Surrogate Predictive and Multi-domain Modelling of Complex Systems by Fusion of Agent-based Simulation, Cellular Automata, and Machine Learning

Stefan Bosse
Dept. Mathematics and Computer Science
University of Bremen
28359 Bremen, Germany
sbosse@uni-bremen.de

Abstract— **Modelling of complex dynamic systems like pandemic outbreaks or traffic flows in cities on macro-level is difficult due to a high variance on entity micro-level and unknown or incomplete interaction models. Agent-based and Cellular Automata (CA) simulations based on micro-level modelling can be used to investigate the outcome of system observables in a sandbox. For a reasonable accuracy a high number of agents, sufficient behaviour variance, high computational times, and calibrated model parameters are required. Surrogate predictive modelling of the multi-agent system can be used to replace time-consuming simulations. In this work we present a hybrid approach combining Agent-based Simulation, probabilistic contextual CA, and Machine Learning (ML). We investigate the replacement of the ABS-CA by surrogate ML models trained by simulation data. The predictive model is state-based and applied to time-series data to predict future development of aggregated system observables. We discuss and show the negative impact of uncalibrated real-world sensor data on time-series prediction and an improvement by surrogate modelling of simulation. A use-case of pandemic simulation using real-world statistical data is used to investigate and evaluate the suitability and accuracy of the proposed methods and to show the high sensitivity of surrogate modelling on distorted and biased data.**

Keywords- Large-scale simulation; Multi-Agent Systems; Cellular Automata; Surrotgae Machine Learning; Data Augmentation.

I. INTRODUCTION

The typical goal of a simulation is the prediction of the behaviour of a complex system by aggregate observables for a particular situation. A simulation can be composed of a set of interacting entities on micro-level, like humans in social sciences, to investigate and predict the outcome of system-level aggregate observables. Machine Learning as well as simulation are used to predict the response of a system to a stimulus that is hard to be studied in the real world and to get macro-level from micro-level observables (aggregates). Both techniques use data analysis and mathematical modelling [1]. In most cases a simulation is composed from elementary cells (holonomic approach). Each cell is defined by a micro-level model and by a set of interaction functions. Agent-based modelling (ABM) and simulation (ABS), and Cellular Automata (CA) are prominent examples of this decomposition approach for large-scale dynamic systems. CA can be considered as a simplified sub-class of ABM/ABS with strictly bounded interaction ranges, better suited and scaling for large-scale problems with a very high number of entities typically required to strength statistical quality. A simulation model is typically a simplification and abstraction of the complex real world that is characterised by the behaviour modelling of single entities (core cell elements of the simulation, e.g., an agent or a cell), the interaction between the elements, the number of elements relative to real world systems, and the variance of behaviour and interaction models. Mostly only an ensemble averaged model is used that is derived from real world observations and sensor data; individualism cannot be covered properly.

The combination of Machine Learning and simulation can improve model and simulation quality, i.e., there is according to:

1. Machine Learning assisted simulation improving the simulation model and quality [1];
2. Simulation assisted Machine Learning improving the prediction or classification model [1];
3. Emulation of the multi-agent behaviour model by an ML derived macro model (surrogate modelling) [2][3];
4. Model calibration using ML [1].

The central concept and novelty of this work is a ML-based ensemble estimator for aggregate observables learned from an incremental hybrid and domain-hierarchical MAS/CA simulation with the aim to improve real-word system time-series data prediction. The CA extension was chosen for efficiency and scaling reasons. A tight coupling of the simulation to real-world entities is an additional feature that ensures real-time updates of the simulation and incremental calibration of the simulation at simulation time, supporting crowd sensing and digital twin methodologies (but not stressed in this work). The work utilises and combines:

1. Hierarchical MAS-CA simulation incorporating real-world data for the parametrisation of the simulation world and agent modelling (digital twin concept) to predict future developments of system state observables from past data;
2. Hierarchical domain-specific modelling and decomposition (with respect to longitudinal and spatial scale);

