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

Challenges and Pitfalls

PD Dr. Stefan Bosse

sbosse@uni-bremen.de

University of Bremen, Dept. Mathematics and Computer Science, Bremen, Germany

Introduction

Topic

Modelling and Simulation of *complex dynamic systems* like pandemic outbreaks or traffic flows in cities

Derivation of **macro-level** aggregate variables (observables) from analytical models and simulation on **micro-level**

Prediction of time-dependent aggregate variables by combination of **Machine Learning and Simulation**

Coupling of simulation with real-world environments including **digital twin** methodology

Introduction

Issues

A high variance on entity **micro-level** and unknown or incomplete entity interaction models

Lack of sensor and model **calibrations** ⇒ affecting functional modelling and simulation

Accessibility of real-world sensor data

High dimensionality, size, and distortion (bias, test coverage, skewness of distributions) of real-world data

High number of micro-level entities in simulation (for statistical strength) ⇒ Computaionial complexity and time

Modelling Levels

Macro level

Description of the system behaviour with aggregate variables (observables) of large-scale ensembles computed typically with analytical functional models deriverd from real-world observations and experiments ⇒ System Level, **Statistical Methodologies**

Micro level

Modelling of complex systems by micro-level entity behaviour and interaction with observation of macro-level system behaviour ⇒ Derivation of aggregate variables by simulation and test (ensemble emergence) ⇒ **Multi-Agent Modelling and Simulation**

Meta level

Description of micro- or macro-level behaviour not based on direct observations ⇒ **Big Data Analysis**

Methodologies

Functional Modelling

Empirical inference from experiments with induction methodologies. Commonly, a mathematical function $f(x)$: $x \rightarrow y$, where x is an input variable (stimulus, state, feature of the environment), and y is a response to this stimulus.

Simulation

Simulation can be used to probe system observables from a large set of interacting micro-level entities

Static Data Modelling

Supervised Machine Learning is used to approximate and develop the mapping function $f(x)$: $x \rightarrow y$

Dynamic Data Modelling

Machine Learning is used to predict the future development of time-series data and observables y

Modelling and Simulation

The combination of Machine Learning and simulation can improve model and simulation quality in different ways:

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

Hybrid Methodology

! 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.

Conceptual Methodology

1. **Hybrid MAS-CA simulation** featuring:
   - Agent-based Simulation
   - Cellular Automata
   - Incorporating real-world data for the parametrisation of the simulation world and agent modelling
   - Digital twin concept
   - Prediction of future developments of system state observables from past data;
2. **Hierarchical domain-specific simulation** and decomposition (with respect to longitudinal and spatial scale);
   - Agents represent spatial domains and controlling CA;
   - Cells of CA represent individual entities in spatial domain

Conceptual Methodology

3. **Predictive modelling** of **time-series data of aggregate variables** using **state-based ML models** trained on real-world and simulation data.
    - Simulation augments or replaces real-world data
    - Augmented data is used to train predictive models, e.g., infection rate development
    - Sequential Time-series predicitive models for surrogate observable variables with auxiliary sensor variables
    - Models are applied to real-world data predicting finally real-world observables!

