Shapley Value based Variable Interaction Networks for Data Stream Analysis

Eurocast 2022 // 2022-02-23

Jan Zenisek, Sebastian Dorl, Stephan Winkler and Michael Affenzeller

Heuristic and Evolutionary Algorithms Laboratory

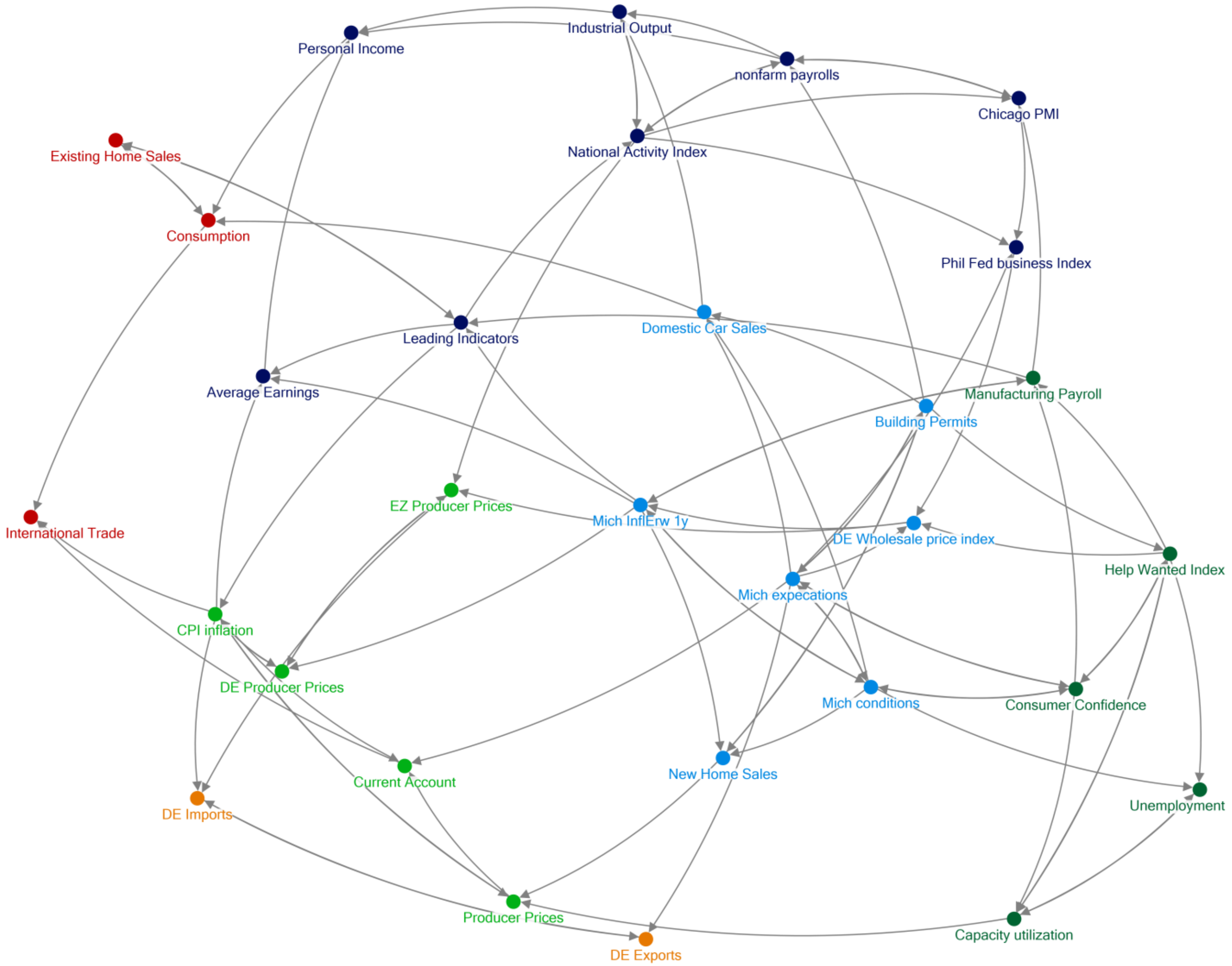
University of Applied Sciences Upper Austria

Institute for Symbolic Artificial Intelligence

Johannes Kepler University Linz

Part I – Method

Variable Interaction Network (VIN)



= directed, weighted graph

Nodes: variables

Edges: impact of variables on others

[1] Kronberger et al. *Genetic Programming: Current Trends and Applications in Computational Finance*, Nova Science Publishers, 2013

[2] Hooker, Giles. *Discovering additive structure in black box functions.* Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

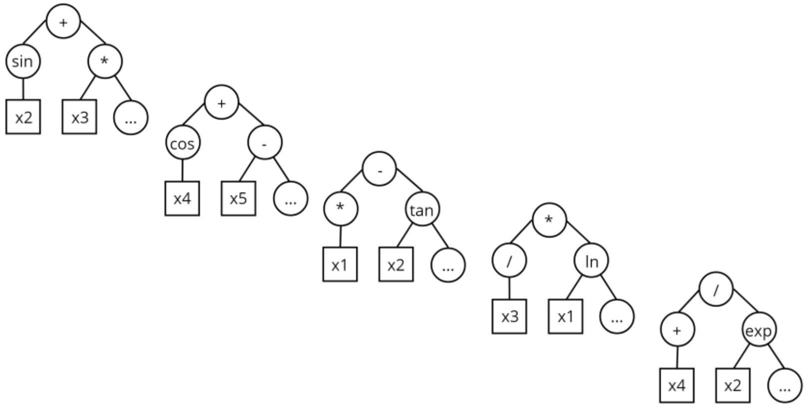
Part I – Method

Variable Interaction Network: Modeling

1. Alternate targets & inputs

target | input variables

x1	x2	x3	x4	x5
1.1	1.4	1.7	1.3	1.2
1.2	1.3	1.4	1.5	1.3
1.2	1.1	1.4	1.9	1.4
1.4	0.9	1.2	1.3	1.4
1.2	1.2	1.6	1.2	1.7



2. Calculate variable impacts

For all models do: For all inputs do:

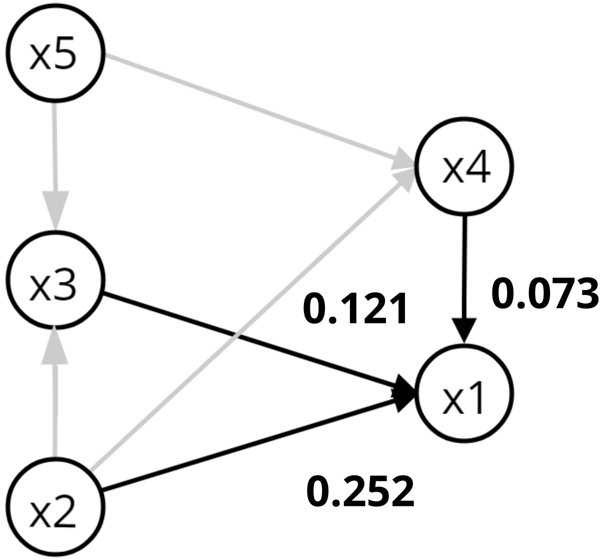
2.1 Remove variable info, e.g. shuffle values

2.2 Recalculate model error, e.g. R^2
→ Error increase = impact

Example calculation for model target=x1:

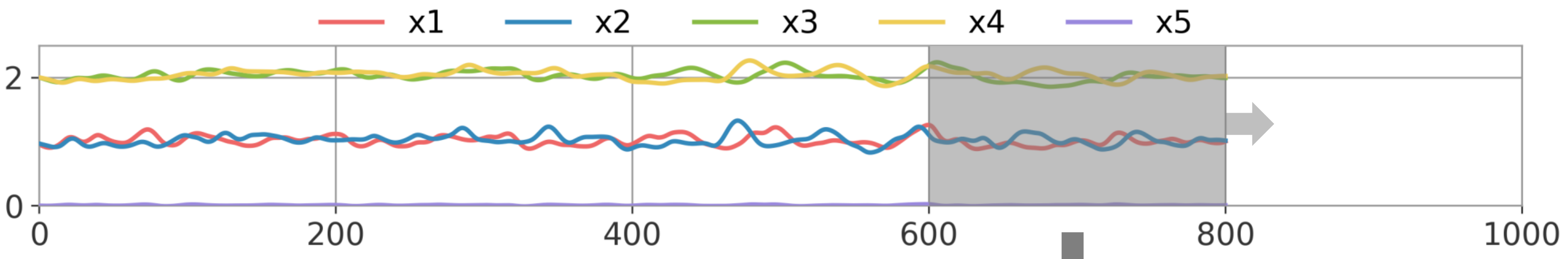
Variable	Impact for x1
x2	0.252
x3	0.121
x4	0.073
x5	0.037

3. Create network



Part I – Method

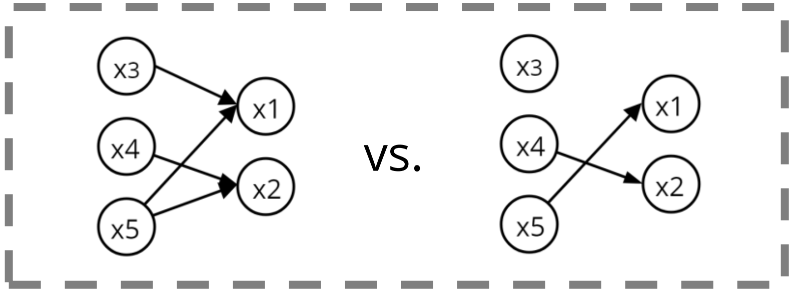
Variable Interaction Network: **Evaluation**



VIN comparison

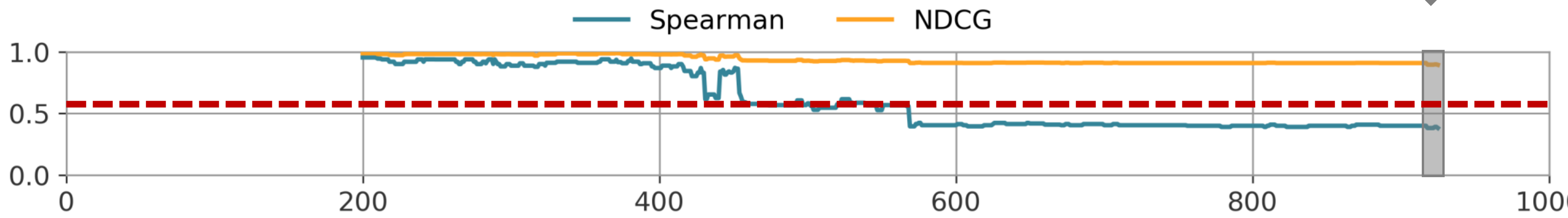
- **Spearman**: Spearman‘s Rank Correlation
- **NDCG**: Normalized Discounted Cumulative Gain (Kekäläinen, 2002)