3. Predictive modelling of time-series data using state-based ML models trained on real-world and simulation data;

The major issue with real-world coupled simulations and predictive machine modelling from simulation is the discrepancy of sensor data (input and output observables) collected in real and simulation domains. Typically, the simulation is almost inaccurate (and wrong) with respect to real world, but the sensor measuring is accurate and exact (all population entities can be accessed and measured directly). In contrast, to the real world domain where measurements are inaccurate and in many cases biased and distorted (or at least not representative), especially on the longitudinal scale. For example, considering traffic simulation, the sensors (counting and tracing traffic flows) are relatively accurate and representative in both domains. But in contrast, observations and simulation of pandemic situations disperse significantly in real and virtual world domains. Therefore, we have chosen the COVID19 pandemic use-case to demonstrate the issues with real-world coupled and data-based simulation and the deployment of predictive machine models derived from inaccurate and biased data.

The surrogate ML models should be able to predict future developments of aggregated macro-level observables from past data, e.g., the accumulative incidence rate of a pandemic situation. ML modelling is already applied in social science and ecological modelling [4]. The application of such learned surrogate models on inaccurate and distorted real-world sensor data will still result in inaccurate prediction results. To solve this issue, a sensor correction and calibration model must be derived by using correct simulation sensor data that is acquired by real-world measuring principles resulting in strongly biased and distorted data. To overcome the computational scaling problem due to a required high number of agents (beyond 100000) a hierarchical hybrid model of agents and contextual cellular automata simulating a lattice gas model is proposed. Fine-grained simulation is performed by spatial and temporal partitioning adapting models and simulations in consecutive time intervals based on changing environmental parameter space. As well simulation as prediction models can be updated incrementally by new measured data (longitudinal extension), e.g., by agent-based crowd sensing [5]. The following sections introduce the hybrid and hierarchical modelling and simulation model, showing results of time-series prediction on real-world data, and finally showing in comparison prelimenary results of time-series prediction from simulation data.

II. The Hybrid and Hierarchical Concept

The hybrid and hierarchical methodology addressed in this work combines MAS-ABS with supervised ML, and the ABS combines two levels of agent behaviour model complexity, state-based reactive agents with complex long-range interaction and CA cells with simple short-range interaction.

The CA is a sub-domain model of the agent model. The simulation framework consists of an agent simulator [6] that is capable to process computational and physical agents (first-level class agents) as well as CA worlds seamlessly. The domain-hierarchical MAS-CA modelling decomposes complex real worlds in simplified organised cell networks on micro-level, the ML methods are used to estimate system-level (ensemble) observables from sensors.

Computational agents are mobile software that can migrate between real- and virtual worlds and they are used for real-world data collection (including mobile crowd sensing) and for creating digital twins in the simulation, whereas physical agents are pure simulation objects that represent physical entities in the simulation world. To reflect spatial variance, the simulation world $\mathbb{S}$ is partitioned into spatial sub-domains $\mathbb{S}=\{S_{\mathrm{di}}\}$, associated with a MAS. Each domain is handled by an agent ag_{di} from the MAS that is a spatial and organisational representation of a large set of simple agents situated in a simplified CA world. Each CA represents a spatial region with a high number of interacting entities (e.g., humans). The MAS reflects the coarse-grained, the CA the fine-grained simulation model. The simulation model is composed basically of mobility, behaviour, and interaction of the observed entities.

First-level domain agents represent larger spatial domains (e.g., terrestrial units or entire cities) and interact with each other to simulate crowd flows, organisation, and networking across spatial domains. Each rectangular CA world *CW* connected to one domain agent consists of cells arranged on a regular two-dimensional grid that is partitioned into logical sub-domains (regions) *ld* associated with specific interaction behaviour and environmental constraints, e.g., living and working areas, $CW=\{ld_{\mathrm{j}}\}$, $ld_{\mathrm{j}}=\{cell \in A_{\mathrm{j}}\}$. The second-level class cell agents within the CA are modelled by a "mobile" data structure bound to one current cell in the CA world and processed by a cell activity function. The mobility of agents within the CA world is modelled with a randomised lattice-gas model by shifting the agent state spatially. A CA cell is occupied by one or no agent. Agents can access neighbouring cells (Moore neighbourhood) and can move to neighbour cells. The hybrid and hierarchical architecture is shown in Fig. 1. The main difference between first- and second level agents is the behaviour function. First-level agents bind each their own behaviour function, whereas second-level agents are represented by on shared behaviour function.