Agent-based Simulation

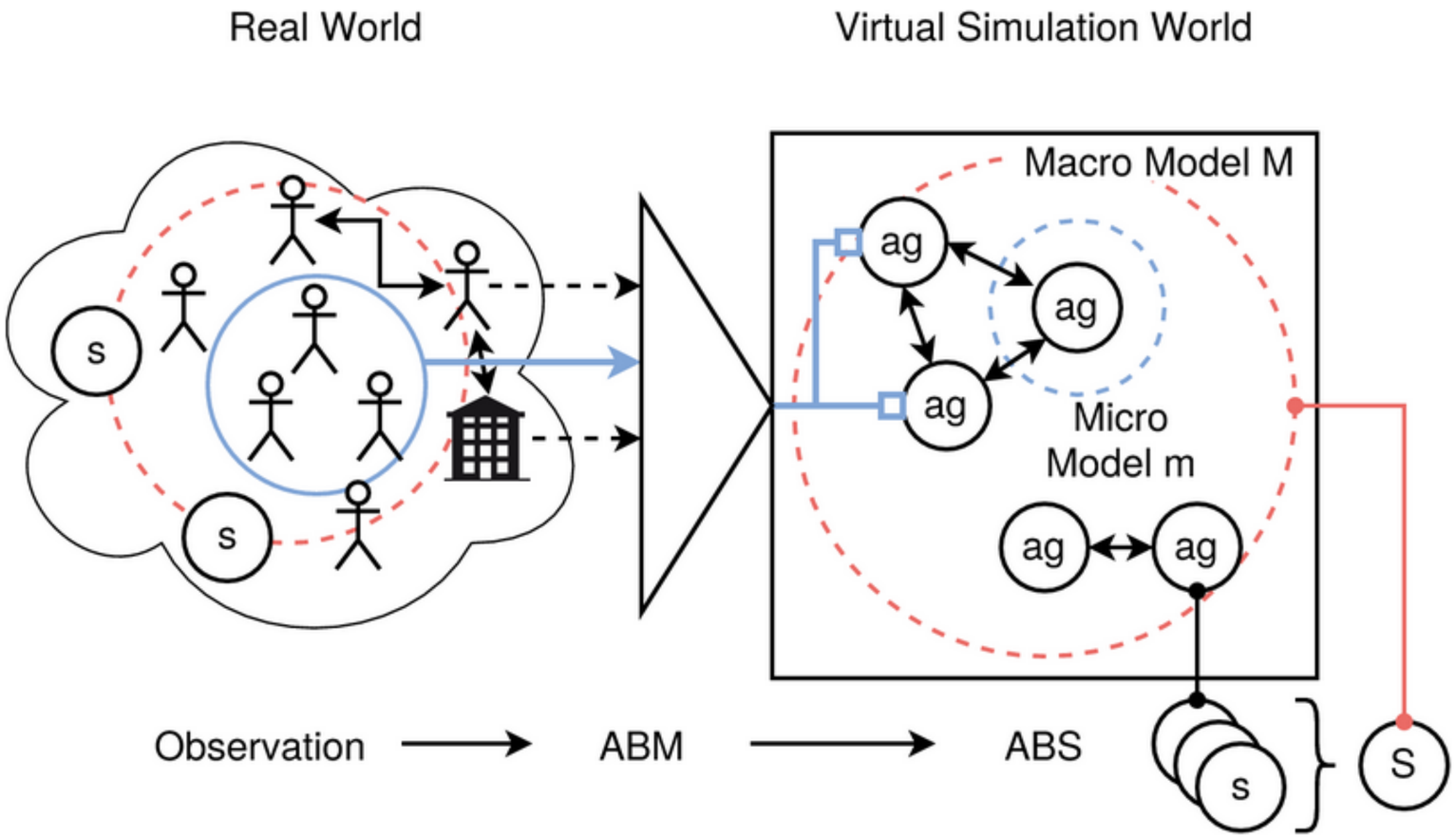


Fig. 1. Simulation on micro-level with computated macro-level observables using sensors (S) and test probing

