generate.

To summarise the results from multiple participants, we adopt a geometric approach. In Fig. 4 we visualise how we can summarise the CCM results from multiple participants, or multiple ecosystems. Looking at the CCM results of each participant, we first calculate two values:

- the difference between the two asymptotes. This is the vertical difference between the red dashed line and the blue dashed line, and is an indicator of the direction of the effect.
- the difference to the raw correlation. This is always calculated as the difference from the top asymptote (in this case the red dashed line) to the correlation (the black dashed line). This is an indicator of the strength of the effect.

Having calculated these two values, we use them as the x and y coordinates respectively in a scatterplot, and we simply add a dot at those coordinates in the scatterplot. In the example in Fig. 4 the difference between asymptotes is about -0.23, while the difference to the correlation is about 0.1. Therefore, we add a datapoint at coordinates (-0.23, 0.1) in the scatterplot. In this scatterplot we use red to denote any data points (*i.e.*, CCM graphs) that are to be discarded because they fail our heuristics.

Next, we calculate the mean x-axis value for all data points that we retain in the scatterplot. This is indicated as the thin vertical dotted line at x = -0.192. This mean value is calculated using only the retained data points, ignoring the discarded data points. This value, along with the standard deviation, is then used to characterise the population and therefore summarise all results for all participants.

As shown in the rightmost of Fig. 4, all points below the horizontal axis are to be ignored, since these represent graphs where the top asymptote is below the correlation line. Additional points may also be ignored if they fail one of the other heuristics (failing the non-linearity test, or lack of convergence), and in our experience this tends to happen with small or non-rich datasets. Furthermore, any data points in the top-left of the scatterplot indicate that variable 1 is stronger, while points in the top-right indicate that variable 2 is stronger.

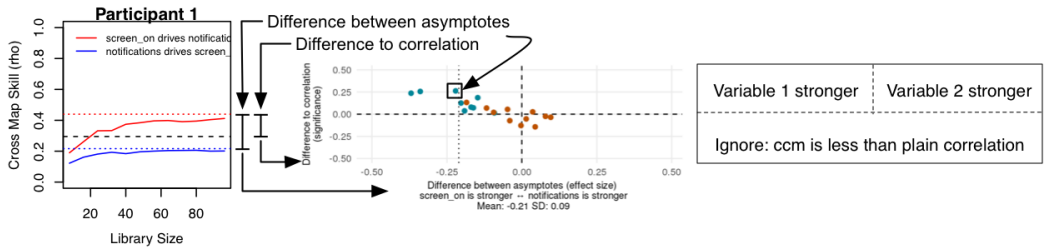

Fig. 4. A visualisation of our geometric approach to summarising CCM results for multiple participants. For each participant we calculate the difference between asymptotes and the difference to correlation. These two values become the (x,y) coordinates of a data in our summary scatterplot. All retained values are used to calculate the x-axis mean, as a means to summarise the overall outcome of the analysis.

Finally, we have developed a method to validate our results. We do so by comparing our results to a null model. We generate null models with the use of "surrogate data" [34]. Surrogates are created by randomly permuting the values of the original time series on a participant level – eliminating temporal dependencies while preserving the histogram of the original data. We expect that if our findings are simply due to broad statistical features of the data in our observations, then these random permutations will produce results that are similar to our actual results. If our actual results are demonstrably different than the random permutations, then we argue that this is evidence that there is something special about the order in which the events took place, and therefore they capture the underlying dynamics of a complex system that evolves over time.

To conduct this validity test, we run CCM on the surrogate data and subsequently store the asymptote differences (*i.e.*, the x-axis coordinates in Fig. 4). The generation of surrogate data and subsequent CCM calculation is independently repeated 25 times per participant. All values are reordered pairwise, thereby maintaining the correlation between the two timeseries. We present a visual comparison of the actual data versus the surrogate data (see Fig. 5 for an example), and consider if the mean asymptote difference that we report in the plots is likely to belong to the distribution of values observed in the surrogate data. We do so by considering the median and 95% confidence interval of each distribution. As can been seen from Fig. 5, the difference between asymptotes for the surrogate data is centred around zero. These results are substantially different from the results obtained from our participant population (shown directly below). It must therefore be the temporal order of the data (as observed in the participant data but randomly permuted in the surrogate data) which causes the difference between asymptotes.

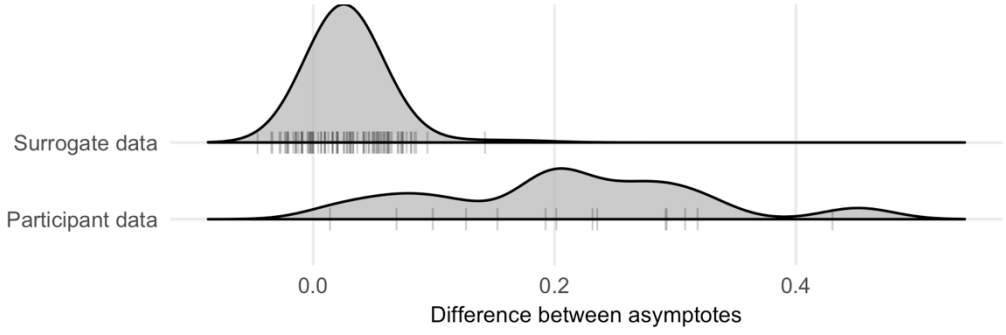

Fig. 5. Comparison of CCM outcomes of surrogate data and participant data. If the two distributions are different then we expect that our findings are not due to chance.

3.3 Results

We now present the results of applying our method to investigate the relationship between a number of variables in our datasets. First, we investigate whether device use is driven by notifications, or the other way around. There is increasing literature suggesting that users have a hard time managing notifications on their mobile devices, and that work has suggested that more notifications may be causing people to use their device more often [25,35]. To analyse this relationship, we analyse these two variables in our two datasets independently. From the first dataset we analyse how often people unlock their screen vs. how many notifications they received on their smartphone. In the second dataset we repeat this analysis for smartwatch users. The results are shown in Fig. 6 and suggest that for both smartphones (top) and smartwatches (bottom) device usage drives the number of notification and not the other way around. The effect is stronger on smartphones (0.208 versus 0.103 for smartwatches). In the case of smartwatches, we observe that several users do tend to experience this relationship reversed (*i.e.*, the notifications appear to be driving their use of the smartwatch). We now compare our results from the datasets against their respective surrogate analysis. For the smartphone dataset, the observed mean value of -0.208 (95% CI [-0.273, -0.141]) differs considerably from the null model's mean value of 0.003 [-0.010, 0.003]. We visualise this comparison for the smartphone dataset in Fig 5. Our surrogate analysis for the smartwatch dataset similarly suggests that the observed value of -0.103 [-0.123, -0.087] is different from the null model 0.000 [0.000, -0.004].

