TIME-SEGMENTED SCATTER PLOTS: A VIEW ON TIME-DEPENDENT STATE RELATIONS IN DISCRETE-EVENT TIME SERIES

Arne Koors and Bernd Page
Department of Informatics
University of Hamburg
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
E-mail: {koors, page}@informatik.uni-hamburg.de

KEYWORDS
Scatter Plots, Discrete Event Simulation, Time Series, Simulation Dynamics, Experiment Analysis.

ABSTRACT

Pairs of discrete event time series are characterised by their asynchronous nature, often hampering direct application of otherwise common analysis methods. For correct application of scatter plots, pairs of discrete event time series first have to be pre-processed and merged into a new synthetic time series of so-called co-observations. While standard scatter plots suggest analysis of global state correlation, connected scatter plots support more sophisticated hypotheses, by including sequence information. The family of time-segmented scatter plots introduced here additionally contributes time information, by dividing co-observations into time-related coloured segments. Time-segmented scatter plots permit to correlate co-observation states, state patterns and state relationships with time intervals, in order to explore time-stability of state relationships, discover otherwise overlooked dynamic patterns and possibly detect underlying processes that shape the formation of co-observation relationships. Enhanced concepts like filtered, tiled or delimited time-segmented scatter plots are available for unfavourable conditions like very high number of co-observations, overplotting, high variance or low autocorrelation. These extensions add visual aids to focus on the basic nature of co-observation relationships and their possible development. All concepts introduced in this paper are illustrated by means of a simple cash and carry warehouse example model.

INTRODUCTION

Discrete event simulation is a methodology that models dynamic systems and runs experiments on these models, in order to gain insights that can be re-transferred to the investigated original system (Page and Kreutzer 2005).
Output time series of discrete event experiments are of non-equidistant and asynchronous nature: it is characteristic for the methodology to allow arbitrary time spans between discrete events, e.g. continuous stochastic transport times. In consequence, many common time series analysis methods are inapplicable here, because they require equidistant and synchronous basic time series.

De facto, analysis of discrete event time series is widely dominated by methods of descriptive statistics, e.g. computation of mean, standard deviation or confidence intervals (cf. e.g. Fishman 2001, Banks 2010, Law 2014). In doing so, sequential and time-related information of discrete event time series is lost; it might be questioned as to how far application of methods designed for static sample distributions can comprehensively represent the true, dynamic behaviour of discrete event processes.
In previous contributions, new methods for analysis of single discrete event time series have been proposed (Koors 2013; Koors and Page 2013). A novel approach for analysis of discrete event time series pairs is described in (Koors and Page 2014). This paper introduces a method family called *Time-Segmented Scatter Plots* (figure 1), conceived for exploration of time-dependent state relationships between two discrete event simulation time series.

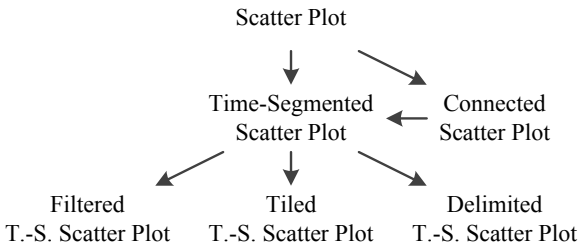

Figure 1: Time-Segmented Scatter Plot Method Family

This paper is structured as follows (cf. figure 1): Next, the common use of scatter plots and phase space diagrams is outlined. The succeeding section describes how pairs of asynchronous discrete event time series have to be pre-processed into new co-observation series, to allow correct application of scatter plots. Afterwards, the example model for the rest of this paper is delineated. The following section presents connected scatter plots. Subsequently, basic time-segmented scatter plots are introduced, extended by sub-sections for filtered time-segmented scatter plots, tiled time-segmented scatter plots and delimited time-segmented scatter plots. The final section summarises and concludes the paper.

Proceedings 29th European Conference on Modelling and Simulation ©ECMS Valeri M. Mladenov, Petia Georgieva, Grisha Spasov, Galidiya Petrova (Editors)
ISBN: 978-0-9932440-0-1 / ISBN: 978-0-9932440-1-8 (CD)

SCATTER PLOTS

Scatter Plots in Statistics

Scatter plots are diagrams used to explore relationships of two variables v_i and v_j in data sets with associated observations. One of the variables is assigned to the abscissa, the other one to the ordinate of a Cartesian coordinate system. Each pair $(s_{i,k}, s_{j,k})$ of associated state observations is represented as a discrete data point p_k in the coordinate system, whose horizontal and vertical location is determined by the observed states $s_{i,k}$ and $s_{j,k}$ of the original variables. An exemplary scatter plot of breast cancer data is shown in figure 2 (data taken from Lichman 2013).

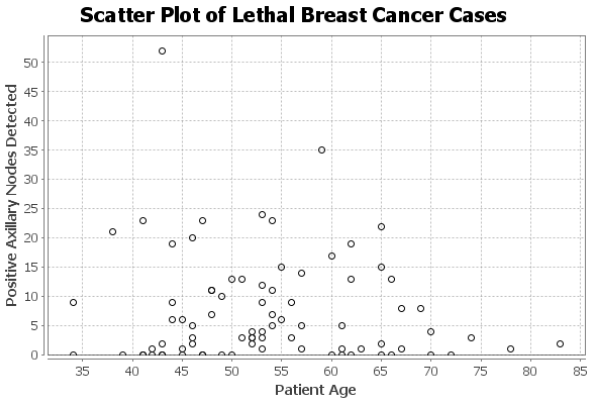

Figure 2: Scatter Plot

For analysis of three associated variables, scatter plots may be extended to the third dimension, by adding a depth axis. 2D or 3D scatter plots can incorporate further variables by mapping their values to colour, shape or size of data points. However, analysis of multi-dimensional data by single scatter plots becomes less intuitive the more dimensions are involved, since human visual and cognitive capability is restricted, especially for distinction of colour and size (Aigner et al. 2011).

Depending on characteristics and quantity of the data given, overplotting may become a problem: a) If more than one pair of observations with identical variable values $s_{i,k} = s_{i,l}$ and $s_{j,k} = s_{j,l}$ exist, corresponding data points will indistinguishably be plotted one on top of the other and therefore be lost visually. b) If there exist many observation pairs with merely small differences $s_{i,k} \approx s_{i,l}$ and $s_{j,k} \approx s_{j,l}$, visual data points may overlap and thus blur the scatter plot.