The aggregated data collected from simulation is used to train a surrogate machine model for time-series prediction. A state-based Long-Short Term Memory (LSTM) artificial neural network architecture was chosen for time-series prediction [7]. A LSTM network is able to predict a variable $\boldsymbol{x}$ for a future sample point $n+\Delta$ with past data $\{\boldsymbol{x}_1,..,\boldsymbol{x}_{\mathrm{n}}\}$.

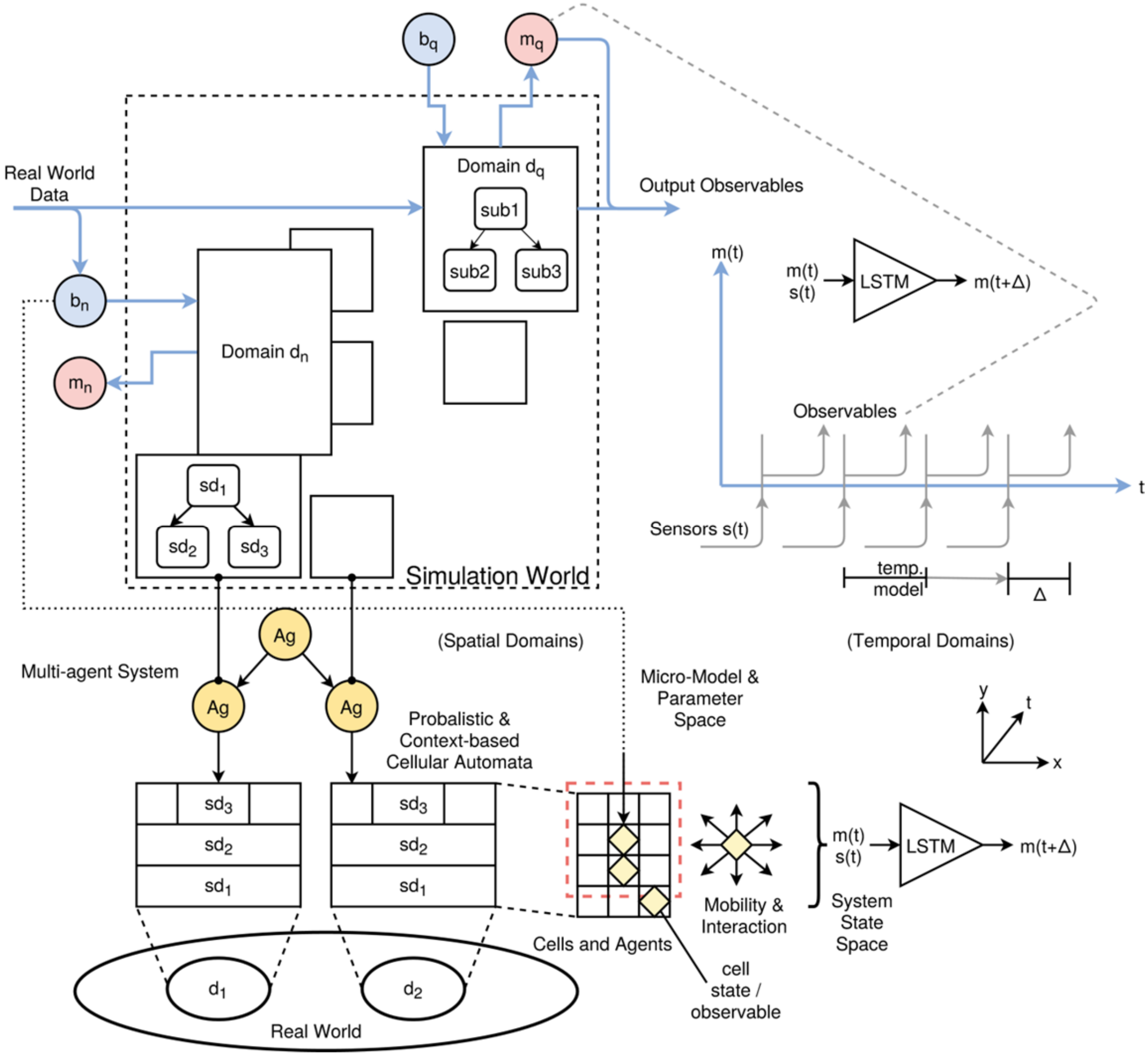

Figure 1. Hybrid simulation with domain-specific MAS-ABS combined with probabilistic CAs. Each CA (bottom) represents a simulation container with simple agents (diamonds) for a spatial domain d_i controlled by a domain agent (circle). Each CA world is partitioned in logical domains sd_j, too (bottom, left). Spatial domains are connected by the domain controller agents (middle,left). The LSTM (bottom, right) is trained with simulation data.

The real-world data is collected remotely by computational agents, e.g., performing WEB scraping to get environmental state information.

The next section demonstrates the novel hierarchical simulation approach for a pandemic use-case. The methodology can also be applied to other fields like traffic flow prediction and optimisation, logistic flows, and long-term prediction with respect to migration and segregation effects (social networking).

III. Use-Case: Pandemic Modelling and Prediction

We demonstrate the proposed hybrid and hierarchical simulation approach of real-world coupled MAS-CA simulation and longitudinal surrogate modelling for the forecasting of pandemic situations. Pure CA-based approaches were already applied to pandemic simulations [8]. This worst-use-case poses a highly unreliable and distorted measuring process, varying on longitudinal scale, and high dynamics based on micro-scale effects.

A. *Simulation and Surrogate Modelling*

Preliminary experiments were performed to investigate the accuracy and generalisation of a domain-specific prediction model from real data with a time-series prediction of infection observables using an LSTM ANN architecture. The input data are weekly infection cases rates of COVID19 pandemic data base from [9], and the output of this model $m_\Delta(t)$ is the prediction of Δ week ahead infection cases rates with respect to spatial domains and population age domains.