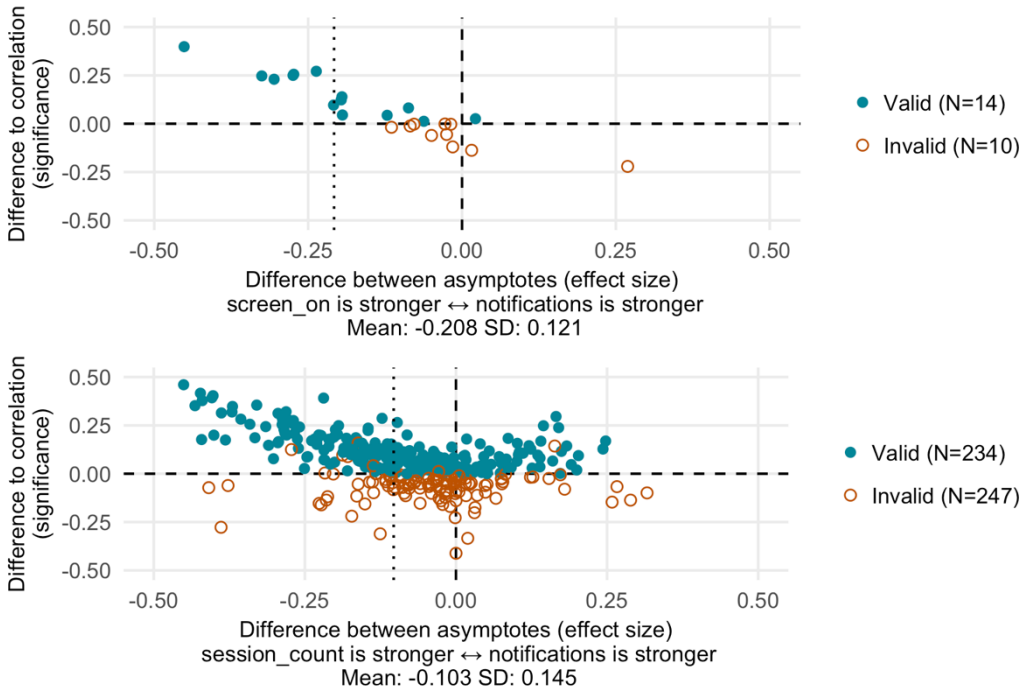

Fig. 6. Results from analysing the relationship between using a device (top: smartphone; bottom: smartwatch) versus receiving notifications on that device.

Next, we present an analysis to determine whether people's use of technology is driven by battery level, or the other way around. There is a growing literature reporting on how people manage power on their mobile devices, charging strategies, and in general how they perceive the autonomy of their devices [7,13,19]. By applying our method (Fig. 7) we show that overall individuals' use of the device has a stronger effect. By inspecting the results we observe that in fact smartphone users (top of Fig. 7) are somewhat spread out, suggesting a weak causal effect. For the smartwatch data (bottom of Fig. 7) the effect is much stronger in favour of device usage (0.155 vs. 0.041 for smartphones). Our surrogate analysis for the smartphone dataset suggests that the observed values of -0.041 [-0.098, 0.019] align to some extent to the null model (0.001 [-0.006, 0.008]), and this is likely because users are spread out in Fig. 5. For the smartwatch dataset, our surrogate analysis suggests that the observed values of -0.155 [-0.178, -0.135] differ from the null model (-0.006 [-0.007, -0.001]).

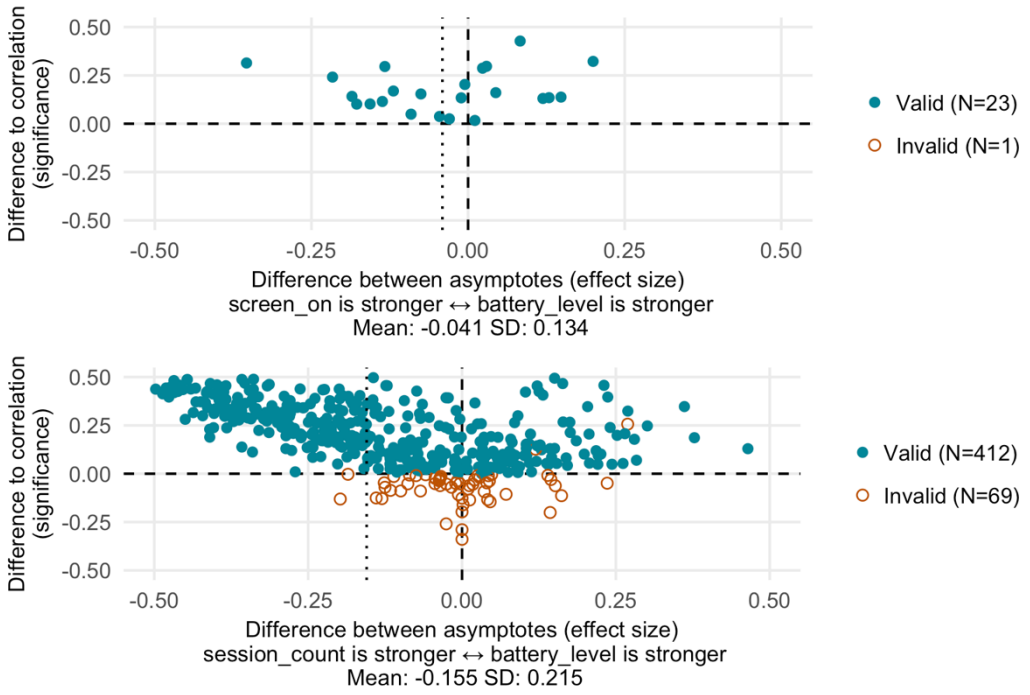

Fig. 7. Results from analysing the relationship between using a device (top: smartphone; bottom: smartwatch) versus the battery level on that device.

Finally, we perform a series of "sanity checks" to confirm that our analysis is sound. Up to this point, it is plausible that the results could be wrong, since we have no objective way of knowing the ground truth (i.e., the data-generating process). In fact, ground truth is impractical to generate here—even if we asked each individual participant to tell us their opinion or provide us some labelled data, we would expect that it is challenging for participants to accurately self-reflect on their past behaviour and precisely quantify the effect of the two measured variables in each direction.

We therefore adopt a number of strategies to further test the validity of our approach: testing for the impossible, and testing for randomness. First, we analyse a pair of variables where we know that the relationship is unidirectional and it is absolutely impossible for the

relationship between the two variables to be bidirectional. If our analysis behaves as expected, then the results should be overwhelmingly in the direction of one variable over another.

To implement this strategy, we consider the following pairs of variables: number of times the device is unlocked versus the hour of the day. Hour of day is a numeric variable ranging from 0 to 23, and device usage is again a numeric variable describing the number of times the device was unlocked for a given hour. Given these two variables, we can speculate that the hour of day may influence how much our participants use their device. We expect device usage to increase during the day and to be mostly absent during the night. However, it is impossible for any effect to be present in the opposite direction—the amount of times a device is used cannot possibly affect the time of day (unless we expect phone use to alter the space-time continuum). The results show that our method produces the expected result: device usage is driven by hour of day (see Fig. 8), and not the other way around. This is in line with expectations, and confirms that our method operates as expected. Our surrogate analysis suggests that the observed distributions (respectively -0.126 [-0.175, -0.087]; and -0.117 [-0.136, -0.100] are different to the null models (respectively 0.001 [-0.008, 0.008] and 0.003 [-0.001, 0.006]).