As a sideline, it should be noted that scatter plots formally only visualise coincidence in observations, not necessarily causality: Apparent relations in scatter plots may be identified a) because there actually *is* an underlying causal dependency between two observational variables or b) just by chance, without fundamental causal relationships (e.g. Leinweber 2007).

Scatter plots are typically used in descriptive and exploratory statistics, to analyse dependencies between two variables, by visually spotting functional relationships in data sets, or for identification of (possibly related) data clusters (Myatt and Johnson 2014).

Regularly, the data sets analysed originate from cross-sectional studies, where all observations are taken at (approximately) the same time instant, e.g. medical or biological studies, public-opinion polls or market surveys. Therefore, scatter plots commonly do not contain sequential or time-related information.

Phase Space Diagrams

Phase space diagrams are similar to scatter plots in that they visualise relations of two variables v_i and v_j. In contrast to scatter plots, the displayed data generally is continuous and not of discrete nature, thus phase space diagrams show continuous curves instead of discrete data points (figure 3).

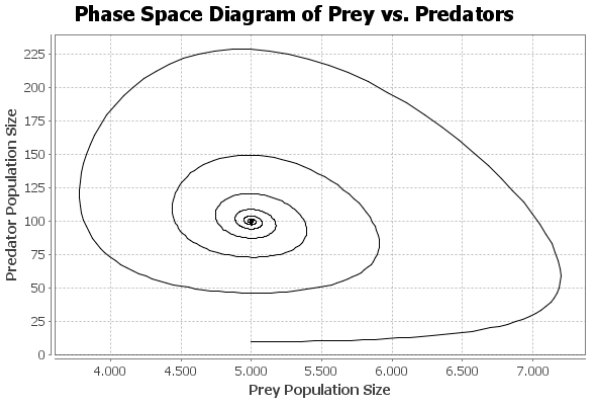

Figure 3: Phase Space Diagram

Phase space diagrams are often used in physics, where dynamical systems are described by equations. Concrete valid dynamic paths $s_*(t)$ of n system variables complying with the underlying equation set are called *trajectories* and can be described in an n-dimensional *phase space*. Phase space diagrams visualise one or a bundle of possible trajectories in phase space as geometric curves in n-dimensional Cartesian coordinate systems.

Here as well, limitations of human visual and cognitive apparatus practically lead to two- or three-dimensional phase space diagrams, illustrating relations only between two or three system variables. Sometimes phase diagram lines are coloured, to display state information of one further variable; however dashing and width of trajectory curves are usually not employed to convey further information.

Besides of physics, phase space diagrams can be used in disciplines where systems are modelled by mathematical equations, e.g. in engineering, biology, ecology or economics. These fields are often application areas for continuous simulation as well; figure 3 shows the phase space diagram of a predator-prey model, output by a continuous simulation experiment.

Although the underlying equation systems consistently include time as independent variable t, time itself is not assigned to phase space diagram axes (since time is not modelled as a state variable). Therefore, phase space diagrams do not indicate *when* certain state combinations are observed. Moreover, if trajectories do not approach fix point attractors but follow periodic orbits or strange attractors, it is not necessarily obvious in which direction the curves evolve. To sum up, phase space diagrams do not

directly contain time-related information and need not contain sequential information.

SCATTER PLOTS FOR DISCRETE EVENT TIME SERIES

Asynchrony of state variables v is one of the distinguishing features of discrete event simulation, compared to other simulation techniques.

Synchronous State Observations

In the majority of simulation approaches, e.g. continuous or agent based simulation, the simulation clock advances in fixed or variable steps of Δt, with step size determined by global settings of the simulation engine. After every time step, all model constructs (e.g. equations or agents) are updated concurrently, and new state observations $o_{*,k} = (t_k, s_{*,t_k})$ are computed quasi-simultaneously. Afterwards the simulation clock is advanced again. Dynamics arise as a sequence of synchronous state observation tuples $(t_k, s_{1,t_k}, \dots s_{i,t_k}, \dots s_{n,t_k})$, where all observed states s_{*,t_k} are tied by the same observation time instant t_k (see figure 4).

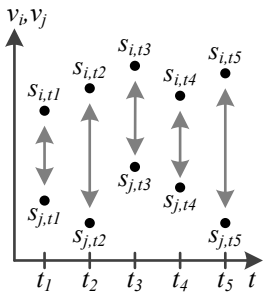

Figure 4: Synchronous State Variable Observations

It is quite natural and formally feasible to use scatter plots or phase space diagrams in case of synchronous state variables, because mapping of each observation $o_{i,k}$ of variable v_i to an observation $o_{j,k}$ of any other variable v_j is well-defined by correspondence in time: Both observations refer to the state of the same system at the same time instant t_k, just from the viewpoint of different state variables.

Asynchronous State Observations

In contrast, advancement of the simulation clock in discrete event simulation is determined by time stamps of scheduled events. Cyclically, the simulator advances the simulation clock to the time stamp of the next event note on the event list, executes the corresponding event routine and then picks the subsequent event note from the event list. Although the simulator executes the flow of events, it has no control of inter-event time spans Δt: time stamps of event notes are set by simulation model event routines, (often stochastically) re-/scheduling other events at arbitrary points in time.

Within event routines, some state variables may change (e.g. length of a waiting queue during the arrival event of a client), but regularly many entities and their state variables are not influenced by particular event routines and thus remain unaltered. In consequence, four state change constellations of two variables v_i and v_j may be observed (figure 5, left):

a) Both variables may change synchronously (t_5). This typically is the case when both variables are affected by the same event or by different events occurring at the same simulation time instant.
b) One variable is modified by an event routine, while the other one remains unchanged (t_2, t_3, t_4). This is the normal mode of progress for most pairs of state variables.
c) At the start of a simulation experiment, one variable may be assigned a (sequence of) value(s), while the other one remains undefined for a certain time (t_1).
d) At the end of a simulation experiment, one variable may have been modified recently, while the other one remained unchanged for a long time (t_6).

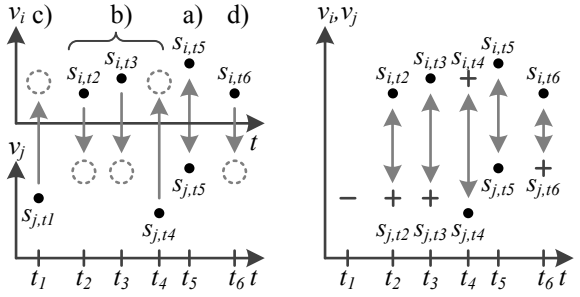

Figure 5: Asynchronous State Variable Observations and Construction of Co-observation Series

Hence, discrete event simulation experiments do not result in a sequence of synchronous observation tuples, but in a set of asynchronous time series. These time series may start and end at different simulation times and consist of a different number of observations; they may have deviating observation densities, and even within one time series event density may change locally. Moreover, between two subsequent observations in one time series (e.g. s_{j,t_1} and s_{j,t_4}), a large number of observations in another time series may have been recorded.