Each spatial domain is trained with its own model. Models are finally exchanged between spatial domains to test generalisation capabilities. The input data was used for seed conditions of the simulation, too.

The simulation world consists of 38 domains of territorial units (TU) of Germany (shown in Fig. 2) with the simulation parameters: Spatial centre location, population statistics, and mobility interconnects between neighbouring TUs. The agent base model is SIRD (susceptible-infected-recovered-dead) population classification. Each spatial domain is represented by its own domain model and parameter set and is simulated independently by a domain agent associated with Lattice Gas Probabilistic and Context-based CA (LG-PCCA), i.e., each domain region is a container for statistical moving and interacting agents, defined by a set of cross-section parameters. Longitudinal day-night cycle simulation is performed. The domain agent is responsible for sensor data acquisition, monitoring, and inter-domain interaction. The CA is partitioned into logical domains, e.g., home, work, outside, school, and culture/sports areas. Sub-agents given by data structures holding parameter and state variables located at cells represent people. Mobility of individuals is given by random walk (gas model), directed diffusion (context model), a mean velocity, and neighbouring and sub-domain constraints. Interaction (infection) is given by a dynamic cross-section and accumulator model, i.e., the integral of mobility and interaction cross-section. Perception and movement of an agent is limited to neighbouring cells (Moore neighbourhood); an agent can change it place (cell) either my moving to a free neighbour cells or by agent-pair swapping. Agent can migrate between different CAs via the domain agents, i.e., domains interact with each other by crowd flows (holiday, travelling, and business).

