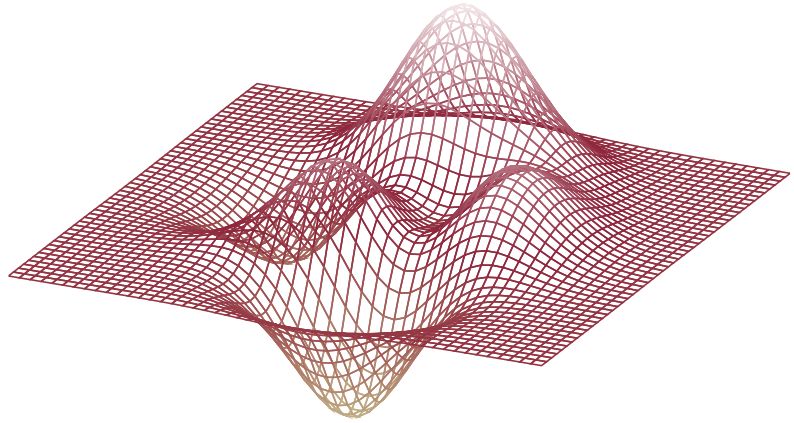


Data Science Research in Economics

Economic Research Seminar Hohenheim

Johannes Bleher
University of Hohenheim

July 2, 2025

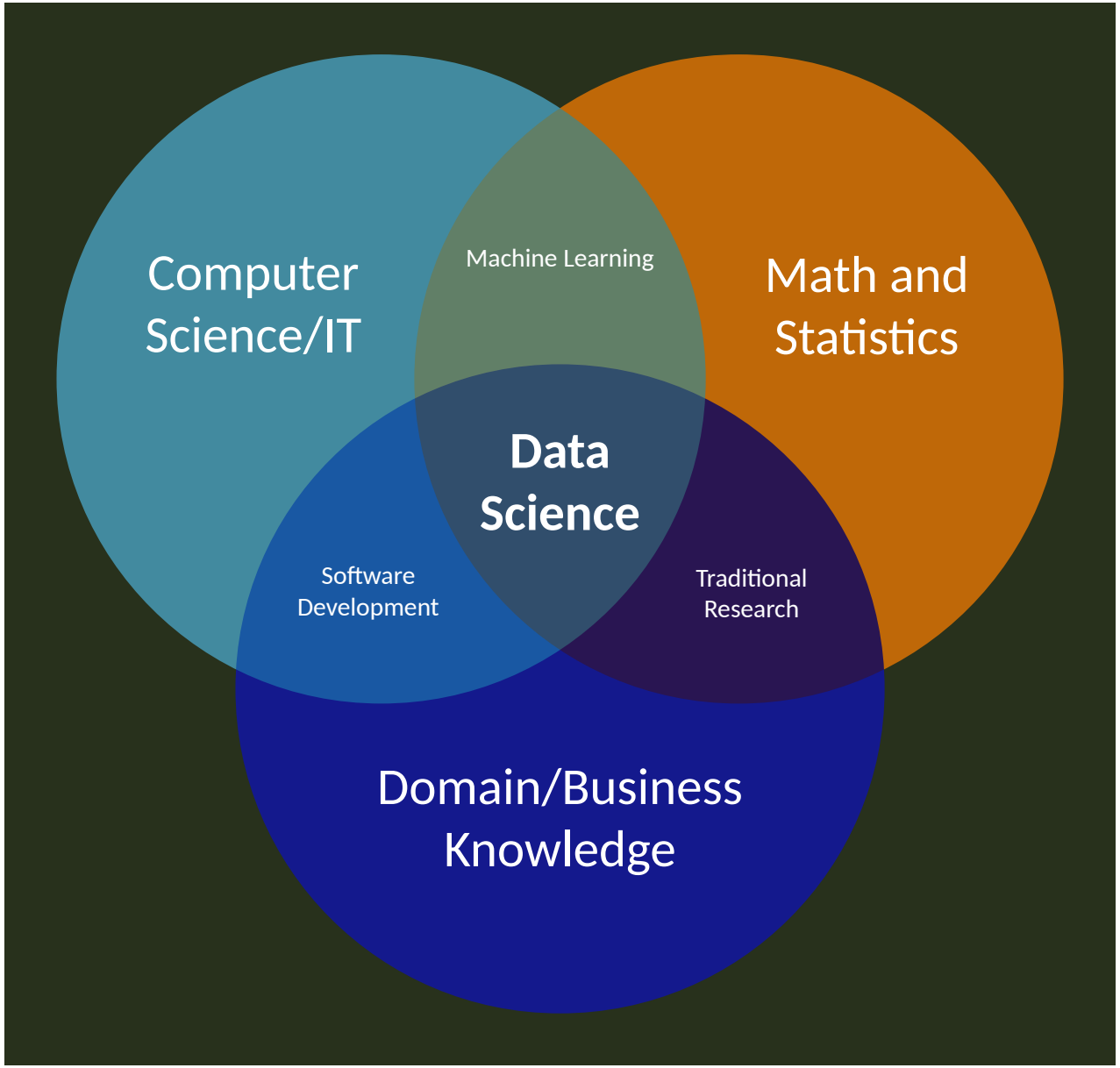
Why Data Science Matters?