Initial VIN



Re-computed VIN

Network Similarity



Threshold

Part I – Method
Variable Interaction Network

Model Type Characteristics

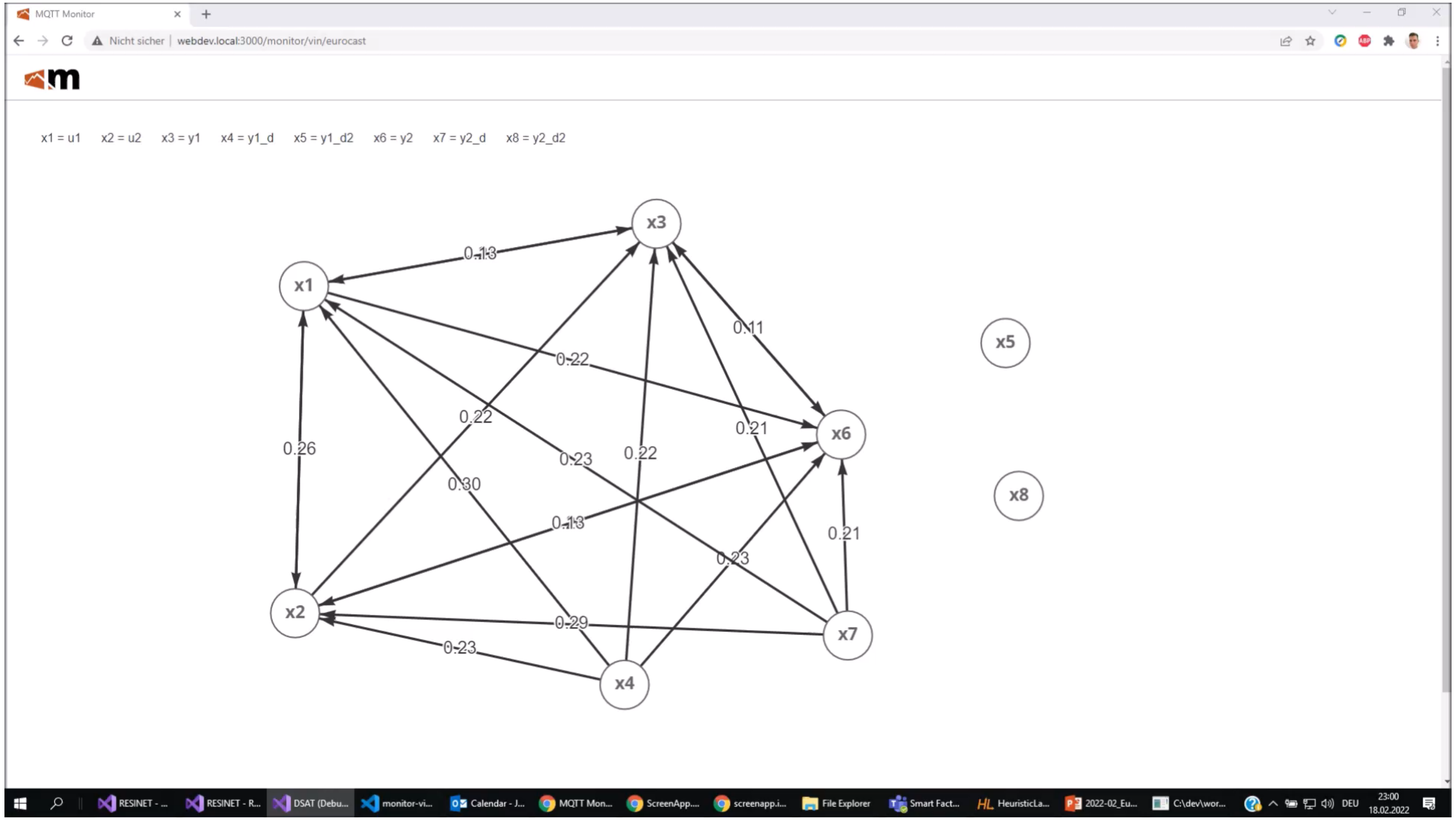
- ↖ Enables holistic system analysis …also on streaming data [3]
- ↖ Agnostic to the regression algorithms / models used as base
- ↖ Fast to create, once regression models are built

- ↙ Infeasible for high-dimensional data without pruning
- ↙ Non-deterministic modeling & evaluating causes network alternatives

[3] Zenisek, J., Kronberger, G., Wolfartsberger, J., Wild, N., & Affenzeller, M. Concept Drift Detection with Variable Interaction Networks. In International Conference on Computer Aided Systems Theory (pp. 296-303). Springer, Cham, 2020.

Part I – Method

VIN Evaluation Instability (Video)



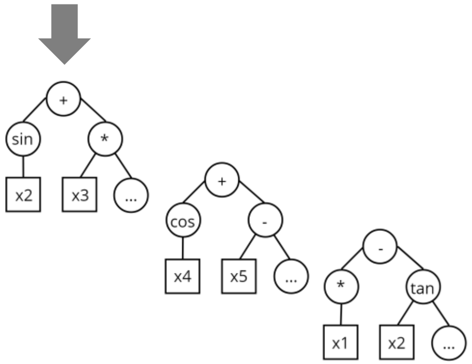
Part I – Method

VI Network: Modeling (cont.d)

1. Alternate targets & inputs

target — input variables

x1	x2	x3	x4	x5
1.1	1.4	1.7	1.3	1.2
1.2	1.3	1.4	1.5	1.3
1.2	1.1	1.4	1.9	1.4
1.4	0.9	1.2	1.3	1.4
1.2	1.2	1.6	1.2	1.7



2. Calculate variable impacts

For all models do: For all inputs do:

2.1 Remove variable info, e.g. shuffle records

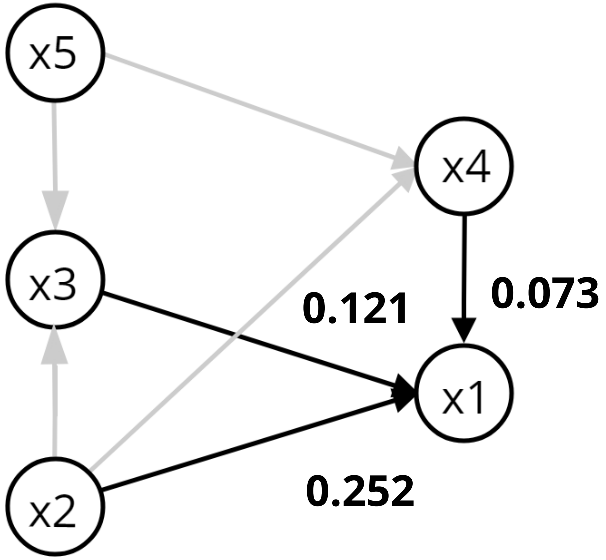
2.2 Recalculate model error, e.g. R^2
→ Error increase = impact

= Permutation Feature Importance [3]

Example calculation for model target=x1:

Variable	Impact for x1
x2	0.252
x3	0.121
x4	0.073
x5	0.037

3. Create network



[4] Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)

Part I – Method

Shapley Value based Variable Impact [5]

Feature contribution to the difference between the actual prediction and the mean prediction

1. Create all possible feature coalitions with and without the targeted feature
2. Calculate „actual prediction – mean prediction“ difference for each coalition
3. Average differences between coalitions with and without the targeted feature

- Coalition game theory (solid mathematical foundation)
- Local and model agnostic
- Computationally expensive

Customization:

- Use of nmse and impact threshold
- Average absolute shapley value of current set (global analysis)
- Normalize outcome

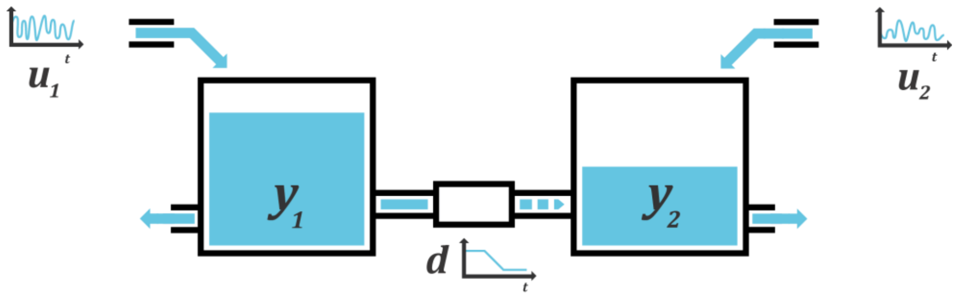
[5] Shapley, Lloyd S. “A value for n-person games.” Contributions to the Theory of Games 2.28 (1953): 307-317

Part II - Experiments

Benchmarking Problems for VINs

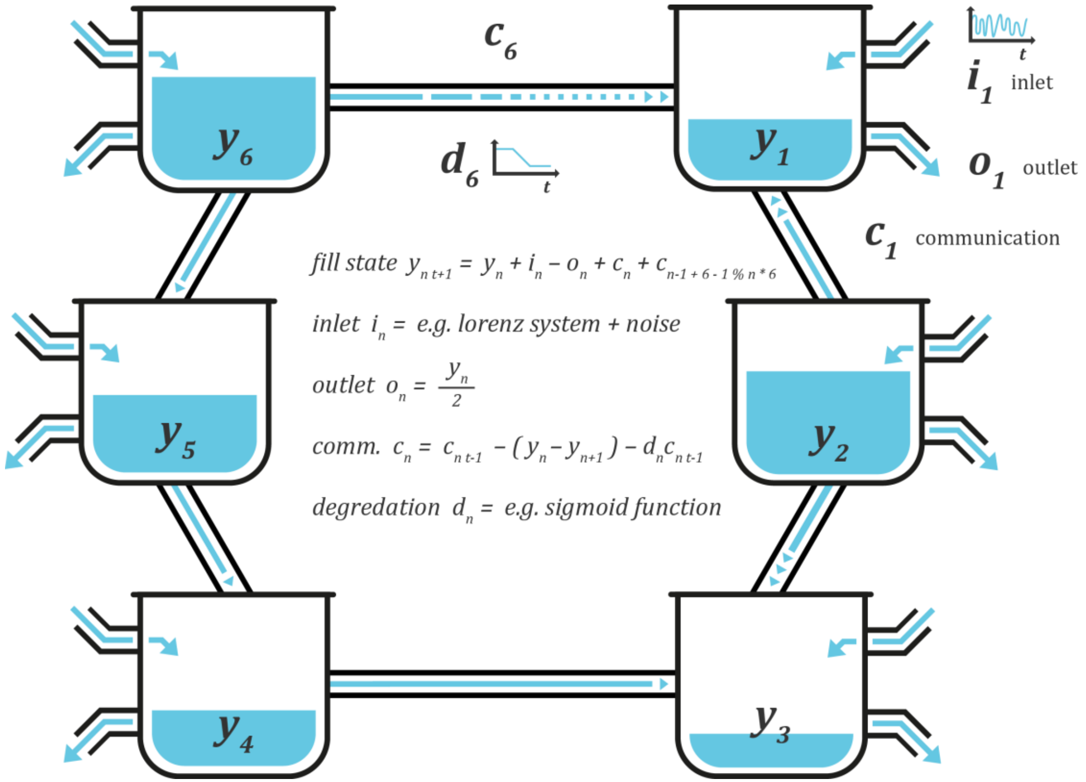
Communicating Vessels

- Inherently stable
- *d* = introduced drift (hidden) to simulate gradually clogging communication path



[3] *Zenisek, J., Kronberger, G., Wolfartsberger, J., Wild, N., & Affenzeller, M. Concept Drift Detection with Variable Interaction Networks. In International Conference on Computer Aided Systems Theory (pp. 296-303). Springer, Cham, 2020.*

Circular Connected CVs (new)



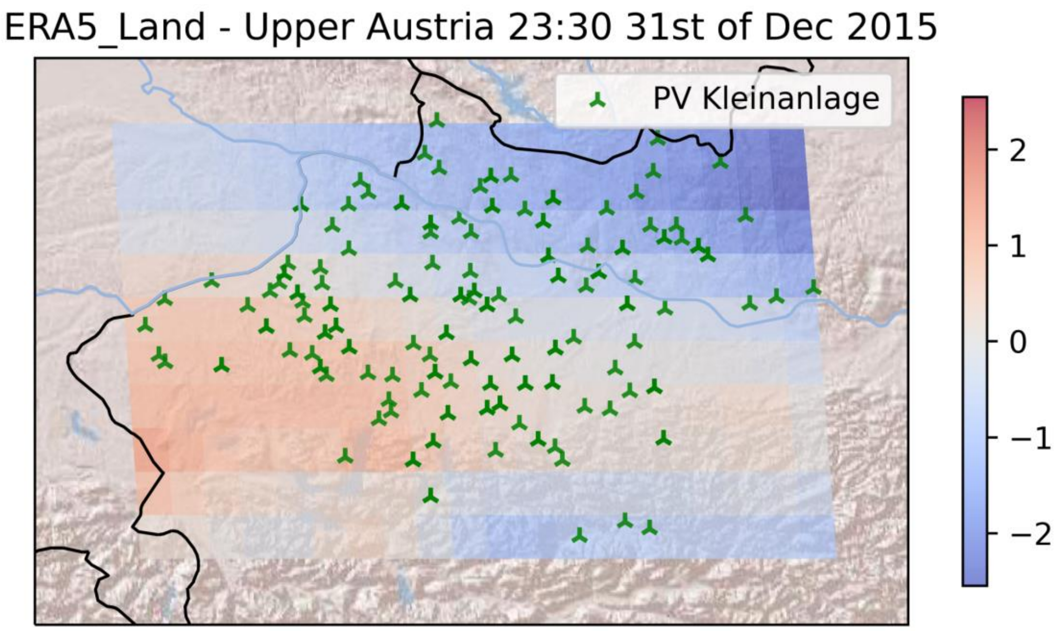
Part II - Experiments

Real World Problem: Photovoltaic Power Production

Available Data:

- 190 privately owned photovoltaic systems
- including battery packs
- Recordings from 2014 to 2019
- Recording interval: 5 min, ~100 Mio. data rows

- Constant parameters: geolocation, manufacturer, capacity,...
- Measured features: PV production, power consumption, grid input/output, battery charge, discharge, SOC



Main goals:

- Prediction models for power production & consumption
- **Network resilience analysis**

Network Resilience Simulations (Resinet)

What happens if…

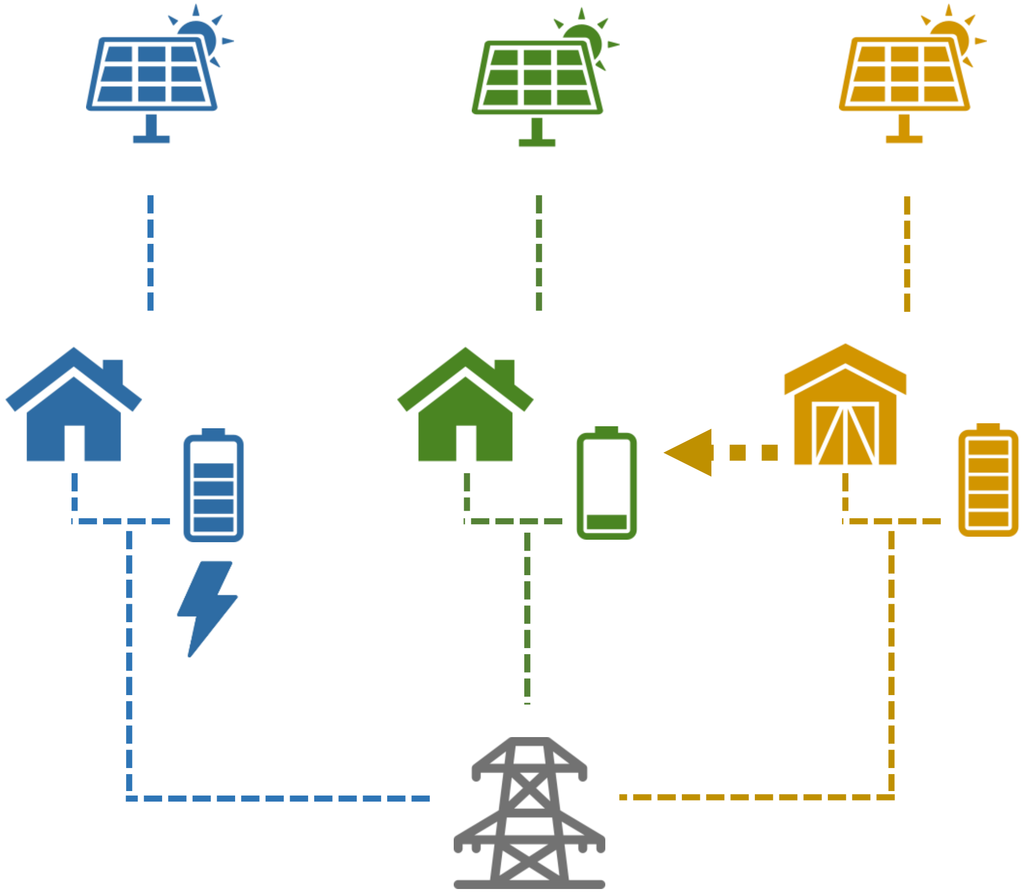
- … weather is sunny (everywhere)?
- … weather is bad (for a long period)?
- … n% of the batteries have an outage?
- … batteries degrade faster than expected?
- … batteries are connected (shared memory)?

How do we detect/predict system drifts?