It is not feasible to construct scatter plots for pairs of "raw" discrete event time series as described above, because no pairs of corresponding observations will exist for most time instants.

At this point, it is helpful to recall that there exist two types of state variables:

The first *persisting* type holds the last observed state indefinitely, until it is superseded by another state observation. For example, a waiting queue has a length of 10 not only at the moment when a new client enters the queue, but for all subsequent time instants, until another client enters or leaves the queue. These state variables regularly have to be time-weighted, because here it is not only important *that* a certain state was observed (at a particular frequency), but *how long* it was observed. Other examples include server utilisation, warehousing KPIs, machine breakdown, but also exposure to harmful substances or performance of investment decisions.

On the other hand, there exists a second *transient* type of often unweighted state variables, where – strictly speaking – observations are only valid for the particular time instant they were observed at. These include all observations of

time (e.g. waiting time, service time, total processing time) and often assessment (e.g. client satisfaction) and will be discussed further below.

Co-Observation Time Series

To overcome the obstacles of constructing scatter plots for pairs of asynchronous discrete event time series, it is suggested to generate synthetic compound *co-observation* time series as follows (figure 5, right):

a) When both original state variables v_i and v_j change synchronously at t_k, a new co-observation $o_k = (t_k, s_{i,t_k}, s_{j,t_k})$ is created that contains the observed state values (figure 5, right, t_5).

b) Whenever an observation for one state variable v_i at t_k has no time-corresponding observation in the other state variable v_j, a new co-observation $o_k = (t_k, s_{i,t_k}, s_{j,t_{k-l}})$ is created, containing the actually observed variable value s_{i,t_k} at t_k and the first existent, immediately preceding value $s_{j,t_{k-l}}$ of the other variable (figure 5, right, t_2, t_3, t_4, t_6; artificially generated state values are marked with "+"). This determination is justified for all state variables of the first persisting type, where state is valid until changed: If a state $s_{j,t_{k-l}}$ was in force at an earlier time instant t_{k-l} in one time series and did not change when a state s_{i,t_k} was observed in the other time series at t_k, this state can be considered to also have been in force at t_k.

c) All initial observations of one state variable v_j for which no preceding observations in the other state variable v_i exist, are discarded (figure 5, right, t_1; marked with "–"). This determination loses part of state observations in the earlier starting time series. However, when no observations have been made for the complementing variable v_i, there is no valid basis that could be merged with leading observations of the variable v_j to form a valid prefix sequence of initial co-observations.

Note that the procedure above constructs a new time series of now synchronised co-observations, but that inter-event time spans remain variable: the event density in the new time series will still vary randomly. Simulation time deliberately has not been discretised, to maintain the characteristic properties of discrete event time series.

With the newly constructed co-observation time series, a valid basis for discrete event scatter plots is at hand.

Handling of Transient Observations

There are two options to incorporate time series containing transient observations:

i) For practical reasons, transient state observations may be handled like persisting state observations. Though, one should be aware that in some cases conceptual inaccuracies might arise: Imagine that a time series of waiting queue length observations is merged with a time series of client satisfaction observations. Then two subsequent queue length observations between two subsequent satisfaction observations will lead to co-observations where both queue lengths are related to the same preceding client satisfaction value. This would be questionable, because the satisfaction a client expresses (probably before finally leaving a system), represents his rating on previously rendered services, at this very moment; thus it should not be used to rate future waiting queue lengths.

ii) Alternatively, another approach may be used: Instead of statistically recording state variable values when they actually occur, these values can temporarily be stored in memory (e.g. in entity data fields) until the values of both corresponding state variables are known. Just then, both time series are updated simultaneously. This will generate synchronous time series pairs consistently containing observations of type a) (cf. figure 5, left), without the need to synthesise additional co-observations from preceding transient state values.

This option has another benefit: Here, the modeller can ensure that pairs of observations indeed refer to state values that conceptually belong together: In the case of client satisfaction and waiting queue length, it would make more sense to record the (e.g. average) actual waiting queue length a client experienced in combination with his rating, and not his rating along with the unrelated waiting queue length observed when he leaves the system.

EXAMPLE MODEL

The types of scatter plot proposed in this paper will be illustrated and motivated by a simple discrete event model of a cash and carry warehouse, implemented in our open source Java simulation framework *DESMO-J* (Göbel et al. 2013):

In a metropolitan area, clients order major household appliances like refrigerators, washing machines or clothes dryers via internet, which in turn are delivered to a central warehouse. Clients are notified by email when their ordered goods are available for collection and payment. When arriving at the warehouse, they park their cars and enter a waiting hall, drawing an electronically time-stamped smart card from a ticket machine. When a warehouse worker becomes available, he calls the next client in chronological order and services him by a) picking his ordered goods from the warehouse, b) handing over a pallet jack carrying the client's goods and c) collecting the purchase price plus a deposit for the pallet jack. Next, the client loads his appliance(s) into the car (probably with the help of a friend), returns the pallet jack and finally inserts the smart card into the ticket machine again. After rating his total experience on a scale of 0 (very poor) to 9 (excellent), the deposit is refunded, and the client leaves the system, now transporting his purchase home.

The smart card ticket machine records clients' total processing times (time span between entering and leaving the system = waiting time + service time) for further analysis, along with the corresponding client ratings.

The warehouse managers suppose that client satisfaction depends on two aspects: a) The longer a client has to wait without service, the lower his rating will turn out, and b) the more accommodating (and longer) a client is serviced, the better his rating is. In order to increase sales by improving customer satisfaction, the managers test a new service strategy: In normal operation mode, clients are serviced as described above. If clients' ratings decrease, additional warehouse workers are deployed to reduce waiting queues.

Moreover, clients who have been waiting (too) long are given special attention by transporting their purchased goods to their cars and assisting them in the loading process. The measured client processing times and corresponding ratings are shown in figure 6.