- Transformation of (Economic) Research
  - New data sources
  - Larger data sources
  - Advanced analytics
  - Automation (e.g., data pipelines, reactive visualization, ...)
- More productivity with Generative AI

  (e.g., AI agents: Wall Street firms already use AI to do 95% of IPO documentation within minutes – Goldman Sachs CEO David Solomon – June 28, 2025)
- Economic theory for (ethical) alignment of AI systems

$\Rightarrow$ This requires an enhanced skillset!

The 3 Circles of Data Science



My Profile

Studies in Tübingen

- International **Economics and European Studies, BSc.** (2007 – 2012)
- **Economics and Finance, MSc.** (2014 – 2016)
- **Doctoral Thesis** in Finance and Econometrics (2016 – 2021)
  - Econometric Method Development (Estimation of Transfer Entropy)
  - Google Trends Analysis
  - Volatility Forecasting Cryptocurrencies
  - Limit Order Book Dynamics

Professional Experience

- Local (2009–2012, 2014–2016) and Accred. Parliamentary Assistant in the EP (2012–2014), Brussels & Strasbourg
- Student Assistant (2008–2012, 2014–2016) and Research Associate (2016–2021) University of Tübingen
- Credit Risk Analyst, Bausparkasse Schwäbisch Hall (2020–2022)
- University of Hohenheim
  - Coordinator AIDAHO
  - CSH-Manager
  - Lecturer
- Freelance Consultant (2022 – present)

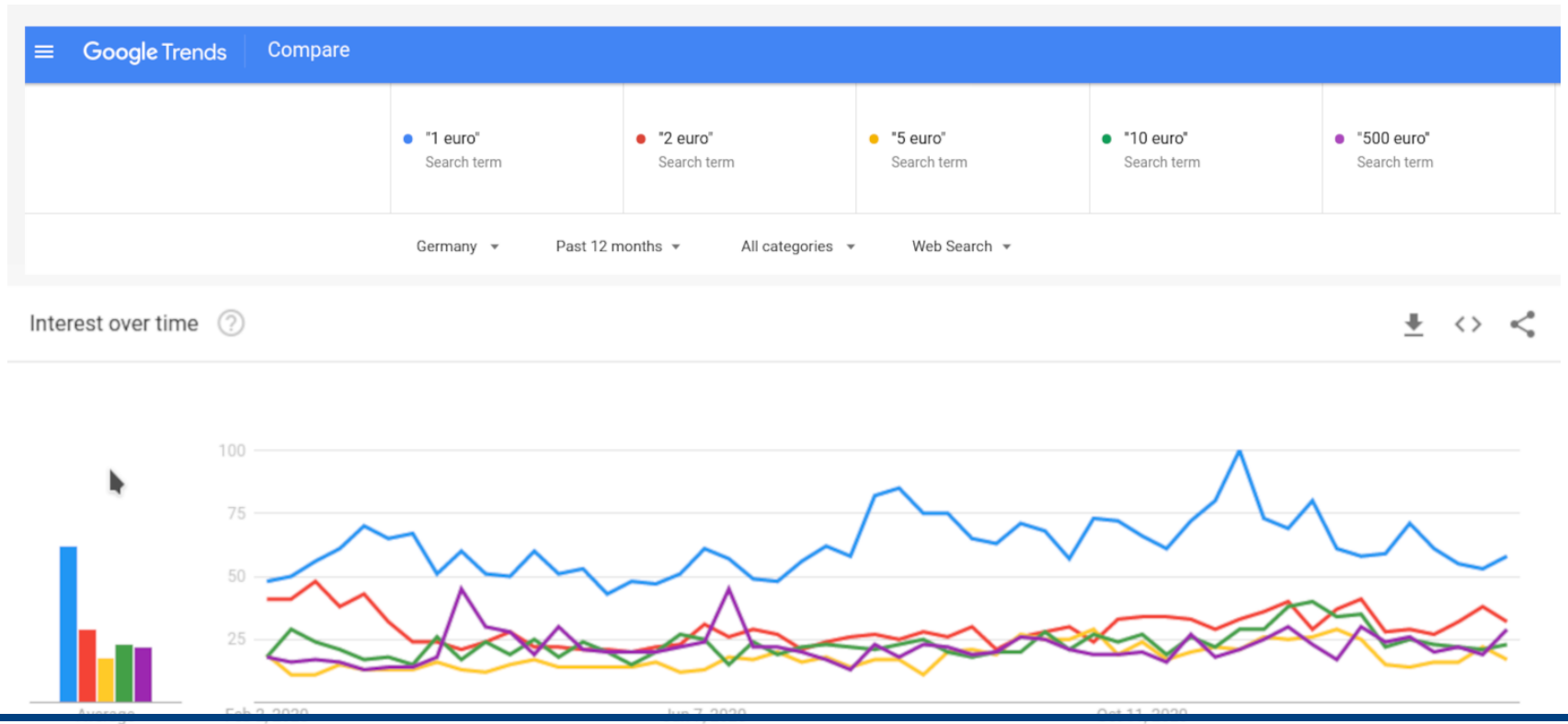
Today: 3 Data Science Projects and an Outlook