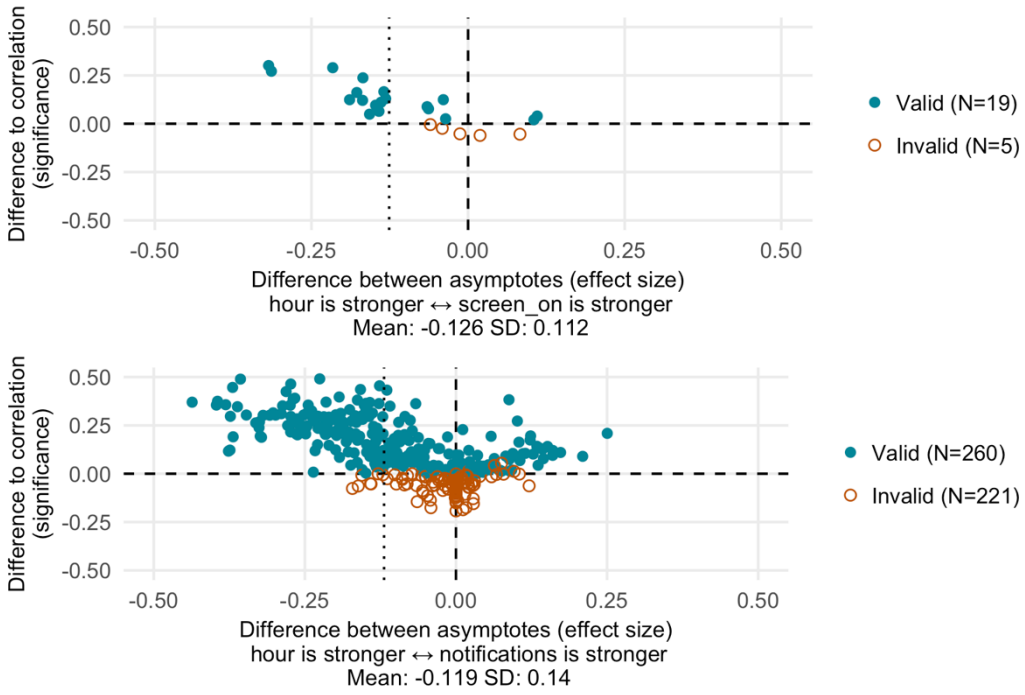

Fig. 8. Results from analysing the relationship between using a device (top: smartphone; bottom: smartwatch) versus the hour of day.

We conduct an additional sanity check by attempting to analyse random data. In this case, we analyse smartphone usage versus a randomly generated number between 0 and 99. We expect that there should be no apparent relationship between these two variables in either direction. The results in Fig. 9 do indeed show that for most participants the results fail the heuristics, and for the few remaining participants the results are small and close to 0. This confirms our expectation that no apparent effect is observed. Our surrogate analysis

suggests that the observed distribution (-0.002 [-0.020, 0.027]) is very similar to the null model (-0.006 [-0.130, 0.001]).

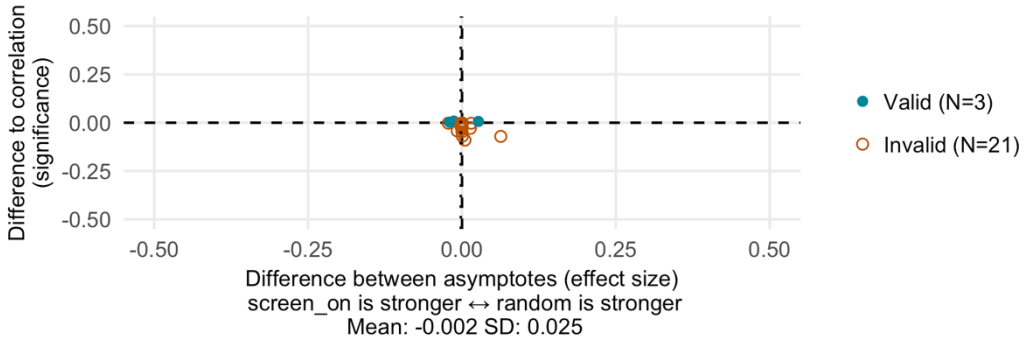

Fig. 9. Results from analysing the relationship between using a smartphone versus a random number.

Our next sanity test, and in many ways a prime demonstration of the benefits of our method, comes from analysing data that is correlated but we know there is no causal effect between these two variables. We created two variables (*i.e.*, columns in a table) in the smartphone dataset, as follows:

- random2 := 0.5*notifications + random1
- random3 := 0.3*notifications + random1

Because of the way 'random2' and 'random3' are generated, they have a very high correlation ($r = 0.97$), however we know that they cannot cause each other because they are only directly affected by the variable 'random1'. This is a typical example of a confounding variable giving rise to an apparent correlation between two other variables. Analysing this data using our method we find that indeed there is no effect between random2 and random3, and generate inconclusive results (Fig. 10). Our surrogate analysis shows that the observed results (-0.021 [-0.058, 0.014]) are very similar to the null model (-0.003 [-0.009, 0.005]), and a visual representation of this comparison is show in Fig 11. Here we can visually confirm that the values for participant data and surrogate data are closely aligned.

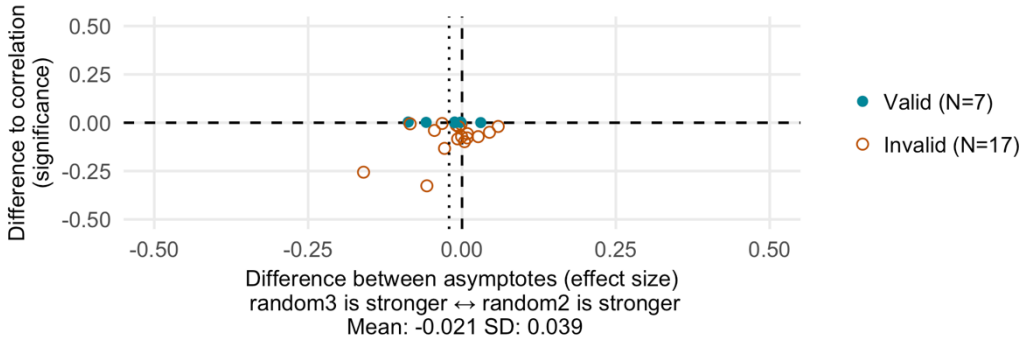

Fig. 10. Results from analysing the relationship between two random variables that are correlated, but do not affect each other.