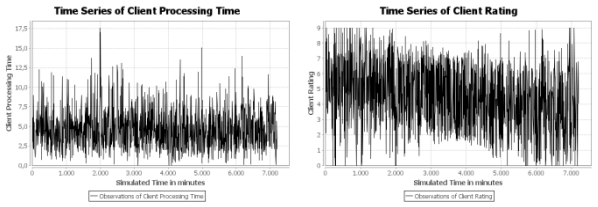

Figure 6: Time Series of Client Processing Times and Client Ratings

This pair of time series is to be explored for relationships between client processing time and client rating, in the context of the new service strategy described above.
Beyond, there exist further time series in the warehouse which the company directors want to be analysed:
1. For the last ten years, the average salary of warehouse managers was recorded, as well as average workplace satisfaction of warehouse workers (figure 7). The company directors are interested in possible relationships between these two observation variables.

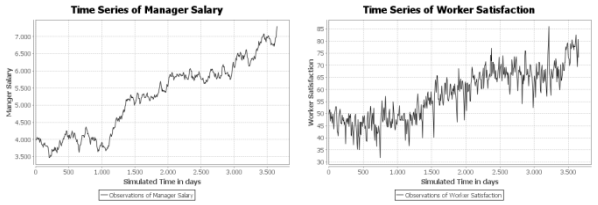

Figure 7: Time Series of Manager Salary and Worker Satisfaction

2. Likewise, a ten year record of manager's average weekly working hours shall be compared with the same manager's average assessment of life quality (figure 8).

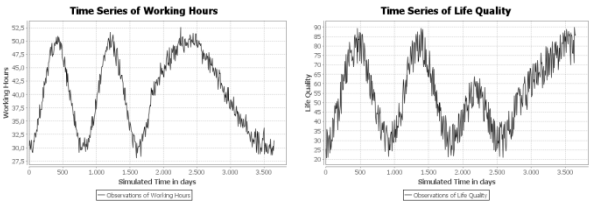

Figure 8: Time Series of Manager Working Hours and Manager Life Quality

In this paper, new advanced versions of scatter plots will be introduced, first exploring exemplarily the manager salary vs. worker satisfaction time series, followed by working hours vs. life quality and concluded by processing times vs. client rating.

CONNECTED SCATTER PLOTS

After a first view on figure 7, one could suppose that manager salary and warehousemen's workplace satisfaction are concordant: In the course of time, both time series rise in a similar manner. To examine this hypothesis, a new time series of co-observations is created as described in the preceding section *Scatter Plots for Discrete Event Time Series*, as basis for a discrete event scatter plot (figure 9).

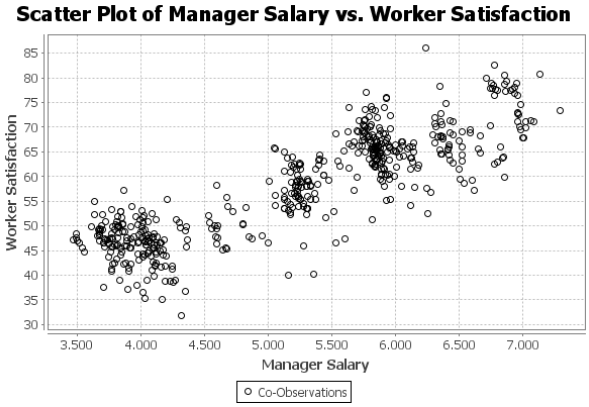

Figure 9: Scatter Plot of Manager Salary versus Worker Satisfaction

The scatter plot shows events when warehouse manager salary changed (pay rise events) related to weekly events of poll outcomes on warehousemen's workplace satisfaction.
On basis of the standard scatter plot, one could assume positive correlation between both observation variables in the last ten years: The higher the average manager salary rose (scale from 3,500 to 7,500), the higher the average worker satisfaction was (scale from 30% to 85%).
As stated in sub-section *Scatter Plots in Statistics*, correlation observations do not necessarily imply causality: Neither workers need to be more satisfied because of manager pay rises, nor manager salary may rise because workers are more satisfied (thus perhaps working more productively). Another simple explanation might just be that life standard rose in the last ten years, and independently managers earned higher wages as well as warehousemen experienced higher workplace quality. However, on the basis of the standard discrete event scatter plot in figure 9, this question cannot be decided.

In scatter plots of common cross-sectional studies, single observations normally are not collected in special order: they are virtually independent of one another. In discrete event scatter plots, observations are linked in an obvious order, defined by the order of co-observation time instants. Therefore, it is quite natural to connect data points according to the sequence in which co-observations were recorded. Figure 10 applies this concept to the salary vs. satisfaction scatter plot.
In the resulting connected scatter plot, an interesting pattern can be discovered, that was not apparent in the standard scatter plot (figure 9): against expectation, worker satisfaction seems to locally *negatively* correlate to manager salary. Observe the pattern marked with a blue ellipse in figure 10: The higher manager salary rises (here from 4,800 to 5,200 currency units), the lower worker satisfaction falls (here from 54% to 40%). Then, worker satisfaction seems to catch up with the global upwards trend, and the same pattern repeats again, from a steadily increasing baseline.

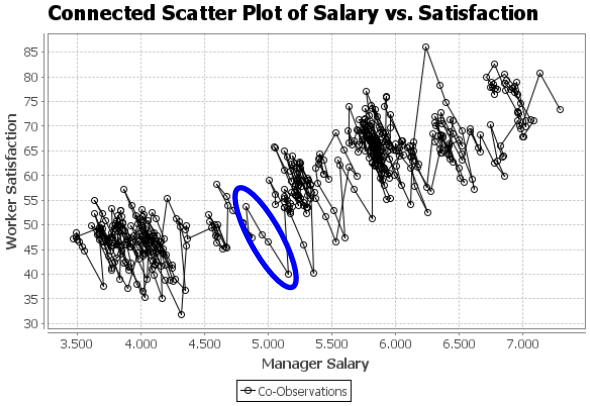

Figure 10: Connected Scatter Plot
of Manager Salary versus Worker Satisfaction

An explanation could be that manger salary is coupled to company profit. Managers may try to increase profit by lowering workplace costs and/or by streamlining warehousemen's work. The resultant measures are often not welcome to workers and could be perceived as decrease in workplace quality. The nonetheless rising global trend may be explained by a combination of two factors: a) increasing life standard within the 10 year observation period b) survivorship bias: workers not getting along with their reorganised workplace leave the company or are made redundant. In consequence, their negative workplace satisfaction is not incorporated in future polls any more, and average workplace satisfaction of the "surviving" workers stays at least at the level observed before workplace reorganisation.