→ **Sliding window based VIN-comparison**

Motivation for VINs: Structure knowledge

Motivation for Shapley Values: Forecasted data



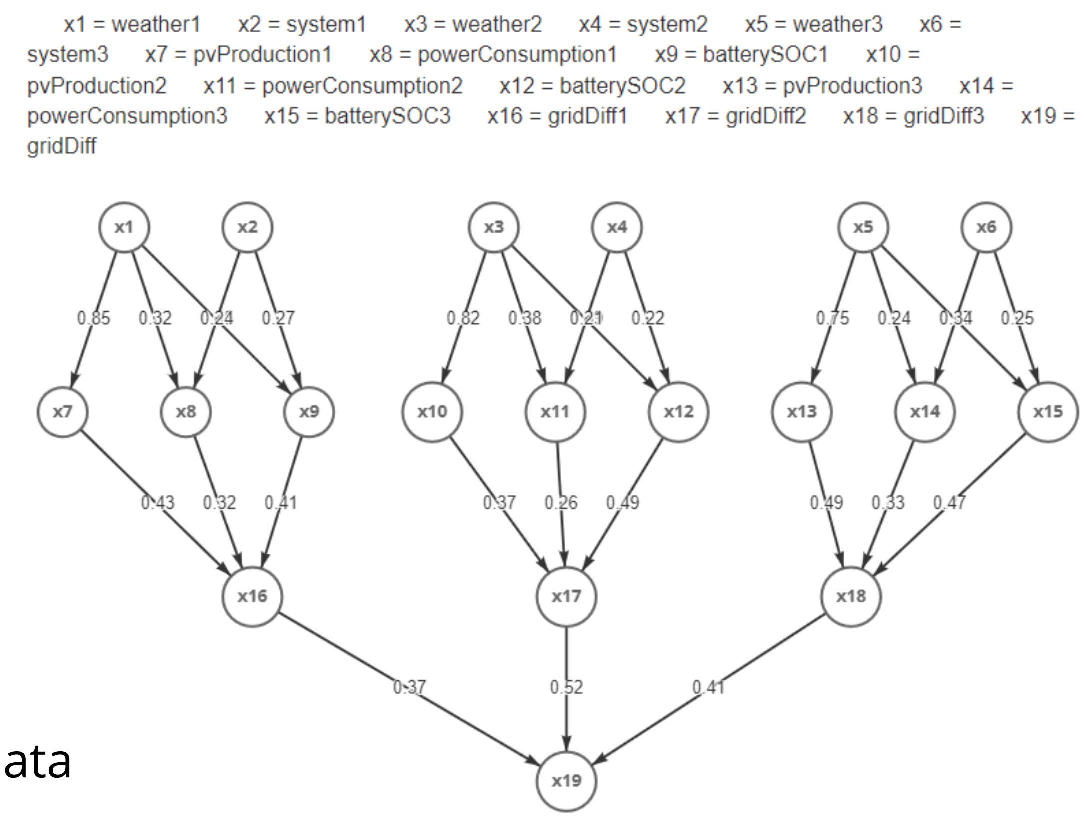
Part II - Experiments

Network Resilience Simulations (Resinet)

What happens if…
- … weather is sunny (everywhere)?
- … weather is bad (for a long period)?
- … n% of the batteries have an outage?
- … batteries degrade faster than expected?
- … batteries are connected (shared memory)?

How do we detect/predict system drifts?