The agent behaviour model covers a wide range of behaviour parameter, i.e., age domains (child, youth, middle, elder people, etc.), activity domains (children, scholars, students, workers, non-workers, retired people), parameter sets (social networking factor, risk, mobility rate, protection, ...), networks (family, temporary groups), infection test coverage and strategies.

Output observables are accumulated monitored infection cases counts (with age distribution?) on daily basis (simulation) and on weekly basis (rates, real-world data). Input sensor variables (for simulation) are population and density distribution, age distribution, start infection count, social networking parameters, social cluster densities, mobility, opening status of domestic and private facilities, social restrictions, lethality, mortality, and the infection reproduction factor adapting the agent cross section and accumulator thresholds. Simulation is synchronised with real-world statistical pandemic data (accumulated, 1 week period). Agent-based WEB Scraping and Mining is used to sense environmental state variables, e.g., closed stores or schools, contact limitations.

Population representation by agents in a CA world is controlled by a domain agent. A typical population-agent scale ranges from 1:1000 to 1:100 depending on population density; if infection probability is low (<0.01), a higher density is required.

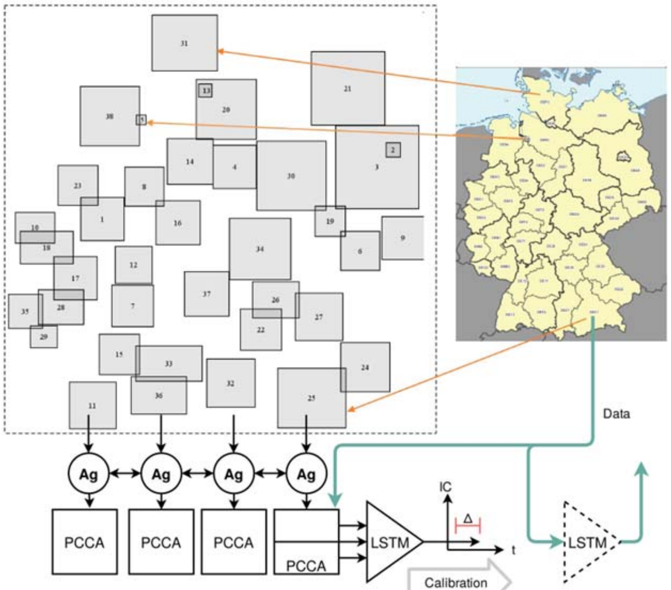

Figure 2. Simulation world partitioned into 38 TUs (NUTS level 2) mapped on 38 CA worlds (left), Cartesian coordinates, not ratio scaled. Size of CA grid is related to TU domain size and population density. Each CA produces data for surrogate modelling by an LSTM (bottom, right).

Model calibration is required for the simulation model (including time-scale calibration) from real to virtual world, and for the surrogate model from virtual to real world.

B. Preliminary Results

1) Real-Data Prediction

Raw real-world data from national RKI data base [9] was chosen to perform preliminary tests for predictive time-series modelling and simulation and to demonstrate the impossibility to predict future developments from past data. The data consists of weekly updated pandemic COVID19 infection cases (positive tests), i.e., infection rates, partitioned horizontally in 5 year age ranges, and vertically in TUs. The accumulated absolute infection cases, i.e., the number of infected persons, cannot be measured accurately and is not used here (in contrast to simulation).

The input sensor variables for the LSTM predictor is the infection rate (IR) grouped in four age ranges $\langle IR(A_{00}\text{-}A_{09}), IR(A_{10}\text{-}A_{19}), \mathrm{IR}(A_{20}\text{-}A_{59}), IR(A_{60}\text{-}A_{99})\rangle$. The output prediction variables (longitudinal extrapolation) are also the infection rates (IR), i.e., $\langle IR_{\Delta}(A_{00}\text{-}A_{09}), IR_{\Delta}(A_{10}\text{-}A_{19}), IR_{\Delta}(A_{20}\text{-}A_{59}), IR_{\Delta}(A_{60}\text{-}A_{99})\rangle$. The LSTM predictor has a layer configuration of [4,8,4] with 8 fully connected LSTM cells [10], a sigmoid transfer function, trained by single-sequence learning. Results of a playback experiment for one TU (Bremen) used to train and predict the infection rate development (Δ=4 weeks) is shown in Fig. 3.