To sum up, connected scatter plots can help identifying local event dynamics patterns that are not visible in the standard unconnected scatter plot, but may confirm or (here) object – respectively even negate – hypothesis merely based on global correlation observations.

TIME-SEGMENTED SCATTER PLOTS

Connected scatter plots visualise *sequence information* that is contained in original time series pairs but disregarded in derived standard scatter plots. As a next step, it is proposed to also visualise *time information* in scatter plots. For this purpose, a basic connected scatter plot is constructed and then enhanced as follows: First, the sequence of co-observations is divided into n adjacent and disjoint segments of a) equal size or b) equal time span. Note that these concepts differ in discrete event simulation, because discrete event time series may have changing local densities: two scatter plot segments with equal number of co-observations m may cover time intervals Δt_1 and Δt_2 of differing duration; on the other hand, two segments of equal duration Δt may consist of different numbers m_1 resp. m_2 of recorded co-observations.
Second, each of the n scatter plot segments is mapped to a simulation time interval, based on the co-observations it contains: Corresponding simulation time intervals start at the observation time of the first co-observation of a segment (inclusively) and end at the observation time of the first co-observation of the following segment (exclusively) or at the simulation end time instant, for the last segment.
Third, the colour spectrum is divided into n segments, and scatter plot segments as well as corresponding time intervals are mapped on associated colours.
Fourth and last, the basic connected scatter plot diagram is augmented by a) replacing the original monochrome co-observation sequence by the n co-observation segments, coloured in their associated colours and b) adding a time scale as bottom legend, where the n adjacent time intervals are coloured according to their associated colours (figure 11). Optionally, the start of every time segment can be highlighted by a small number icon, showing the ordinal number of the segment.

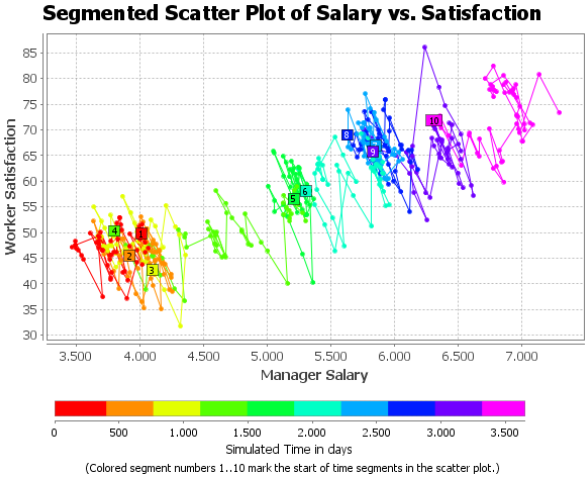

Figure 11: Time-Segmented Scatter Plot
of Manager Salary versus Worker Satisfaction

Comparable to figure 10, figure 11 shows manager salary vs. worker satisfaction, now containing additional colour-coded information *when* pay rises resp. worker polls were recorded.
Generally, the number of time segments is freely selectable. Here, it was decided to divide the scatter plot into ten segments, because original observations spanned an interval of ten years (3650 days). Thus, the small time interval ordinal numbers (1..10) in the diagram denote the start of one year segments. Without the need to resort to the original time series diagrams, it can quickly be spotted that in years 1-3 and 7-8 salary and satisfaction were rather stable. Years 4 and 10 saw a fast rise in salary, whereas in years 3, 4 and 9 the detected phenomenon of negative correlation between satisfaction and salary was particularly pronounced.

Time-segmented scatter plots document the historical development process of standard scatter plots by colour codes. They often relieve the modeller from the overhead of resorting to the original pair of time series diagrams, ending up analysing three diagrams side by side. Time-segmented scatter plots facilitate relating (connected) scatter plot patterns to simulation time intervals. Vice versa, they can help detecting temporary relationships between two variables at given sub-intervals of overall simulation time. In this respect, time-segmented scatter plots can help determine whether the relationship between two variables is either characterised by stable patterns in time or subject to consecutive regime changes.

Filtered Time-Segmented Scatter Plots

Next, analysis focuses on the ten year recording of manager's average weekly working hours compared to their average assessment of life quality. Average manager working hours were recorded in weekly events every Friday evening; average life quality assessment was recorded in manager interview events, irregularly carried out by the personnel department, on average four times per month. The scatter plot of manager working hours vs. perceived life quality is shown in figure 12.

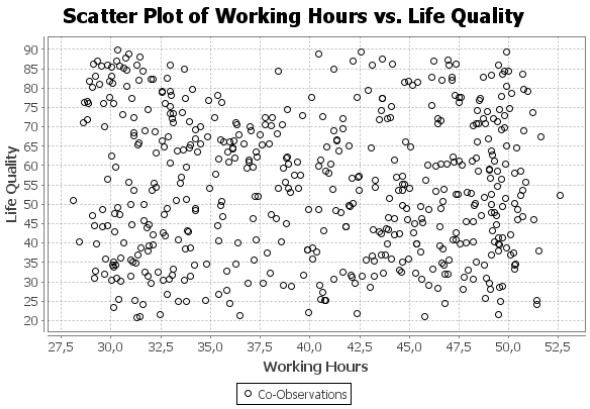

Figure 12: Scatter Plot of Manager Working Hours versus Manager Life Quality

At first glance, the scatter plot appears unremarkable, and working hours and life quality seem to be uncorrelated. For arbitrary weekly working hours (from 28 to 52 hours per week), the whole range of life quality (from 20% to 90%) has been recorded.
However, employing the new approach of time-segmented scatter plots gives a different picture (figure 13; ten time segments again): The sequence of co-observations follows a certain path, and some scatter plot regions can easily be assigned to time intervals, e.g. low working hours and high perceived life quality exclusively in year 10.