Models and Architecture

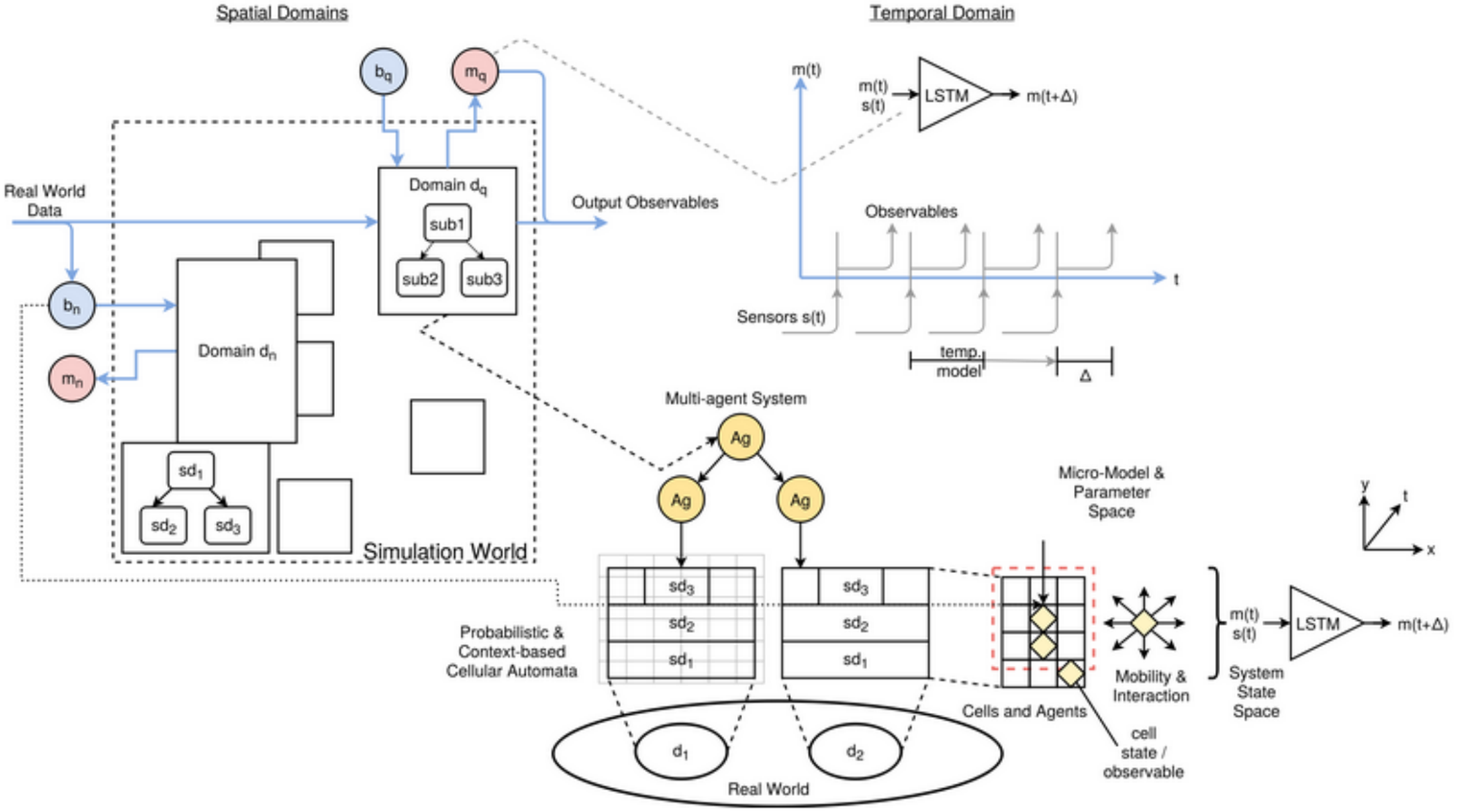
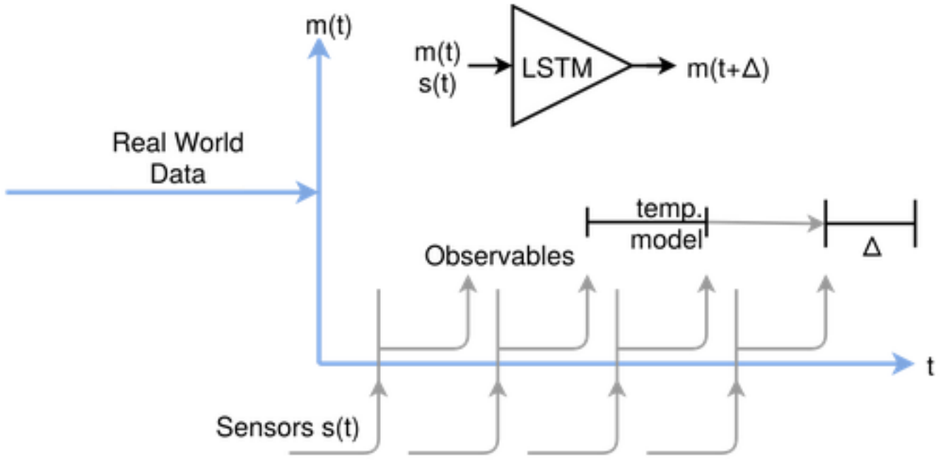