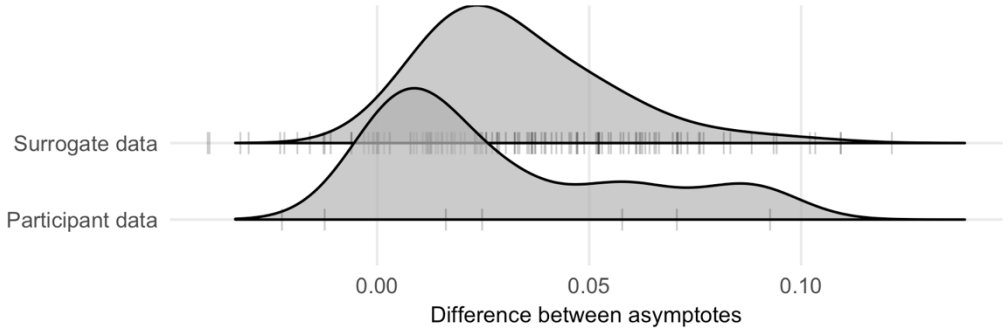

Fig. 11. Comparison of MCCM outcomes for the relationship between random2 and random3. Surrogate data and participant data are closely aligned.

Next, we analyse physiological data from Dataset 3. Specifically, we analyse the relationship between finger temperature and battery temperature. The correlation between this data is high ($r = 0.85$) since both variables are affected by the ambient temperature. However, we expect the variables not to affect each causally. Our analysis (Fig. 11) confirms that indeed there is very weak causality between these two variables. Our surrogate analysis shows that the observed values of 0.010 [-0.029, 0.007] are similar to the null model (-0.001 [-0.01, 0.008]). This case highlights an example where the data are nonlinear but not causally related.

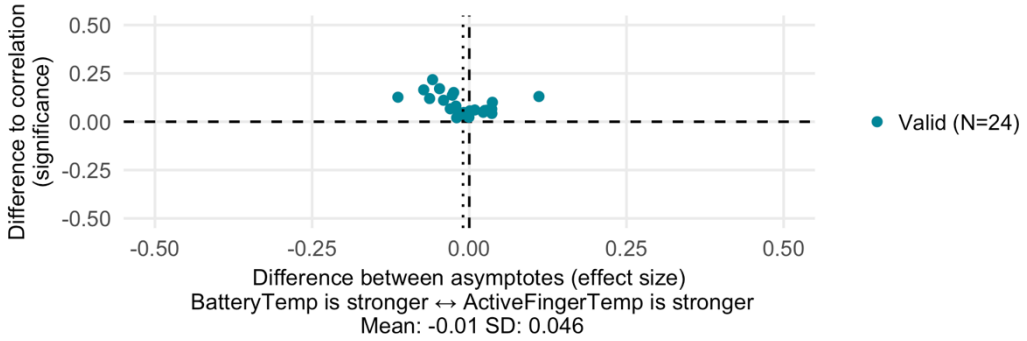

Fig. 11. Results from analysing the relationship between battery temperature and finger temperature as recorded in a laboratory study. The two variables are correlated, but do not affect each other.

Finally, we present a comparison between the aforementioned test results and their respective surrogate results in Fig. 12, containing the asymptote differences for all original and surrogate data. This overview visualises the aforementioned mean values and confidence intervals, and provides further evidence for our method of analysis. Tests which report no clear causality have a strong overlap with the causal data (*e.g.*, tests with random data or the 'BatteryTemp' ↔ 'ActiveFingerTemp' test) – whereas test which report a strong causal relationship have no overlap with the surrogate data. Close alignment with the surrogate data indicates that the order of the data is unable to reveal causal information on the variables of interest.

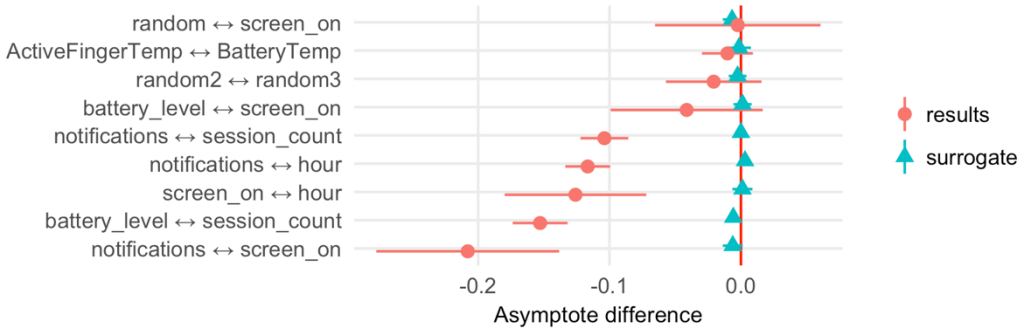

Fig. 12. Overview of asymptote difference between the original data and the surrogate data.

4 DISCUSSION

HCI researchers have typically drawn on a variety of methods for analysing their study results [22]. Even though the dictum 'correlation does not imply causation' is well known within our discipline, only a handful of previous work has aimed to tease apart correlation and causality. Here, we present a novel causality test from nonlinear dynamic system analysis, apply it to multiple recent datasets, and develop a new way to interpret the results for multiple participants.

Although previous work has applied this method to singular entities (*e.g.*, [8,37,46]), HCI researchers typically analyse larger groups of participants rather than individual entities. Our proposed method of analysis differs from the existing work in HCI on identifying causal relationships (*e.g.*, [26,39]). First, our approach does not use correlation to determine pairs of variables worthy of further investigation. As we have shown in our results, correlation and causality can be quite independent, and therefore using correlation as a precondition for further analysis can lead to unreliable results. Second, previous approaches have often assumed that all variables of interest (including confounding variables) are accounted for, but this not required with CCM. Given Takens' Theorem suggesting that it is possible to reconstruct a complex system based on a single variable's time-series, the role of any unobserved variables is captured even when that variable is not directly observed. This is important, as it is unlikely that we can capture all variables that may be related to our variable of interest when studying participants in-the-wild.

We analyse a variety of datasets to demonstrate the applicability and validity of our method. Our results show that smartphone and smartwatch usage drive incoming notifications, and not the other way around. As shown in Fig. 6, the effect is stronger for smartphone users than for smartwatch users. Similarly, we analyse the effect of device usage on battery level and again find differences between smartphones and smartwatches. The behaviour of smartphone users is not one-sided, whereas the smartwatch data indicates that device usage drives battery level rather than the other way around (Fig. 7). Following this, we present a series of sanity checks to verify the correctness of the presented method. We show that the hour-of-day is not driven by device usage (which would be impossible), but the causal relationship is in fact the other way around (Fig. 8). We generate synthetic data to show that even highly correlated data does not necessarily give rise to causation in the presence of a confounding third variable (Fig. 10). Finally, our results show that the

time-series based CCM method can also be used in task-based laboratory studies by considering each task as an element in a series. As shown in Fig. 11, although finger and battery temperature are highly correlated [33], they do not have a causal relationship. This shows that CCM is an applicable method not only for in-the-wild studies, but can also be applied in laboratory-based studies. CCM could therefore be of potential use in classic low-level ergonomic experiments, which are the foundations of much of today's HCI research.