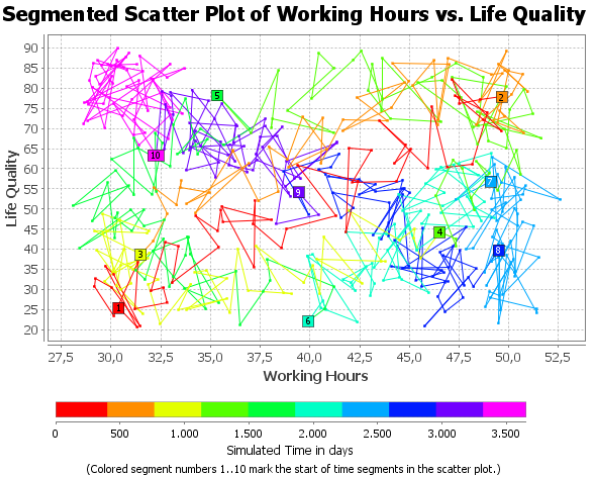

Figure 13: Time-Segmented Scatter Plot of Manager Working Hours versus Manager Life Quality

Unfortunately, the increased number of observations and their stochastic nature hamper discovery of a clear path which life quality vs. working hours may follow.
At this point, it is proposed to first filter (i.e. smooth) the original co-observation sequence and only afterwards time-segment the new filtered co-observation sequence, as introduced in the previous sub-section. To avoid visual blur, original co-observations remain unconnected, but still coloured according to their correspondent time segment. Figure 14 shows the result of this procedure.

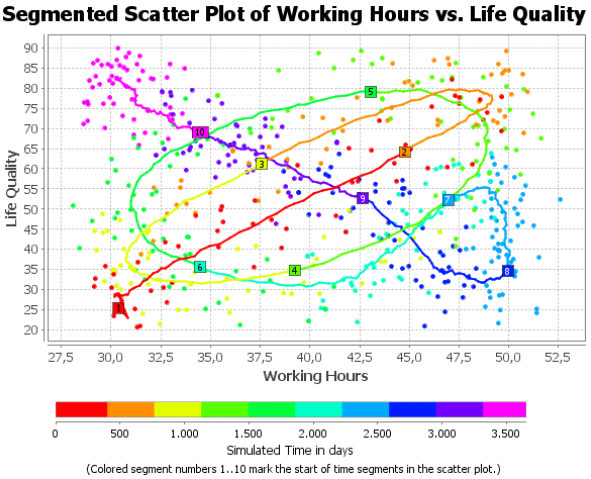

Figure 14: Filtered Time-Segmented Scatter Plot of Manager Working Hours versus Manager Life Quality

In this example, filtering is performed by a simple moving average on the co-observation components (i.e. abscissa and ordinate values are averaged independently). The moving average window size w is automatically set to $\lfloor 5\% \text{ of co-observations} \rfloor$, but may be set manually to any natural number. The moving average function itself may be substituted by alternative functions which implement a given software interface.
In order not to drop the first window of w co-observations (the w-moving average is undefined before), the w-moving average of the first v co-observations ($v < w$) is replaced by the respective v-moving averages. Therefore, the first w filtered values are less smooth then the rest of the series, but not lost.

When now looking again at the filtered time-segmented scatter plot, a clear development of life quality vs. working hours is recognisable over time: In the first five years, life quality and working hours were positively correlated: The longer managers worked, the higher their perceived life quality was. Presumably life quality was linked to income, which again might have been dependent on hours worked. The positive correlation gradually became weaker from year 1 to 5, and almost vanished in year 6. Year 7 saw a turning point: With work load remaining very high, perceived life quality decreased. In years 8 to 10 life quality increased again as work load tended to lower levels. Perhaps in the last four years, managers had reached a life standard where additional material income did not further impact on perceived life quality, and spare time became more a defining factor of life quality than higher wages.

Employing the proposed filtered time-segmented scatter plot, the apparently uncorrelated scatter plot of figure 12 turned out to have been shaped by three consecutive phases, inconsistently relating working hours and life quality over time: The role of working hours in definition of life quality radically changed over the examined ten year period, from strong positive correlation to strong negative correlation. As a consequence, further analysis of these two factors, alone or in combination, should be handled with special attention.
It should be noted here that similar conclusions as above also could have been drawn by directly comparing the original time series from figure 8 side by side; however with less comfort and precision, and only due to the simple nature of this example. The more complex relationships resp. dynamics of time series become, the more beneficial the integrative time-annotated scatter plot approach turns out.

Filtered time-segmented scatter plots superimpose an extra polyline of time-segmented filtered co-observations onto an unconnected cloud of time-coloured co-observations. In presence of many and/or stochastic co-observations, they help working out basic relationships of two variables over time, where otherwise too many co-observations would distract from the underlying nature of relationships.

Tiled Time-Segmented Scatter Plots

Last, the relationship between client processing time and client rating will be analysed.
Recall that clients were requested to rate their total experience at the smart card ticket machine, when service had ended. Client ratings were recorded along with their total processing times (= waiting time + service time). Both client processing times and client ratings are transient observation types, without a persisting impact on their state variables, beyond the instant when they were recorded. With regard to sub-section *Handling of Transient Observations* the data constellation is unproblematic, because both observations logically become available at the same time (system exit) and are synchronously recorded as well.
Also note that average client rating decreases over time (figure 6, right), which causes management to deploy more and more workers according to their new service strategy.
The scatter plot of client processing time vs. client rating is shown in figure 15.

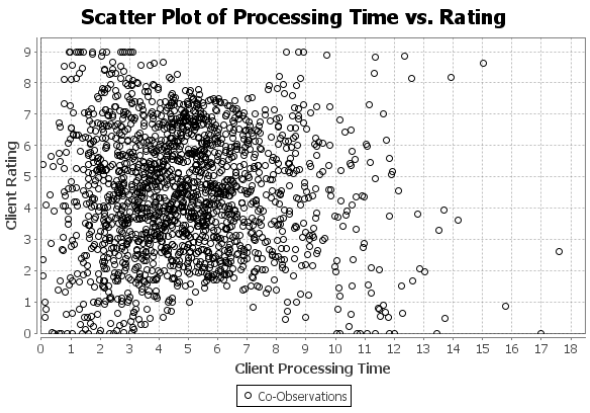

Figure 15: Scatter Plot of Client Processing Time versus Client Rating

Overplotting poses a problem here (cf. sub-section *Scatter Plots in Statistics*), due to the high number of nearly two thousand co-observations, which were recorded in two six-day weeks of ten-hour working days. In this context, time-segmented scatter plots cannot clearly reveal a time-dependent structure, because connection lines also suffer from heavy overplotting (figure 16, left). Filtered time-segmented scatter plots are just as unhelpful: Compared to the highly autocorrelated time series of the previous example, observations now are not locally correlated, but widely dispersed. Thus, all (smoothed moving) averages of the scattered time segments consistently gravitate towards their geometrical centre (figure 16, right).