Fig. 2. The hybrid overall architecture and methodology for predictive surrogate modelling of time-dependent system observables (here infection rates of a pandemic situation) using state-based machine learning models

Machine Learning

- 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
- A LSTM network is able to predict a time-dependent variable $\vec{x}(n)$ for a future sample point $n+\Delta$ with past data $\{\vec{x}(1),..,\vec{x}(n)\}$.



Case-study: Pandemic Outbreak

Goal: Future prediction of system observable infection rate (or cases) from past data with a machine model trained with real and/or simulation data

! The main issue with pandemic data bases is the high bias and distortion of sampled population data (infection cases) due to uncalibrated sensors and unknown test strategy (cross section)

! Developing time-series prediction models for pandemic observables from population data on a long-term time scale is nearly impossible!

Simulation Model and Partitioning

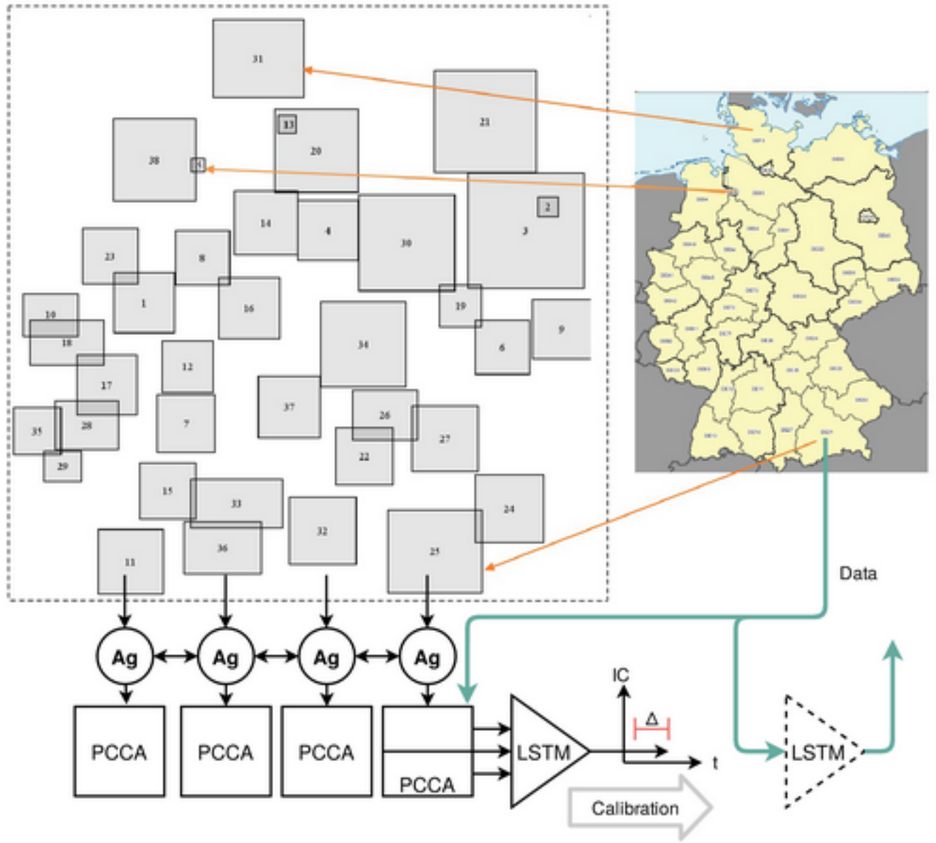


Fig. 3. Simulation world (Germany) partitioned into 38 TUs (NUTS level 2) mapped on 38 CA worlds, Cartesian coordinates, not ratio scaled. Size of CA grid is related to TU domain size and population density.