The entire data set was used for training and prediction (playback from start to end). A very high accuracy of prediction results were achieved (error below 10%).

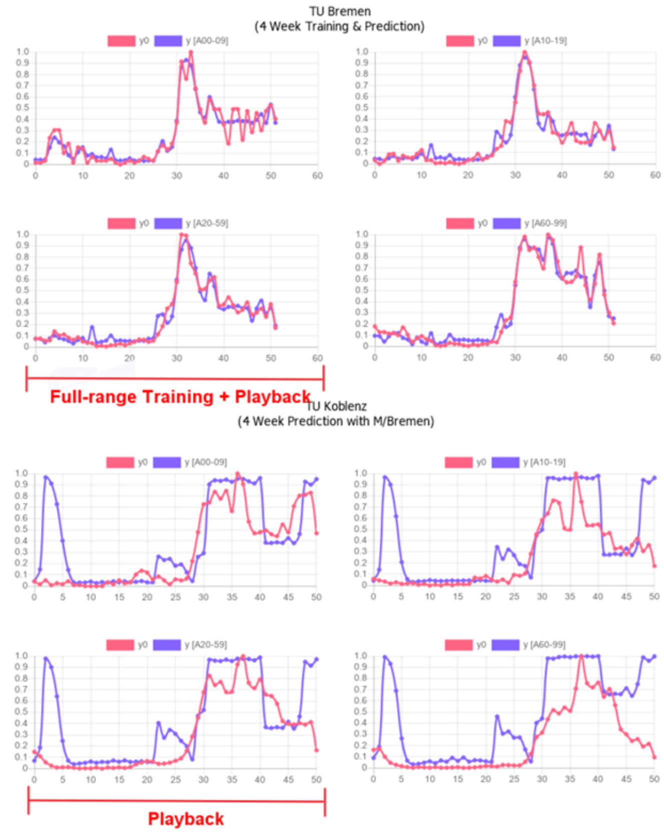

Figure 3. Playback of a domain-specific predictive system-level model derived from entire data series of longitudinal infection case rate development from real-world data (Top) Four week prediction (Δ=4w) for TU Bremen with respect to four population age ranges (Bottom) Model trained with TU Bremen data and applied to data from TU Koblenz [x: week, y: normalised infection case rate numbers, y0: reference data , y: predicted]

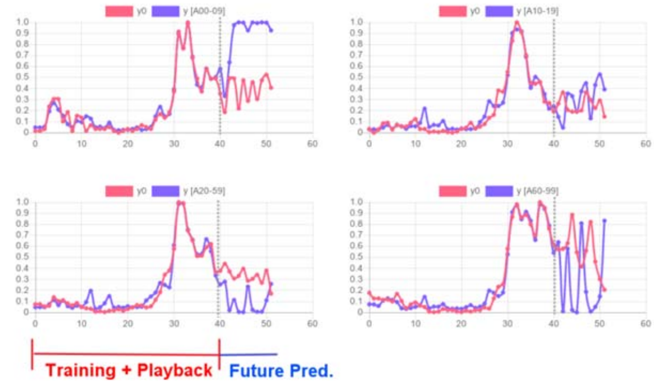

Figure 4. Future estimation of the same domain-specific predictive system-level modelling derived from the half of the data series (cut-off at week 40) of longitudinal infection case rate development from real-world data (Top) Four week prediction (Δ=4w) for TU Bremen with respect to four population age ranges [x: week, y: normalised infection case rate numbers, y0: reference data , y: predicted]

But if a model trained for one domain is applied to data of another domain the prediction shows very high prediction errors and peaks, shown for the TU Koblenz. This result shows the requirement for domain-specific simulation and surrogate modelling, and that the surrogate prediction model learned some longitudinal data structure that is not related to any pandemic model and behaviour (black box pitfall)! But the aim of the predictive modelling of aggregate variables is future prediction. To illustrate the impossibility of long-term future prediction the experiment was repeated but with a training only using the first half data set, show in Fig. 4. The predictor function diverges quickly after the last trained point and tends to oscillate.