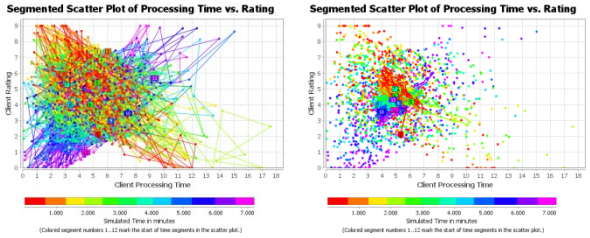

Figure 16: Time-Segmented Scatter Plot and Filtered Time-Segmented Scatter Plot of Client Processing Time versus Client Rating

In this situation, it is proposed to replace the multitude of data points by coarser structures that possibly reveal inherent relationships with less noise: The plane is divided into squares of equal parametrisable size, called *tiles*. Within every tile, the number of co-observations per time segment is counted and ranked. The time segment with the maximum number of co-observations per tile passes its associated colour to the whole tile (figure 17).

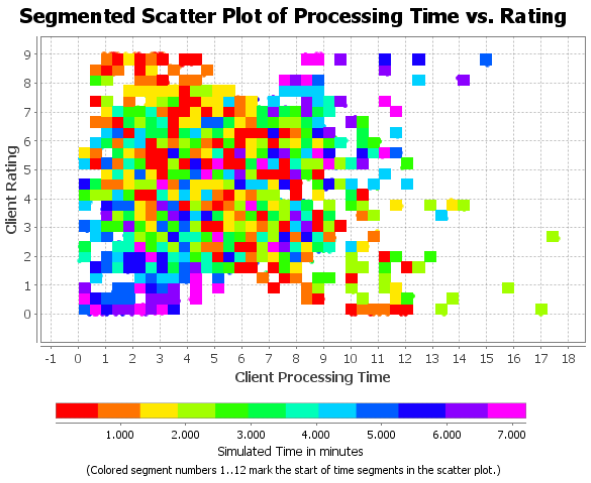

Figure 17: Tiled Time-Segmented Scatter Plot of Client Processing Time versus Client Rating

In this way, local concentrations of time segments become more recognisable. Compared to standard time-segmented scatter plots, not the last (topmost) drawn segments resp. segments with long diameters determine overall colour patterns (figure 16, left), but the very time segments actually dominating certain scatter plot regions.

Tiles are drawn on top of an unconnected time-segmented scatter plot. By setting a threshold level, the user can require a minimum number of "winning" co-observations for a tile before it is displayed. In this manner, tiles may be overlaid only in highly frequented scatter plot regions, whereas seldom frequented regions will still display their original (small) data points.
An alternative option of colouring tiles (not implemented yet) could guarantee each time segment the same fixed number of tiles, e.g. ⌊number of populated tiles / number of time segments⌋. By distributing tiles to segments (and not segments to tiles) it could be prevented that certain time segments never show up, because they never reach the topmost rank in any tile. Apart from that, further tile colouring schemes are conceivable, e.g. cellular automata-like rules etc.

To explore possible temporal relationships in the warehouse managers' new service strategy, the basic scatter plot (figure 15) is divided into twelve time segments, one per operating day within the examined two-week period (see figure 17). At a glance, a clear time-dependency between client processing time and client rating is observable: In the first days of normal operation mode, client processing time and client rating are negatively correlated. The longer a client waits for hand-over of the pallet jack, the lower his rating turns out (red, orange and yellow tiles). When the new service strategy becomes effective due to decreasing client ratings (figure 6, right), client waiting times shorten and service time becomes the major component of total processing time. In consequence, client rating schemes change fundamentally towards the end of the observation period (blue, purple and pink tiles): Eventually client processing time and client rating are positively correlated, as more intensive and longer service is rewarded by higher client ratings.
A first formal result is that the basic time series (cf. figure 6) are not suitable for standard descriptive statistics as a whole, because they were generated by a non-stable underlying process.
In substance, the warehouse managers' new strategy does not prove successful: Gradually, clients take instant service for granted and commence assigning poor marks for shorter service, despite high costs incurred by additional workers. Moreover and despite all efforts, the average client rating still decreases over time.

Tiled time-segmented scatter plots support identification of state regions where certain time segments are predominant, in circumstances where overplotting, high variance and/or low autocorrelation would blur standard, connected or (filtered) time-segmented scatter plots. Analysis of tile patterns can support detection of inherent state-time relationships that may be hard to recognise otherwise.

Delimited Time-Segmented Scatter Plots

Tiled scatter plots concentrate on indicating highly frequented state regions. At the same time, seldom frequented time segment sections are layered by tiles of other colours and are visually lost (which is intended). Consequently, tiled scatter plots hide the full state region in which time segments extend.

Sometimes, knowledge of the full area covered by time segments is expedient, e.g. when it can be shown that their form is not constant over time but changes according to an underlying process. For this kind of scenarios it is suggested to first generate an unconnected time-segmented scatter plot and to subsequently construct the convex hull for every time segment. Then, every convex hull is coloured according to the time segment it delimits (figure 18).

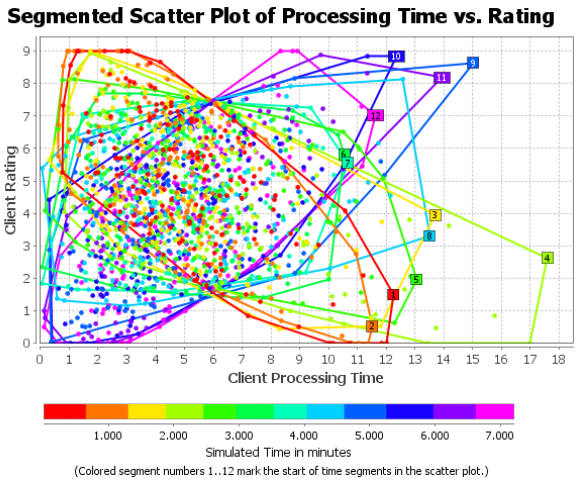

Figure 18: Delimited Time-Segmented Scatter Plot of Client Processing Time versus Client Rating

Following the colour-ordered sequence of convex hulls can reveal development tendencies in the cloud of co-observations. Object of investigation may e.g. be a) location b) form c) orientation d) perimeter and diameters and e) centroids of convex hulls.
It is conceivable to exclude outlier co-observations from convex hulls, to keep delimited areas more compact and concentrate on their basic form. On the other hand, protruding co-observations can well be distinguishing marks of certain hulls and carry non-negligible information, like relative position to the main cloud of a time segment, thereby indicating the (history of) direction(s) in which unusual behaviour might be expected.
As an option and in case of abundant co-observations in the scatter plot, a large number of shorter time segments may be chosen. Then, each time segment would receive its own delimited time-segmented scatter plot, containing only its own co-observation cloud and the corresponding hull. Stringing this sequence of snapshots together (e.g. in a .gif-file) would result in a film that could be traversed forward and backward as desired. Doing so would add a new analysis dimension, because in this manner the original scatter plot formation process could comfortably be tracked and possibly better comprehended.