Cellular Automata

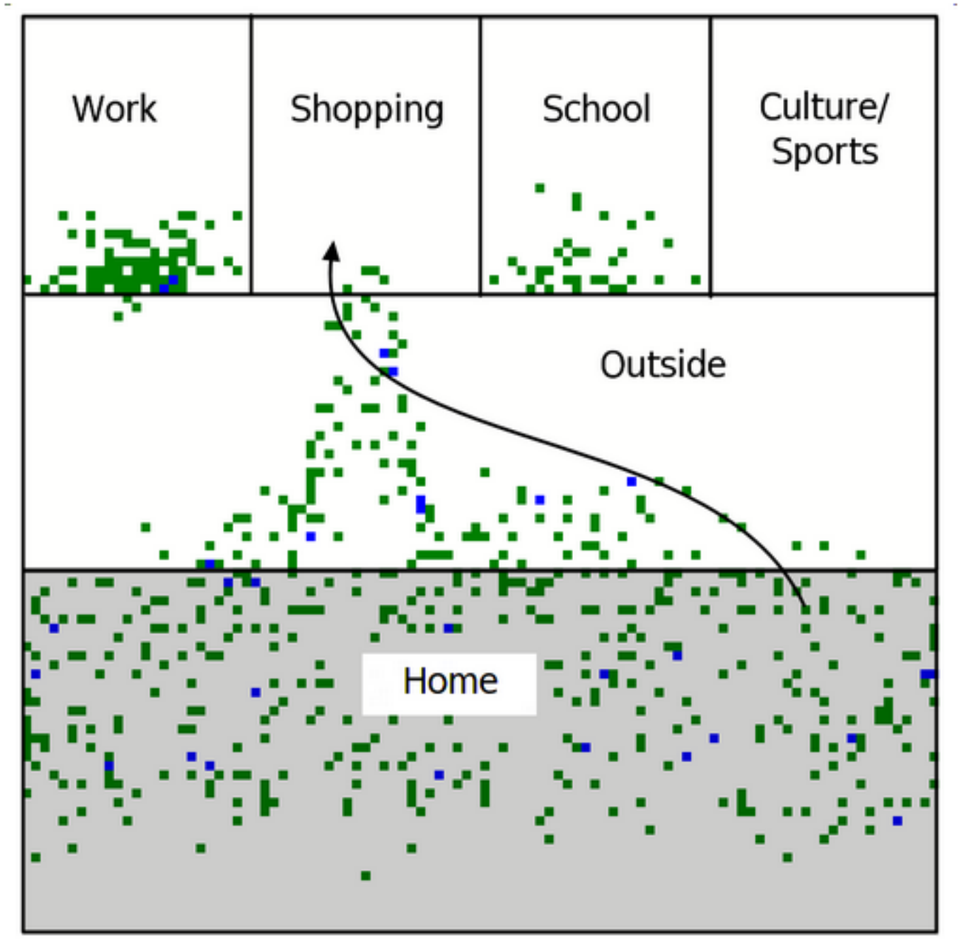


Fig. 4. Each TU is simulated with a CA partitioned in sub-domain areas

Results 1: Prediction on Population Data

- Time-series prediction by a machine learning LSTM model trained with population data and predicting the future development of infection rate
- There is a set of independent predictive models $M=\{m_d\}_{d=1}^{38}$, one for each terrestrial unit domain (TU)
- Data: Population data from Robert Koch Institute (uncalibrated, as-is data)
  - Daily submitted infection cases of COVID19 virus infections
  - Time range: about 52 Weeks
  - Sub-data split: TU (38) and age distribution in 5 years intervals

Evaluation strategies

A. Training of a model in one TU_i over full longitudinal range (52 weeks)
B. Application of model to TU_i over full longitudinal range (playback)
C. Training of a model in one TU_i over partial longitudinal range (40 weeks)
D. Application of partially trained model to TU_i over full longitudinal range (playback)
E. Application of model TU_i to another domain TU_j