→ **Sliding window based VIN-comparison**

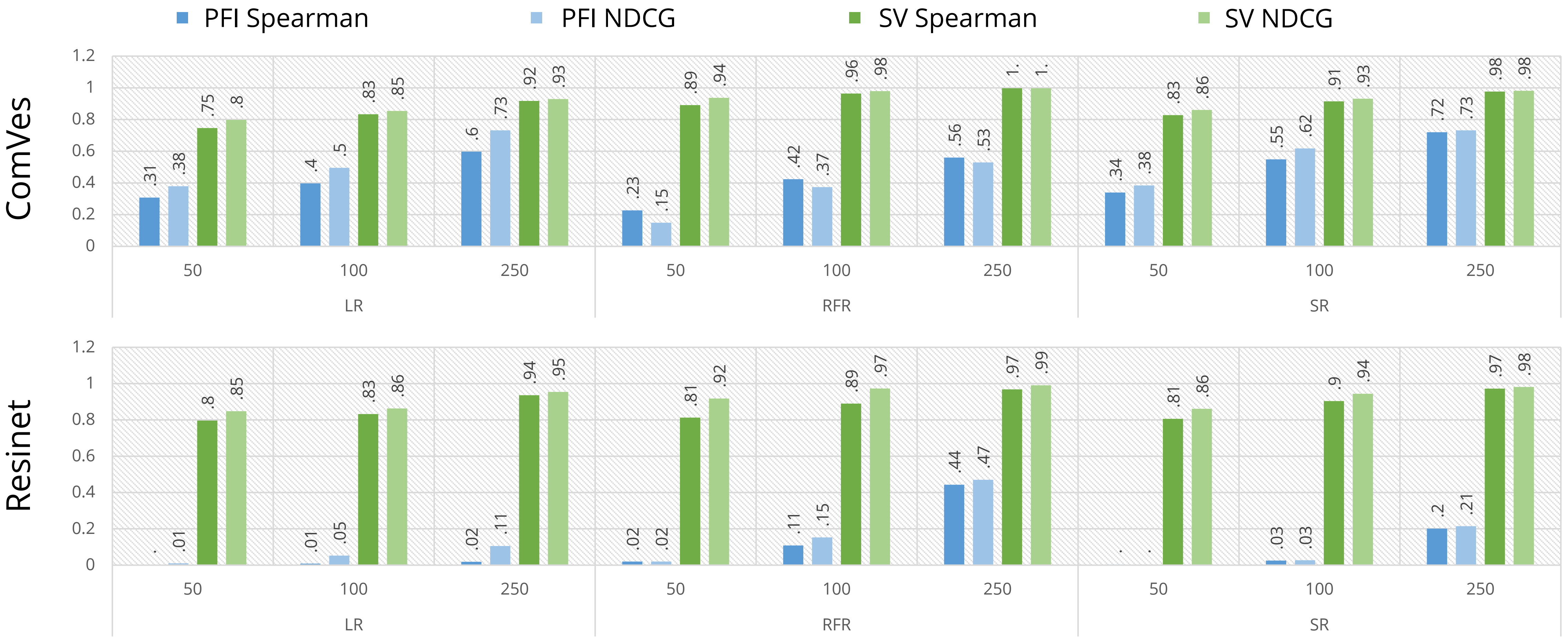


Motivation for VINs: Structure knowledge

Motivation for Shapley Values: Forecasted data

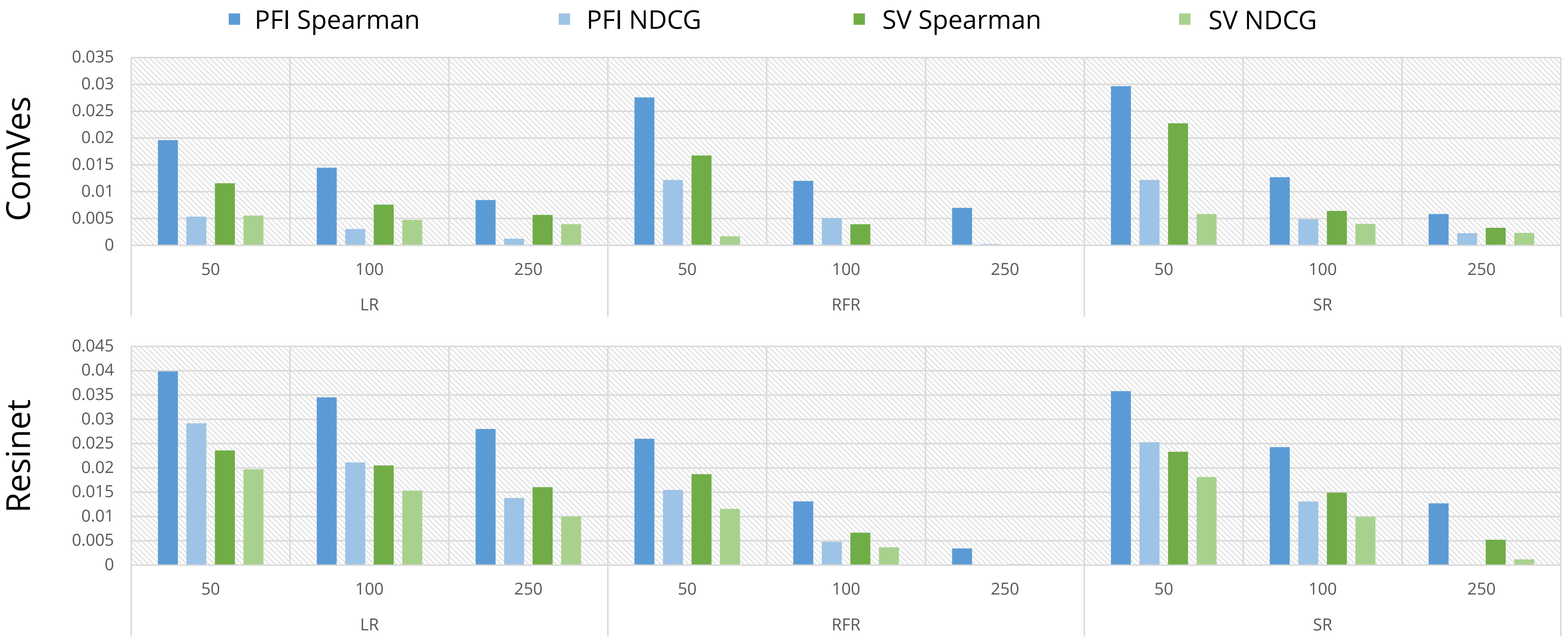
Part II - Experiments

Stability Test Results: Stability Ratio



Part II - Experiments

Stability Test Results: SD of Changes

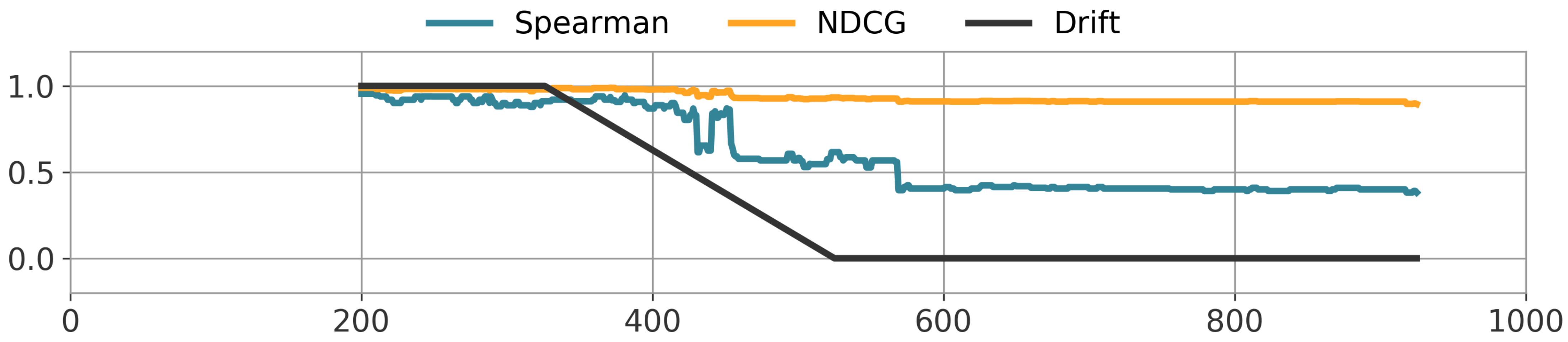


Introducing Drift

Communicating Vessels: clogging communication paths

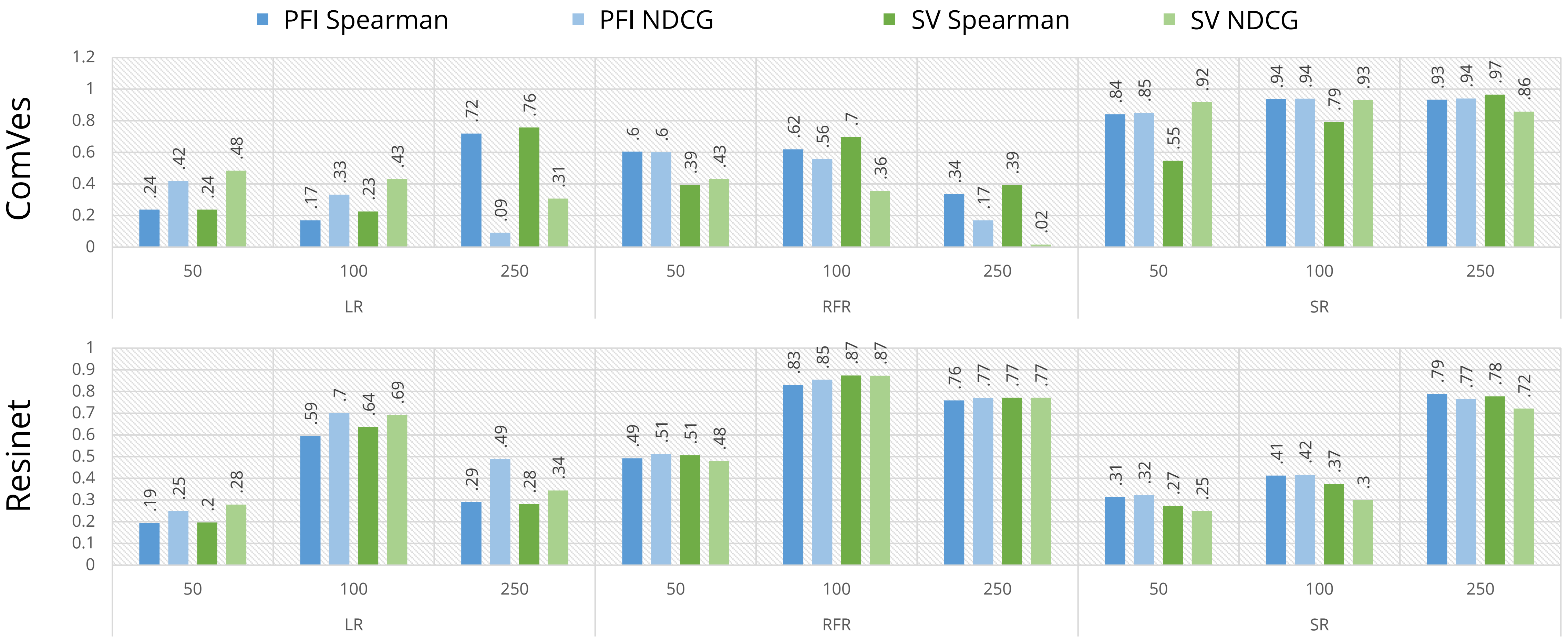
Photovoltaic Network: shared batteries + individual outages