In the context of the warehouse model, the sequence of convex hulls 1..12 in figure 18 shows the state space development of processing time vs. client rating over time. The same conclusions as for tiled time segmented scatter plots may be drawn, but additional insight can be gained: On days 6 and 7, client ratings are rather moderate between 1 and 8. There are no extreme ratings, and maximum processing time is quite short, below 11 minutes. The two convex hulls are rather compact and have horizontal orien-

tation. This means that clients appreciate comparatively short, reliable processing times: in these cases they assign moderate, lower-variance ratings which are virtually independent of processing time – even though a number of clients do not receive costly extra services.
As a consequence, the warehouse managers could be advised to focus on the parameters and processes of days 6 and 7 (number of workers, waiting time / rating threshold for additional services, etc.), to base further optimisation measures on these configurations.

Delimited time-segmented scatter plots show convex hull boundaries of time segments. Analysis of their developmental process can yield insight into possibly time-varying relationships of the original variables, even under difficult conditions where other methods described before may fall short or mask relevant observations.

SUMMARY AND CONCLUSION

Discrete event time series are characterised by their asynchronous nature, often hampering direct application of downstream analysis methods used by other simulation techniques. In order to correctly use common scatter plots in discrete event simulation, pairs of asynchronous discrete event time series first have to be pre-processed and merged into new synthetic time series of compound co-observations. On this basis, standard scatter plots can promote state pattern discovery, with extended connected scatter plots optionally complementing sequence information.
The family of time-segmented scatter plots introduced here additionally contributes time information, by dividing co-observations into time-related coloured segments. Time-segmented scatter plots permit to correlate scatter plot patterns respectively regions with time intervals, in order to a) confirm or object relationship hypothesis gained from e.g. standard scatter plots b) analyse time-stability of state relationships c) discover unexpected relational patterns or event dynamics, of local or global type in state/time d) possibly detect and describe an underlying process that drives the formation of co-observation relationships.
Advancing from basic time-segmented scatter plots, enhanced concepts like filtered, tiled or delimited time-segmented scatter plots are recommended for unfavourable conditions like very high number of co-observations, overplotting, high variance or low autocorrelation. These extensions add visual aids, to concentrate on the basic nature of co-observation relations and their possible relationship development.
The standard, connected and time-segmented scatter plot classes described in this paper have been implemented as extensions of our open source discrete event simulation framework DESMO-J. DESMO-J is developed and maintained in the Informatics Modelling and Simulation Workgroup at the University of Hamburg and free for download at www.desmo-j.de.
Concluding, the method family of time-segmented scatter plots often can relieve the modeller from the overhead of complementing standard scatter plots with their two original time series diagrams. It substitutes three traditional diagrams by one integrative approach, adding supplementary information to more deeply explore time-dependency of co-observation states, state patterns and state relationships.

REFERENCES

Aigner, W.; S. Miksch; H. Schumann; and C. Tominski. 2011. *Visualization of Time-Oriented Data.* Springer London, Guildford, Surrey.

Banks, J. 2010. *Discrete-event system simulation.* Pearson, Upper Saddle River, N.J, Singapore.

Fishman, G.S. 2001. *Discrete-event simulation. Modeling, programming, and analysis.* Springer, New York.

Göbel, J.; P. Joschko; A. Koors; and B. Page. 2013. "The Discrete Event Simulation Framework DESMO-J: Review, Comparison to other Frameworks and Latest Development". In *Proceedings of 27th European Council for Modelling and Simulation,* European Council for Modelling and Simulation (Aalesund, Norway, 27th-30th May 2013).

Koors, A. 2013. "Assessing Risk in Discrete Event Simulation by Generalized Deviation". In *Proceedings of the 8th EUROSIM Congress on Modelling and Simulation,* K. Al-Begain, D. Al-Dabass, A. Orsoni, R. Cant; and R. Zobel (Eds.) (Cardiff , Wales - UK, 10th-13th September 2013), 336–344.

Koors, A. and B. Page. 2013. "Application and Visualization of Financial Risk Metrics in Discrete Event Simulation - Concepts and Implementation". In *Proceedings of The International Workshop on Applied Modeling and Simulation 2013,* A. Bruzzone, C. Frydman, S. Junco, E. Cayirci; and C. Zanni-Merk (Eds.) (Buenos Aires, Argentina, 25th-26th November 2013), 118–130.

Koors, A. and B. Page. 2014. "Analysis by State: An alternative View on Discrete-Event Time Series". In *Proceedings of 28th European Simulation and Modelling Conference,* EUROSIS (Porto, Portugal, 22nd - 24th October 2014).

Law, A.M. 2014. *Simulation modeling and analysis,* Boston, Mass.

Leinweber, D.J. 2007. "Stupid Data Miner Tricks". *The Journal of Investing* 16, No.1, 15–22.

Lichman, M. 2013. "UCI Machine Learning Repository". Available at http://archive.ics.uci.edu/ml/datasets/ Haberman%27s+Survival. Accessed 10th April 2015.

Myatt, G.J. and W.P. Johnson. 2014. *Making sense of data I. A practical guide to exploratory data analysis and data mining.* John Wiley & Sons Inc, Hoboken, New Jersey.

Page, B. and W. Kreutzer. 2005. *The Java simulation handbook. Simulating discrete event systems with UML and Java.* Shaker, Aachen.

AUTHOR BIOGRAPHIES

ARNE KOORS obtained his master degree in Computer Science from University of Hamburg, Germany. Since then he has been working as a software developer and management consultant in the manufacturing industry, primarily in the field of demand forecasting and planning. Furthermore, he works as a research associate and on his PhD thesis on analysis and visualisation of discrete event simulation dynamics in the Modelling & Simulation research group led by Prof. Dr. Page.

BERND PAGE holds degrees in Applied Computer Science from the Technical University of Berlin, Germany and from Stanford University, USA. As professor for Modelling & Simulation at the University of Hamburg he researches and teaches in Computer Simulation and Environmental Informatics. He is the head of the workgroup which developed DESMO-J and the author of several simulation books.