Fig. 5. Training of predictive time-series model for observable infection rate for one TU (Bremen) over full longitudinal range (52 weeks); y_0: original data, y: predicted data

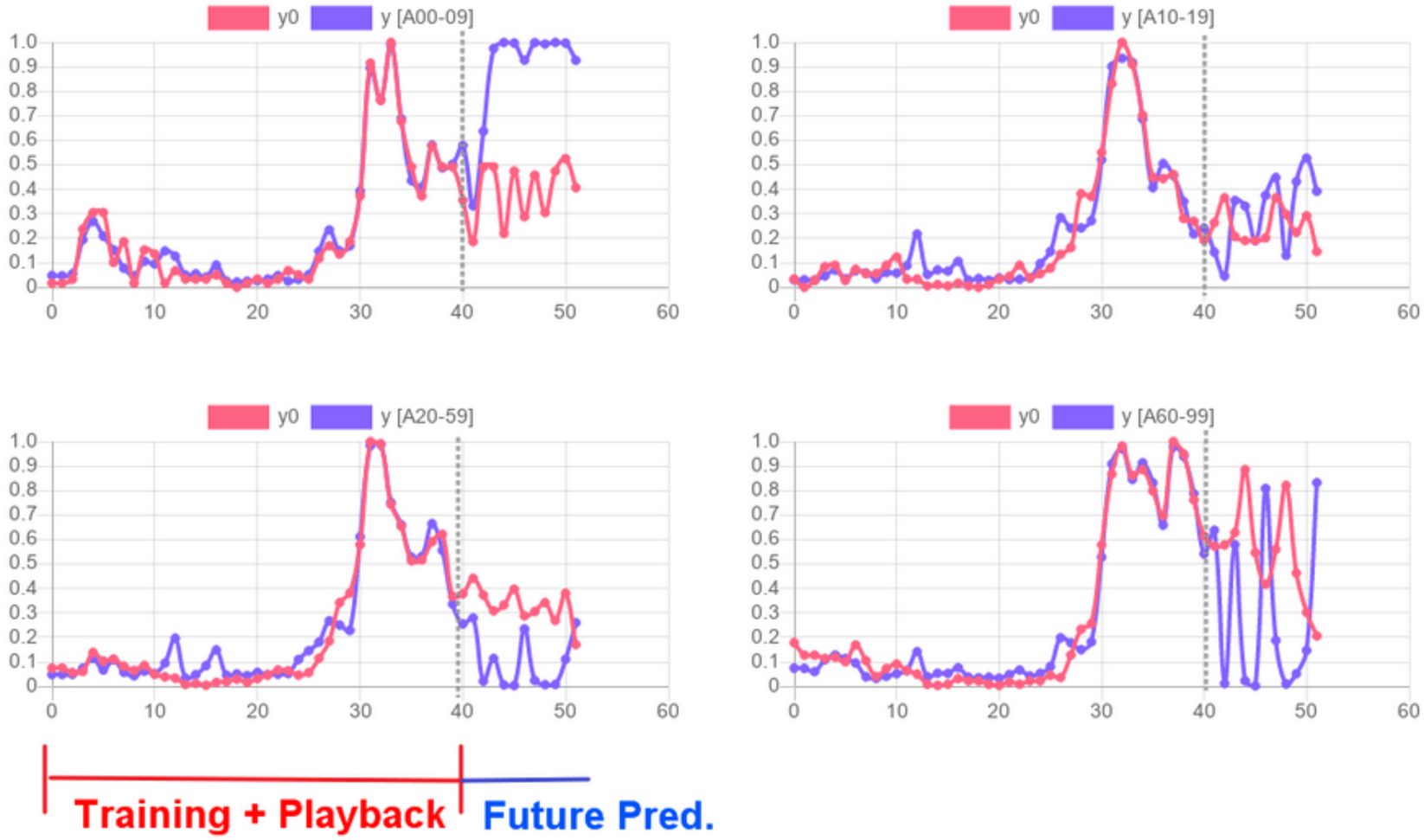


Fig. 6. Training of predictive time-series model for infection rate observable for one TU (Bremen) over partial longitudinal range (40 weeks); y_0: original data, y: predicted data