2) *Simulation and Prediction*

The simulation was performed with the probabilistic and contextual CA representing one artificial TU domain. The CA was spatially partitioned into 6 logical regions, shown in Fig. 5 (a): Home, outside, working area, shopping area, schools, and culture/sports. Agents that want to change the region always pass the centred outside region. Each region is defined by a mobility scaling factor. The agent movement is either randomised or directed. The simulation addresses day-night cycles.

All agents return to their root home position at night. Fractions of agents migrate to different regions at different time slots. In contrast to the real-world prediction, the normalised accumulated infection case number is the aggregated system state variable that is measured and predicted by the trained surrogate model. The sensor input variables is the infection count IC (full age distribution) with an auxiliary variable, the derivation: $\langle IC, \delta IC/\delta t\}$. The output prediction variable is again IC_{Δ}. The LSTM model has a layer configuration of [2,7,7,1] with two × 7 fully connected LSTM cell layers [10] (each cell with `memoryToMemory`, `inputToOutput`, and `inputToDeep` gates control), a sigmoid transfer function, and was trained periodically with multi-sequence learning.

A high prediction accuracy for Δ=4 (arb. units) was achieved in playback mode (i.e., full-range training and replay prediction), as shown in Fig. 5 (b). But in contrast to the highly distorted and temporally biased real-word data predictions (with useless results), future prediction (of a second infection raise) can be predicted with high accuracy just by using past date only (cut-off point is here 30), as shown in Figure 5 (c). To conclude, the surrogate modelling of the CA/MAS system poses a high degree of generalisation (on the longitudinal scale), in contrast to the same model trained on real-world data.

The seed of the simulation was a population of 600 agents with a share of 5% infected agents. The cell placement is randomised. Some simulation runs (with same seed parameters) did not show a pandemic development. Without the (dependent) auxiliary variable, the prediction model could not be trained (no training convergence).

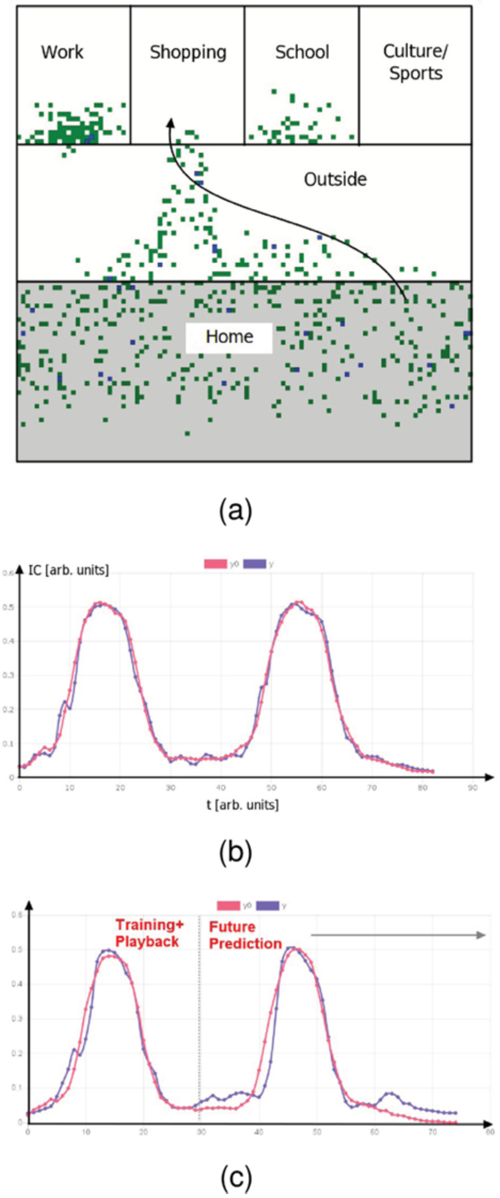

Figure 5. (a) CA simulation world with logical regions (b) Playback of predictive modelling of longitudinal infection cases development; full-range training (Δ=4 arb. units) from simulation data (c) Partial-range training and future prediction [x: time (arb. units), y: normalised infection case numbers, y0: reference data , y: predicted] Using the Template

In the real-data prediction case, there were already four correlated input variables (age range variables). There is still no longitudinal updated simulation (with real world data) and surrogate model calibration. The time scale is artificial and arbitrary.