One important detail that is not apparent in our results is the computational intensity of our method. The computational complexity grows linearly with the addition of additional participants (all of which are considered as individual ecosystems). The analyses presented in this paper take 4 days to complete on a single 3.2 GHz processor. In our analysis script we implemented parallelisation which allows the analysis to complete significantly faster: using a 32-core machine the analysis time was reduced to less than 4 hours. Thankfully the analysis lends itself to parallelisation, since each participant's data can be analysed independently, and at the end all results are combined to generate our plots.

Finally, we highlight that interpretation of the results, and the quality of the results, depends substantially on the sample size. In Fig. 8 top, we observe that for one participant (in the top-right quadrant) we have obtained a seemingly impossible result: device use affects the hour-of-day. The presence of this datapoint suggests that if our sample consisted of that sole participant, then we would be seemingly faced with an impossible result. Therefore, it is important to interpret the sample as a whole, and that is why we have decided to not simply report mean values but also to visualise the results of all participants. This situation bears great resemblance to the work by Bennett *et al*. [1] who reported in an fMRI study the surprising result of brain activity in a dead salmon. The salmon was 'presented' a set of photographs depicting humans in social situations and asked to identify the emotion of the human shown in the photo. Due to the large number of analyses completed in an fMRI study, some of the tests turned out positive despite controlling for multiple comparisons in the fMRI results. These results would indicate that there was in fact actual brain activity in the dead salmon. Earning an *IgNobel* prize for their study, Bennett *et al*. [1] showcase how the multiple comparison problem can lead to incorrect interpretation of results. Analysing multiple deceased salmons would have indicated that their initial results were in fact noise rather than actual brain activity. Similarly, an increase in sample size in HCI studies will strengthen the reliability of the results and avoid misleading conclusions due to noisy small samples. A strength in our analysis is the fact that participants are treated as a potentially unique dynamic system, while summarising these various ecosystems (*ergo*, participants) in one figure, rather than just a single number. This allows for a rigorous inspection of outliers and interpretation of the general trend(s) between two variables across participants. In addition, our comparison between the original data and generated surrogate data further demonstrates the reliability of our results (Fig. 12).

4.1 Data Analysis in HCI

Traditionally in HCI we conduct controlled experimental studies in which two (or more) systems are compared in terms of multiple variables. By strictly controlling the experiment and ensuring that the only difference notable to participants is in the presented systems, researchers aim to explain the effect of the system's differences on the user's performance or attitude. As a result, the relationships we analyse are restricted to a single direction: how

does the system affect the user (*e.g.*, user performance, user preference)? However, the relationship between user and system is bidirectional rather than unidirectional [3], similar to the bidirectional ecosystem relationship between wolves and sheep. For example, the usability of a system may attract users to use a system more frequently, and this increased usage will in turn also affect the user's interaction with the system.

Analysis of a study in which two systems are compared typically relies on *t*-tests, ANOVAs, or related non-parametric tests (*e.g.*, Wilcoxon signed-rank test) to investigate whether an effect is likely. However, these tests are unable to provide an indication on the relation of causality of the relationship. As shown in Fig. 11, it is possible for two highly correlated variables ($r = 0.85$) to have limited causality – indicating that the variables do not affect each other in any way. Determining the existence and direction of a cause-and-effect relationship between two variables is helpful in a wide variety of HCI studies, and the method presented here can achieve just that.

4.2 Causality in HCI

Convergent Cross Mapping is neither the first nor the only method to infer causality. In fact, the study of causality has brought forward a variety of statistical approaches aimed at this goal [30]. Such approaches are typically based on a combination of a model and corresponding measurements of the system [27]. However, as indicated by Mønster [27], in many cases such a model of the system is not available, or the multiple available models provide conflicting information – this is especially true in the field of complex natural, technical, and social systems. We argue that such problems are also faced in HCI, where for example the use of a technological artefact can be considered as a dynamic complex system which cannot be fully captured in any single model or set of models.

We therefore turn to the use of model-free methods in establishing causality. Granger causality, originally published in 1969 [18], is likely the most widely known method used to determine the relation between two timeseries. Other methods include the use of lagged correlation and Bayesian networks [20]. CCM, the method we apply in this paper, was proposed as an alternative to these methods, most prominently as an alternative to Granger causality. Granger causality is used for the analysis of two easily separable variables in a linear system (*e.g.*, stock market performance and a country's economic growth). CCM on the other hand, is suitable for the analysis of weakly coupled variables in a non-linear dynamic system [27,37]. Furthermore, Granger causality assumes that cause comes before effect [18], whereas both the 'sheep and wolves' example and some of our results indicate that this assumption is not warranted. The analysis presented in this paper (*e.g.*, the causal relationship between battery level and smartphone usage) are typical of the research questions in HCI. MCCM allows us to analyse the 'messiness' of real-world user interaction across large and divergent participant samples.

Furthermore, DeAngelis and Yurek [9] point out the central role of equations in modern science, stating that "*mathematics has not had the "unreasonable effectiveness" in ecology that it has had in physics*" [9]. This stems from the fact that it is near impossible to parameterise all aspects of an ecological system in a single model. As such, rather than formulating equations to construct a model, the authors state that the collected data should directly determine the model [9]. This notion forms the basis of Ye *et al.*'s [46] equation-free ecosystem forecasting using empirical dynamic modelling. Equation-based modelling in HCI faces the same problems as identified in ecological modelling. Capturing and

measuring all aspects of interaction between a user and an artefact, including a complete overview of the user's context, is near impossible regardless of the care a researcher takes in controlling a study. Takens' theorem describes how the future state of a complex dynamic system can be predicted using time series data of only a single variable of that system [38]. This is an important property for the analysis of observational, in-the-wild studies. Given the nature of in-the-wild studies, researchers are unable to control for all confounding variables which may potentially affect the variable of interest. Takens' theorem suggests that these latent variables nevertheless leave an imprint on the variables captured by the researcher. Returning to the example of wolves and sheep introduced at the onset of this paper, it is easy to imagine that the availability of grass affects the sheep population. Even though the variable 'grass' may not be measured by the researcher, changes in the availability of grass are reflected in changes in the sheep population. As such, the analysis can determine whether there is a relationship between wolves and sheep without necessarily measuring the amount of grass, rain, or other potentially confounding variables.

The implication for HCI researchers is that when using our proposed method it is not necessary to capture all aspects of the context of the participant, which would be impractical, but that sampling can be limited to those variables of interest that can be captured *reliably*.

4.3 Study Designs in-the-wild

The paradigm shift of conducting research in-the-wild rather than in a laboratory has resonated strongly with the HCI community [4,32]. In transitioning from laboratory studies to real-world observations, HCI researchers have often relied on the lab-based practice of introducing conditions to their study designs. We summarise common study configurations for lab-based and 'in-the-wild' studies in Fig. 13. Introducing conditions 'in-the-wild' does however introduce an interesting incongruity: imposing artificial study conditions upon participants as we attempt to study them in a naturalistic environment. This level of 'undisturbed' observation is typically seen in ethnographic research but rarely in empirical work.