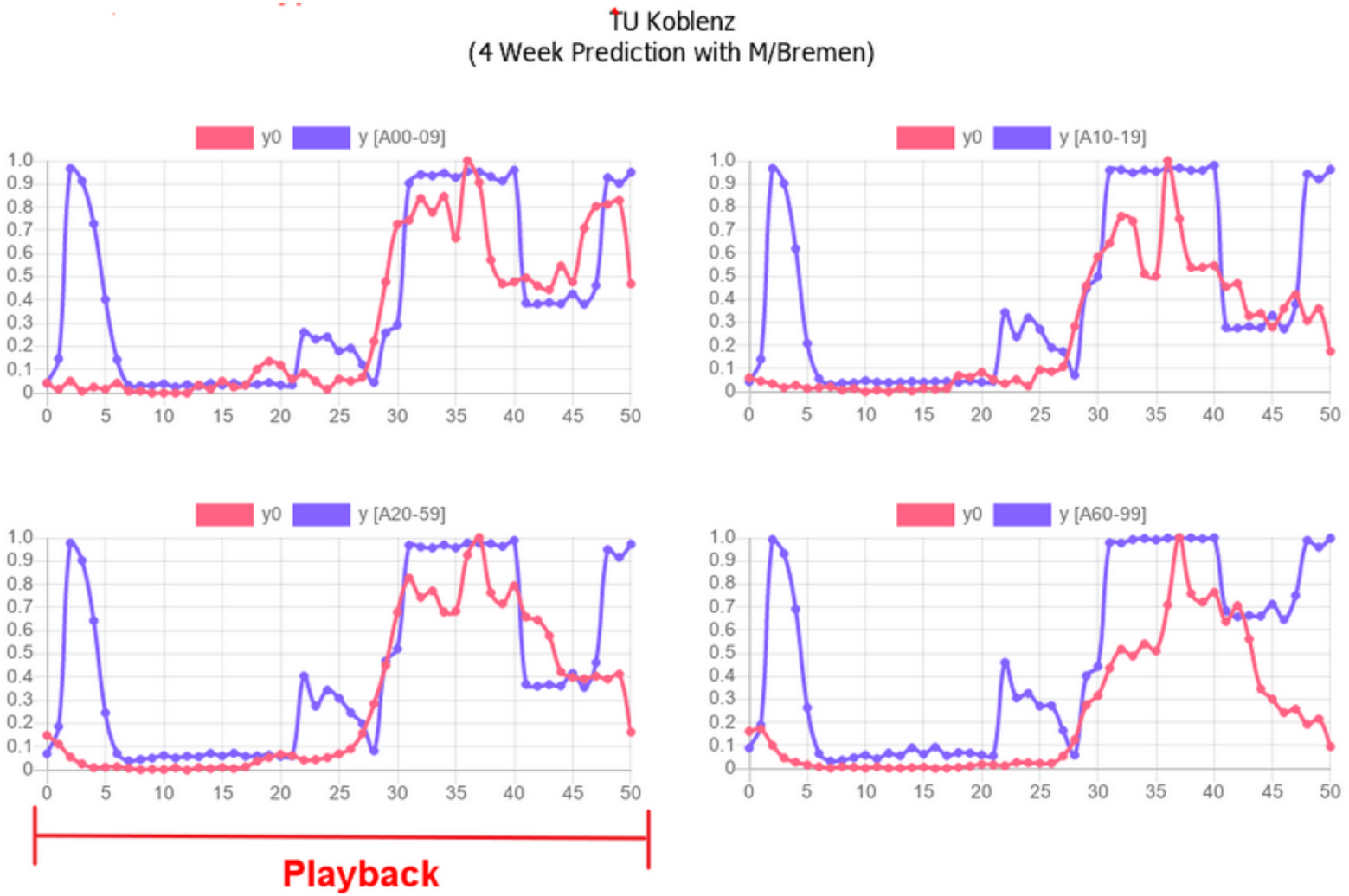


Fig. 7. Application of predictive time-series model TU Bremen for infection rate observable to TU Koblenz; y_0: original data, y: predicted data