Detection scoring: *Pearson R* of (hidden) drift and network similarity



Part II - Experiments

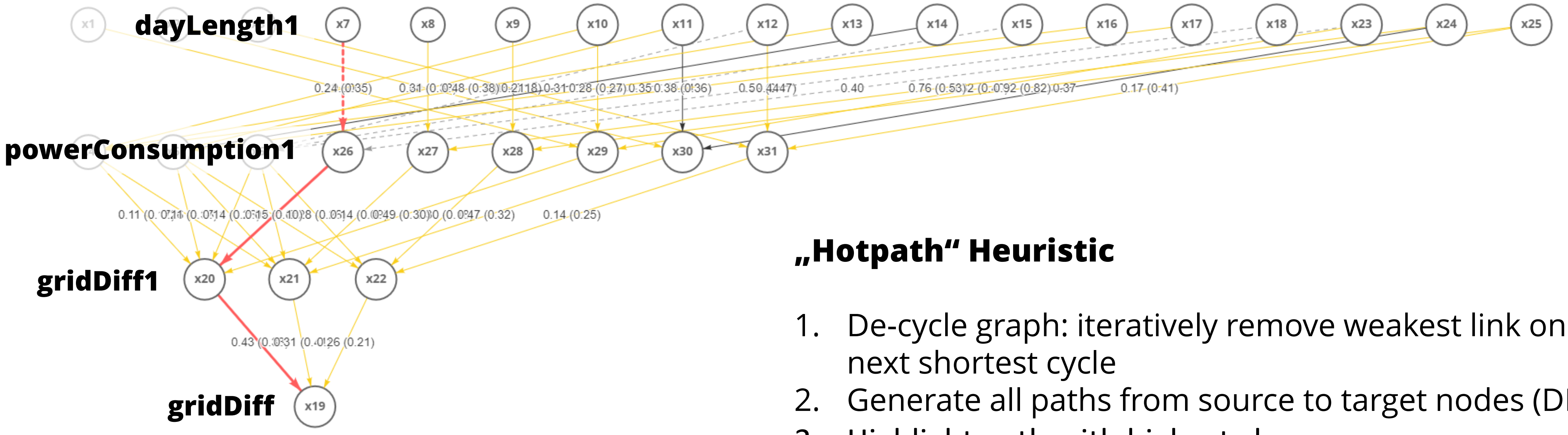
Drift Test Results: Pearson R



Part II - Experiments

Implementation and „Root-Cause“ Analysis

x1 = age1 x2 = age2 x3 = age3 x4 = batterySOCWh1 x5 = batterySOCWh2 x6 = batterySOCWh3 x7 = dayLength1 x8 = dayLength2 x9 = dayLength3 x10 = globalRadiation1 x11 = globalRadiation2 x12 = globalRadiation3 x13 = globalRadiationSum2h1 x14 = globalRadiationSum2h2 x15 = globalRadiationSum2h3 x16 = globalRadiationSumFrame07to12h1 x17 = globalRadiationSumFrame07to12h2 x18 = globalRadiationSumFrame07to12h3 x19 = gridDiff x20 = gridDiff1 x21 = gridDiff2 x22 = gridDiff3 x23 = hoursAfterSunrise1 x24 = hoursAfterSunrise2 x25 = hoursAfterSunrise3 x26 = powerConsumption1 x27 = powerConsumption2 x28 = powerConsumption3 x29 = pvProduction1 x30 = pvProduction2 x31 = pvProduction3



„Hotpath“ Heuristic

1. De-cycle graph: iteratively remove weakest link on next shortest cycle
2. Generate all paths from source to target nodes (DFS)
3. Highlight path with highest change sum

Take-Home Messages and Outlook

Variable Interaction Networks (VIN)

- enable holistic system analysis (also on streaming data)
- enable knowledge integration (i.e. network structure)
- currently underrepresented in the field Explainable / Interpretable AI

	Evaluation	Precision	Stability	Data Access
– based on PFI:	fast	high	mediocre	input, true outcome
– based on SV:	slow (bulk), fast (streaming)	high	high	input

Further leads: extend root-cause analysis

- „Hotpath“ improvement, e.g. add memory (find most stable over time)
- Classification approach

Q & A Shapley Value based Variable Interaction Networks for Data Stream Analysis

Eurocast 2022 // 2022-02-23

Jan Zenisek, Sebastian Dorl, Stephan Winkler and Michael Affenzeller

Heuristic and Evolutionary Algorithms Laboratory

University of Applied Sciences Upper Austria

Institute for Symbolic Artificial Intelligence

Johannes Kepler University Linz

References and Acknowlegdements

[1] Kronberger et al. Genetic Programming: Current Trends and Applications in Computational Finance, Nova Science Publishers, 2013

[2] Hooker, Giles. Discovering additive structure in black box functions. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004

[3] Zenisek, J., Kronberger, G., Wolfartsberger, J., Wild, N., & Affenzeller, M. Concept Drift Detection with Variable Interaction Networks. In International Conference on Computer Aided Systems Theory (pp. 296-303). Springer, Cham, 2020.

[4] Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)

[5] Shapley, Lloyd S. "A value for n-person games." Contributions to the Theory of Games 2.28 (1953): 307-317

Funding thankfully received from:

ober österreich

Europäische Union Investitionen in Wachstum & Beschäftigung. Österreich.

EFRE

Weather data provided by:

IMPLEMENTED BY ECMWF

ZAMG Zentralanstalt für Meteorologie und Geodynamik