We believe that using the analysis method presented in this paper, researchers can achieve a better understanding of their participants' interaction with technology without the need for experimental conditions. This approach, sometimes labelled as computational ethnography, compels us to rethink the design and goals of experiments in HCI. Conceptualising the participant's world as a dynamic system allows us to study this environment without introducing artificial conditions in the participant's world to determine significant effects. Our method allows us to identify relationships between variables and obtain a higher-level understanding of the participant's interaction with technology. Arguably this approach is more compelling to use when studying macro-level behaviours, such as in the case of digital phenotyping, rather than micro-level behaviours, such as text entry usability. We therefore consider this approach highly attractive for the study of complex, high-level concepts – which are by definition intertwined with other variables. Although these 'macro-level' studies could broaden our understanding of the effect of technology on peoples' lives, these types of study are currently underexplored. 'Traditional' analysis techniques (*e.g.*, ANOVA, linear modelling) require researchers to either reduce the complexity of the outside world (*i.e.*, laboratory study) or collect data on

all potentially influencing variables – a non-viable approach. The method we have presented, on the other hand, intrinsically captures the effect of these latent variables on the collected variables.

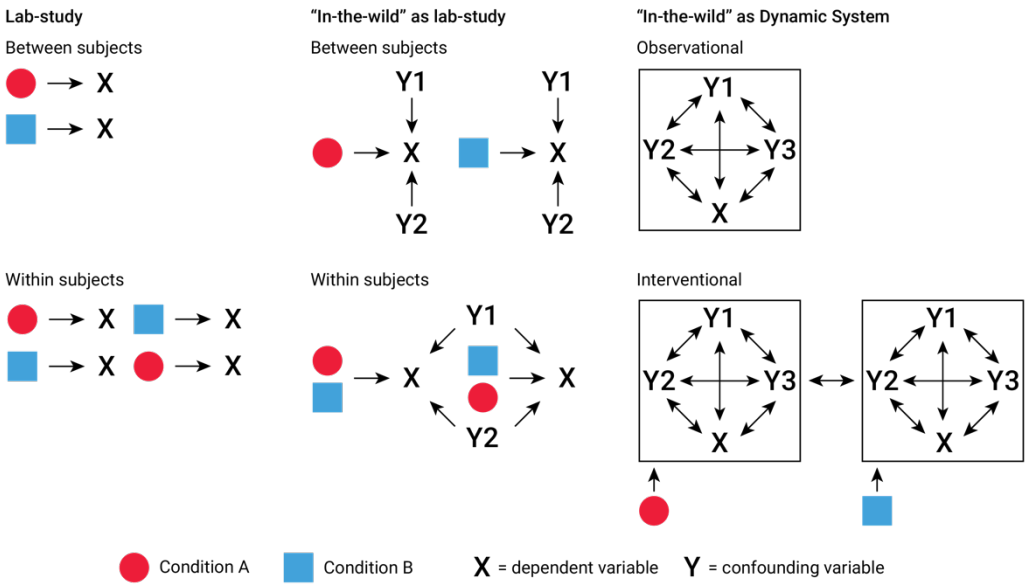

Fig. 13. Summary of lab-based studies, condition imposed-'in-the-wild' studies, and observational/interventional 'in-the-wild' studies as dynamic system.

Although a limitation of the presented analysis technique is the requirement for repeated data collection (*e.g.*, timestamped data), there are a number of advantages which we have not yet discussed. MCCM allows for the identification of cause-and-effect relationships between two variables (given that they meet the time series criteria). MCCM does not offer a one-stop solution to analysing more than two variables simultaneously, as is for example the case in multiple regression. In essence, multiple regression allows researchers to identify which variables affect the variable of interest and to get a sense for the size of their effect. To construct an answer on questions involving three or more variables, MCCM requires independent pairwise analyses on the various combinations of variables. Following this, the individual test results can be used to answer the overarching question. While being more laborious than a single test, the MCCM will provide richer insights into which variables drive which. Similarly, MCCM cannot be used to analyse the full causal relationship of control variables. To clarify, consider an analysis with variable A (control) and variable B (dependent). As the researcher controls variable A, variable B cannot simultaneously control variable A. The effect can only be in one direction (variable B affects variable A). This is similar to our analysis of causality between hour of day and device usage, as it is simply impossible for device usage to drive the hour of day (Fig. 8).

Studies in which the goal is to compare two or more artefacts are however not always feasible without introducing any conditions. For example, in one of our previous studies we analysed the effect of gamification on the quality and quantity of mobile self-report data [42]. The study featured a between-subjects design, in which half of our participants installed a gamified application (*i.e.*, leaderboard, points, etc.) and the remaining half installed a non-gamified application. This allowed us to verify the effect of gamification

without informing those in the non-gamified condition that we are tracking their scores. The presented study design is typical in the current HCI landscape, in which two artefacts are compared by analysing their respective effect on participants in-the-wild. As MCCM does not allow for a direct analysis of variables across conditions, one can run a separate analysis for each category and a variable of interest. Doing so for a binary categorical variable (in our example: gamified or non-gamified) will generate two separate plots, identifying the relationship of interest for each condition. Then, based on these plots and summarised results, it is possible to compare the direction and effect of a variable of interest between two conditions. We label this approach as 'interventional' in Fig. 13. The same approach can be used to analyse differences in causal relationships between other categorical variables (*e.g.*, testing for differences in gender, geography, or other demographics).

Rather than analysing categorical variables which have already been established (*e.g.*, conditions, gender, etc.), the proposed visualisation method can be used to visually identify unknown clusters in the dataset if they exist. For example, we know from literature that people use their mobile devices differently, and researchers have applied clustering to identify these differences [17,45,48]. Clusters can emerge in the analysis as a group of participants for which the relationship between variables is in the opposite direction or off different strength. For example, while for a large majority of participants in our smartwatch dataset their battery level is driven by smartwatch usage – a cluster of participants emerges with an opposite relationship in which battery level drives device usage.

4.4 Weaknesses and Limitations

The analysis technique presented in this paper relies on time series data. The collection of such a dataset requires repeated measurements over a period of time. Therefore, this method of analysis is only suitable for studies in which participants are observed and (continuously) tracked for an extended period of time. Data which is collected through one-off surveys, interviews, or otherwise missing a repetitive nature of data collection cannot be used in combination with CCM. The suitable granularity for time series analysis is dependent both on the richness of the data and the total duration of the study. In our analysis of Dataset 3, we show how this method can be applied in laboratory studies on a task-based granularity as opposed to a time-based delimiter (*e.g.*, minutes, hours).

In addition, the technique requires adequate volumes of data for each participant. In our analyses we use a threshold of at least 10 data points per participant, but more stringent requirements might increase this to 20 or even 30 to obtain adequate coverage of a dynamic systems for each individual. These may be the number of points in a time series dataset (per each variable), or can be the number of distinct tasks that the participant was observed doing. In our experience, the datasets tend to fail the heuristics when they have fewer data, and additionally the number of observations limits the search space for E (embedding dimension).