Results 2: Prediction on Simulation Data

- Time-series prediction by a machine learning LSTM model trained with simulation data and predicting the future development of infection cases
- There is a set of independent simulations $U=\{u_d\}_{d=1}^{38}$ and predictive models $M^U=\{m_d^u\}_{d=1}^{38}$, one for each terrestrial unit domain (TU)
- Each TU is represented by an agent controlling a Cellular Automata (CA):
  - About 500-1000 cell agents / TU simulation
  - Agent behaviour (mobility, interaction, health, ..) is given by an average standard model (classes: worker, kids, family, ..) and individual behaviour by digital twin creation
  - Measurement of infection case observable

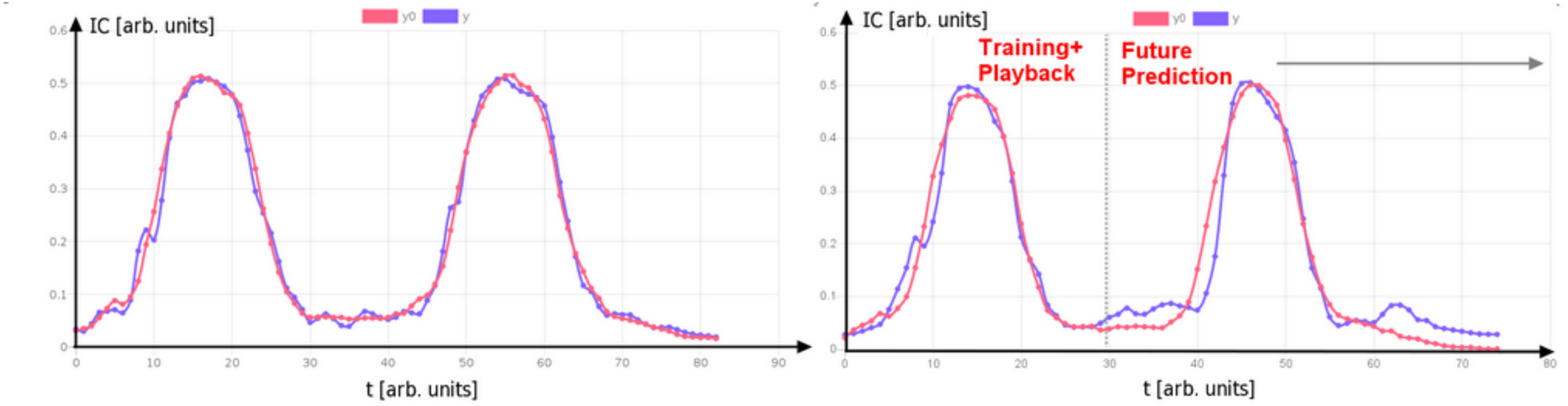


Fig. 8. Training of predictive time-series model for infection rate observable for simulated TU (Bremen) over full and partial longitudinal range (70/30 weeks); y_0: original data, y: predicted data

Summary

Prediction of time-dependent population observables from domestical population data mostly fails due to uncalibrated and distorted sensors

Simulation of large-scale populations like in pandemic situations is a challenge due to high number of entities and high degree of domain-dependent and individual behaviour variance typically not covered

Digital Twin concepts can improve simulation by introducing micro-level variance

Surrogate modelling by using simulation data replace computational complex agent-based simulations

A hybrid simulation model of agent-based and probabilistic cellular automata methodologies is a good trade-off

Spatial domain paritioning can further improve prediction accuracy

Questions and Comments are welcome!

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

Challenges and Pitfalls

PD Dr. Stefan Bosse

sbosse@uni-bremen.de

University of Bremen, Dept. Mathematics and Computer Science, Bremen, Germany