IV. Conclusion

The acquisition of real-world sensor data and the derivation of time-dependent system state observables can be a challenge. The measurement and the test sample distribution of real-world sensors are often distorted and biased, or sensor variables are nor accessible (on spatial and/or longitudinal scale). Pandemic situations are prominent examples. Simulations rely on accurate data for simulation world parametrisation and model calibration. Time-series prediction of system state variables is of high relevance for political and domestic decision making processes. We evaluated time-series prediction on real data from a RKI data base containing infection cases data rates of the COVID19 pandemic (54 weeks) using a LSTM neural network. Firstly, we showed a high prediction accuracy on the longitudinal axis (4 week prediction) in playback mode, but very low accuracy on spatial scale, i.e., by applying a trained model to another spatial domain, and for future predictions. Secondly, we concluded that the trained model do not base on any reasonable pandemic model and that the original RKI data base contains highly distorted and biased data (especially on longitudinal scale). In the next step we introduced a multi-domain hybrid and hierarchical agent-cellular automata simulation approach. The CA was partitioned into logical regions and agent mobility and interaction bases on a constrained lattice-gas model. The data collected from the simulation was again used for time-series prediction using a LSTM-ANN providing a surrogate model for the system state variable infection cases of the MAS-CA simulation. Again, a high accuracy for playback and forward predictions was achieved. But the simulation model cannot actually be applied to real-world data, and sensor calibration addressing longitudinal, measuring, and pandemic parameters have to be performed in future work to achieve a transfer to real-world data prediction. Finally, domain-specific variance must be improved and derived from real-world data. The surrogate modelling of the MAS-CA system poses a high degree of generalisation (on the longitudinal scale), in contrast to the same model trained on real-world data.

References

[1] L. von Rueden, L., S. Mayer, R. Sifa, C. Bauckhage, J. Garcke1, “Combining Machine Learning and Simulation to a Hybrid Modelling Approach: Current and Future Directions”, IDA 2020, LNCS 12080, 2020.

[2] C. Angione, E. Silverman, E. Yaneske, “Using Machine Learning to Emulate Agent-Based Simulations”, arXiv preprint arXiv:2005.02077. 2020.

[3] S. van der Hoog, “Surrogate Modelling in (and of) Agent-Based Models: A Prospectus”, Comput. Ecom., vol. 53, 2019.

[4] F. Recknagel, “Applications of machine learning to ecological modelling,” Ecological Modelling, vol. 146, 2001.

[5] V. C. Raykar et al., “Learning From Crowds,” Journal of Machine Learning Research, vol. 11, 2010.

[6] S. Bosse, U. Engel, “Real-time Human-in-the-loop Simulation with Mobile Agents, Chat Bots, and Crowd Sensing for Smart Cities”, Sensors (MDPI), 2019, doi: 10.3390/s19204356

[7] S. Saadatnejad, M. Oveisi, M. Hashemi, “LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices”, IEEE Journal Of Biomedical and Health Informatics, 2019.

[8] S. Ghosh, S. Bhattacharya, “Computational model on COVID-19 Pandemic using Probabilistic Cellular Automata”, SN COMPUT. SCI. 2, 230, 2020

[9] Robert-Koch Institute, Germany, “Infection case data base”, accessed on 19.3.2021, https://survstat.rki.de/Content/Query/Create.aspx

[10] Neataptic, https://github.com/wagenaartje/neataptic, last accessed 1.7.2021