Finally, we have described how CCM can only analyse pairwise relationships, unlike multiple linear regression where multiple variables can be considered. In the presence of multiple variables, one has to conduct multiple pairwise analyses between the outcome variable and each of the variables of interest.

5. CONCLUSION

In this paper we present the use of Multiple Convergent Cross Mapping (MCCM) for the analysis of human behaviour in HCI. While the basis of MCCM has been in active use in other scientific fields, most prominently in Ecology, it has never been applied to HCI. By analysing time-series data of two variables, MCCM is able to detect the causality between these two variables. Whereas CCM has previously been applied to analyse single ecosystems, MCCM can summarise and visualise a large number of ecosystems (*ergo*, participants), as user studies in HCI typically involve multiple participants. Our analysis shows how these results could reveal interesting information about a participant population and present results on two in-the-wild studies, a laboratory study, and an artificial dataset. We revisit previously carried out data analyses and find examples where a high correlation does not result in causation. In addition, we show that our analysis method can reveal differences in the population (*e.g.*, clustering behaviour). Finally, we present various sanity checks to verify the validity of our approach.

We believe our analysis method will be useful for a variety of study designs in HCI. For laboratory studies, we show how our technique can be useful to establish causality between two variables even in the absence of explicit time series data by analysing experimental data on a task-level. For in-the-wild studies, we identify two additional changes in perspective. First, rather than imposing research conditions, and thus affecting the *in situ* observation of participants, the presented analysis method allows researchers to better understand the complex interaction between people and technology without inferring about the participant's low-level interaction with technology. Second, instead of utilising a unidirectional analysis such as regression (in which one variable is the dependent variable), CCM allows for a bidirectional analysis without the problematic assumptions associated with Granger causality. Crucially, the presented analysis accounts for latent variables, and therefore researchers can be selective in which variables they capture or analyse. We argue that, similar to the relationships identified in ecological systems, the relationship between a human and a technological artefact may often be bidirectional rather than unidirectional. Using the presented EDM and CCM analysis methods, researchers can identify these relationships and achieve a better understanding of the interactions between humans and technology while making minimal assumptions about the form these interactions take over time.

REFERENCES

[1] C. M. Bennett, M. B. Miller and G. L. Wolford. 2009. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction. *NeuroImage*, 47. S39-S41. DOI: http://dx.doi.org/10.1016/S1053-8119(09)71202-9

[2] S. R. Carpenter and W. A. Brock. 2006. Rising variance: a leading indicator of ecological transition. *Ecology Letters*, 9 (3). 311-318. DOI: http://dx.doi.org/10.1111/j.1461-0248.2005.00877.x

[3] John M. Carroll. 2000. *Making Use: Scenario-Based Design of Human-Computer Interactions*. MIT press.

[4] Alan Chamberlain, Andy Crabtree, Tom Rodden, Matt Jones and Yvonne Rogers. 2012. Research in the Wild: Understanding 'In the Wild' Approaches to Design and Development. In *Proceedings of the Designing Interactive Systems Conference (DIS)*, Newcastle Upon Tyne, United Kingdom, ACM, 795-796. DOI: http://dx.doi.org/10.1145/2317956.2318078

[5] Chun-Wei Chang, Masayuki Ushio and Chih-hao Hsieh. 2017. Empirical dynamic modeling for beginners. *Ecological Research*, 32 (6). 785-796. DOI: http://dx.doi.org/10.1007/s11284-017-1469-9
[6] Adam Thomas Clark, Hao Ye, Forest Isbell, Ethan R. Deyle, Jane Cowles, G. David Tilman and George Sugihara. 2015. Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, 96 (5). 1174-1181. DOI: http://dx.doi.org/10.1890/14-1479.1
[7] Russell B. Clayton, Glenn Leshner and Anthony Almond. 2015. The Extended iSelf: The Impact of iPhone Separation on Cognition, Emotion, and Physiology. *Journal of Computer-Mediated Communication*, 20 (2). 119-135. DOI: http://dx.doi.org/10.1111/jcc4.12109
[8] Angélique O. J. Cramer, Claudia D. van Borkulo, Erik J. Giltay, Han L. J. van der Maas, Kenneth S. Kendler, Marten Scheffer and Denny Borsboom. 2016. Major Depression as a Complex Dynamic System. *PLoS ONE*, 11 (12). e0167490. DOI: http://dx.doi.org/10.1371/journal.pone.0167490
[9] Donald L. DeAngelis and Simeon Yurek. 2015. Equation-free modeling unravels the behavior of complex ecological systems. *Proceedings of the National Academy of Sciences*, 112 (13). 3856-3857. DOI: http://dx.doi.org/10.1073/pnas.1503154112
[10] E. R. Deyle, R. M. May, S. B. Munch and G. Sugihara. 2016. Tracking and forecasting ecosystem interactions in real time. *Proceedings of the Royal Society B: Biological Sciences*, 283 (1822). DOI: http://dx.doi.org/10.1098/rspb.2015.2258
[11] Alan Dix, Janet E. Finlay, Gregory D. Abowd and Russell Beale. 2003. *Human-Computer Interaction (3rd Edition)*. Prentice-Hall, Inc.
[12] Zoltán Dörnyei, Alastair Henry and Peter D MacIntyre. 2014. *Motivational Dynamics in Language Learning*. Multilingual Matters.
[13] D. Ferreira, A. K. Dey and V. Kostakos. 2011. Understanding human-smartphone concerns: a study of battery life. In *International Conference on Pervasive Computing*, 19-33. DOI: http://dx.doi.org/10.1007/978-3-642-21726-5_2
[14] Sandro Galea, Matthew Riddle and George A. Kaplan. 2010. Causal thinking and complex system approaches in epidemiology. *International Journal of Epidemiology*, 39. 97-106. DOI: http://dx.doi.org/10.1093/ije/dyp296
[15] William W. Gaver. 1991. Technology affordances. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New Orleans, Louisiana, USA, ACM, 79-84. DOI: http://dx.doi.org/10.1145/108844.108856
[16] James J Gibson. 1979. *The Ecological Approach to Visual Perception*. Psychology Press.
[17] Jorge Goncalves, Simo Hosio, Niels van Berkel, Furqan Ahmed and Vassilis Kostakos. 2017. CrowdPickUp: Crowdsourcing Task Pickup in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1 (3). 51:51-51:22. DOI: http://dx.doi.org/10.1145/3130916
[18] C. W. J. Granger. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37 (3). 424-438. DOI: http://dx.doi.org/10.2307/1912791
[19] S. Hosio, D. Ferreira, J. Goncalves, N. van Berkel, C. Luo, M. Ahmed, H. Flores and V. Kostakos. 2016. Monetary Assessment of Battery Life on Smartphones. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 1869-1880. DOI: http://dx.doi.org/10.1145/2858036.2858285
[20] Kevin B. Korb and Ann E. Nicholson. 2008. The Causal Interpretation of Bayesian Networks. In Holmes, D.E. and Jain, L.C. eds. *Innovations in Bayesian Networks: Theory and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, 83-116.
[21] Diane Larsen-Freeman. 2015. Ten 'Lessons' from Complex Dynamic Systems Theory: What is on Offer. In Dörnyei, Z., D. MacIntyre, P. and Alastair, H. eds. *Motivational Dynamics in Language Learning*, 11-19.
[22] Jonathan Lazar, Jinjuan Heidi Feng and Harry Hochheiser. 2017. Research Methods in Human-Computer Interaction. In *Research Methods in Human-Computer Interaction*, Wiley Publishing, 25-44.
[23] B. Lu, E. Zanutto, R. Hornik and P. R. Rosenbaum. 2001. Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse. *Journal of the American Statistical Association*, 96 (456). 1245-1253. DOI: http://dx.doi.org/10.1198/016214501753381896
[24] Michael J. Mauboussin. 2002. Revisiting Market Efficiency: The Stock Market as a Complex Adaptive System. *Journal of Applied Corporate Finance*, 14 (4). 47-55. DOI: http://dx.doi.org/10.1111/j.1745-6622.2002.tb00448.x

[25] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Santa Clara, California, USA, ACM, 1021-1032. DOI: http://dx.doi.org/10.1145/2858036.2858566
[26] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley and Mirco Musolesi. 2017. MyTraces: Investigating Correlation and Causation between Users' Emotional States and Mobile Phone Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1 (3). 83:81-83:21. DOI: http://dx.doi.org/10.1145/3130948
[27] Dan Mønster, Riccardo Fusaroli, Kristian Tylén, Andreas Roepstorff and Jacob Friis Sherson. Year. Inferring Causality from Noisy Time Series Data - A Test of Convergent Cross-Mapping. In *COMPLEXIS*.
[28] Donald A. Norman. 1986. Cognitive Engineering. In Norman, D.A. and Draper, S.W. eds. *User Centered System Design: New Perspectives on Human-Computer Interaction*, Hillsdale, NJ: Lawrence Erlbaum Associates, 31-61.
[29] Donald A. Norman. 2002. *The design of everyday things*. Basic Books.
[30] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press.
[31] Gabriel Popkin, A Twisted Path to Equation-Free Prediction. Accessed from https://www.quantamagazine.org/chaos-theory-in-ecology-predicts-future-populations-20151013/
[32] Yvonne Rogers. 2011. Interaction Design Gone Wild: Striving for Wild Theory. *Interactions*, 18 (4). 58-62. DOI: http://dx.doi.org/10.1145/1978822.1978834
[33] Zhanna Sarsenbayeva, Niels van Berkel, Aku Visuri, Sirkka Rissanen, Hannu Rintamaki, Vassilis Kostakos and Jorge Goncalves. 2017. Sensing Cold-Induced Situational Impairments in Mobile Interaction using Battery Temperature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 1 (3). 98:91-98:99. DOI: http://dx.doi.org/10.1145/3130963
[34] M. Small and C. K. Tse. 2003. Detecting Determinism in Time Series: The Method of Surrogate Data. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 50 (5). 663-672. DOI: http://dx.doi.org/10.1109/TCSI.2003.811020
[35] Cary Stothart, Ainsley Mitchum and Courtney Yehnert. 2015. The attentional cost of receiving a cell phone notification. *Journal of Experimental Psychology: Human Perception and Performance*, 41 (4). 893-897. DOI: http://dx.doi.org/10.1037/xhp0000100
[36] George Sugihara. 1994. Nonlinear forecasting for the classification of natural time series. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 348 (1688). 477-495. DOI: http://dx.doi.org/10.1098/rsta.1994.0106
[37] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty and Stephan Munch. 2012. Detecting Causality in Complex Ecosystems. *Science*, 338 (6106). 496-500. DOI: http://dx.doi.org/10.1126/science.1227079
[38] Floris Takens. 1981. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, Springer Berlin Heidelberg, 366-381.
[39] Fani Tsapeli and Mirco Musolesi. 2015. Investigating causality in human behavior from smartphone sensor data: a quasi-experimental approach. *EPJ Data Science*, 4 (24). 1-15. DOI: http://dx.doi.org/10.1140/epjds/s13688-015-0061-1
[40] Fani Tsapeli, Mirco Musolesi and Peter Tino. 2017. Model-free Causality Detection: An Application to Social Media and Financial Data. *Physica A: Statistical Mechanics and its Applications*, 483. 139-155. DOI: http://dx.doi.org/10.1016/j.physa.2017.04.101
[41] Anastasios A. Tsonis, Ethan R. Deyle, Hao Ye and George Sugihara. 2018. Convergent Cross Mapping: Theory and an Example. In Tsonis, A.A. ed. *Advances in Nonlinear Geosciences*, Springer International Publishing, Cham, 587-600.
[42] Niels van Berkel, Jorge Goncalves, Simo Hosio and Vassilis Kostakos. 2017. Gamification of Mobile Experience Sampling Improves Data Quality and Quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1 (3). 1-21. DOI: http://dx.doi.org/10.1145/3130972
[43] Niels van Berkel, Jorge Goncalves, Lauri Lovén, Denzil Ferreira, Simo Hosio and Vassilis Kostakos. 2018. Effect of Experience Sampling Schedules on Response Rate and Recall Accuracy of Objective Self-Reports. *International Journal of Human-Computer Studies*, 125. 118-128. DOI: http://dx.doi.org/10.1016/j.ijhcs.2018.12.002
[44] A. Visuri, Z. Sarsenbayeva, N. van Berkel, J. Goncalves, R. Rawassizadeh, V. Kostakos and D. Ferreira. 2017. Quantifying Sources and Types of Smartwatch Usage Sessions. In *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems (CHI), 3569-3581. DOI: http://dx.doi.org/10.1145/3025453.3025817

[45] A. Visuri, N. van Berkel, J. Goncalves, C. Luo, D. Ferreira and V. Kostakos. 2017. Predicting Interruptibility for Manual Data Collection: A Cluster-Based User Model. In *Proceedings of the ACM Conference on Human-Computer Interaction with Mobile Devices and Services*, ACM, Article 12. DOI: http://dx.doi.org/10.1145/3098279.3098532

[46] Hao Ye, Richard J. Beamish, Sarah M. Glaser, Sue C. H. Grant, Chih-hao Hsieh, Laura J. Richards, Jon T. Schnute and George Sugihara. 2015. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences*, 112 (13). 1569-1576. DOI: http://dx.doi.org/10.1073/pnas.1417063112

[47] Hao Ye, Adam Clark, Ethan Deyle, Steve Munch, Oliver Keyes, Jun Cai, Ethan White, Jane Cowles, James Stagge, Yair Daon, Andrew Edwards and George Sugihara, rEDM: Applications of Empirical Dynamic Modeling from Time Series. Accessed from https://CRAN.R-project.org/package=rEDM

[48] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan and Anind K. Dey. 2016. Discovering different kinds of smartphone users through their application usage behaviors. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, ACM, 498-509. DOI: http://dx.doi.org/10.1145/2971648.2971696