- **New data**
  - Index of Prices Searched Online
- **Big(ger) data**
  - German hydrology data and fuel prices
- **New methdology**
  - Conditional Density and Transfer Entropy Estimation
  - A Bayesian-Frequentist Approach to Post-Double Selection Regressions for Survey Data with Multiple Imputation
- **Automation with AI agents** (for another talk)
  - Queing System for GenAI APIs on AIDAHO servers
  - Database of Austria-Hungarian census data

Bleher and Dimpfl (2022)

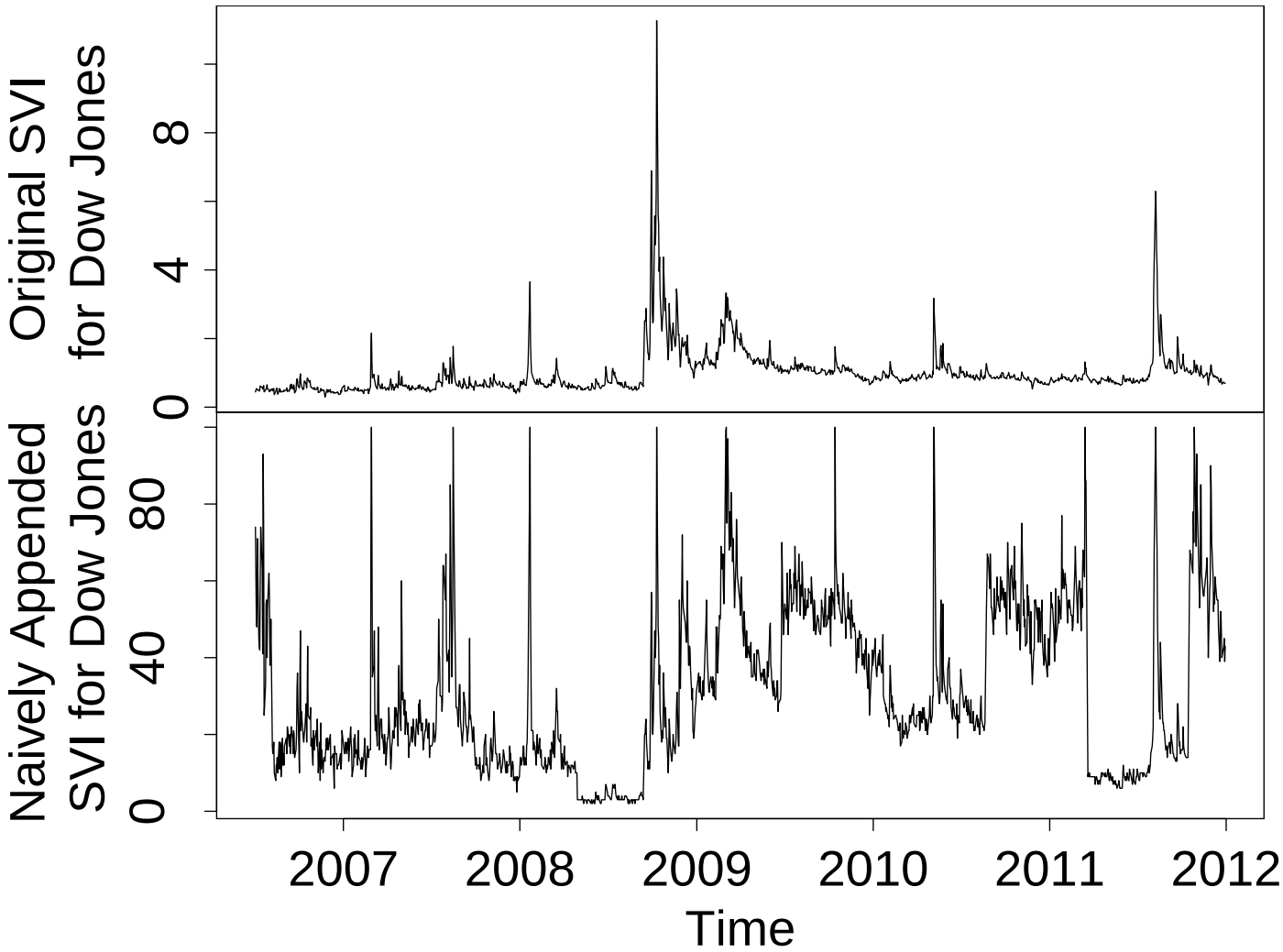
Google's Search Volume Index

- Want to know about *XYZ*? Google it! – Google records it.
- Web, News, Images, Youtube.
- Regional data available.
- No official API, Access limited – selected frequencies, compare up to 5 queries.



Yes you can use daily frequency, just do it.

(cp Dastgir, Demir, Downing, Gozgor and Lau 2019)



Let's unlock the full potential of Google Trends!

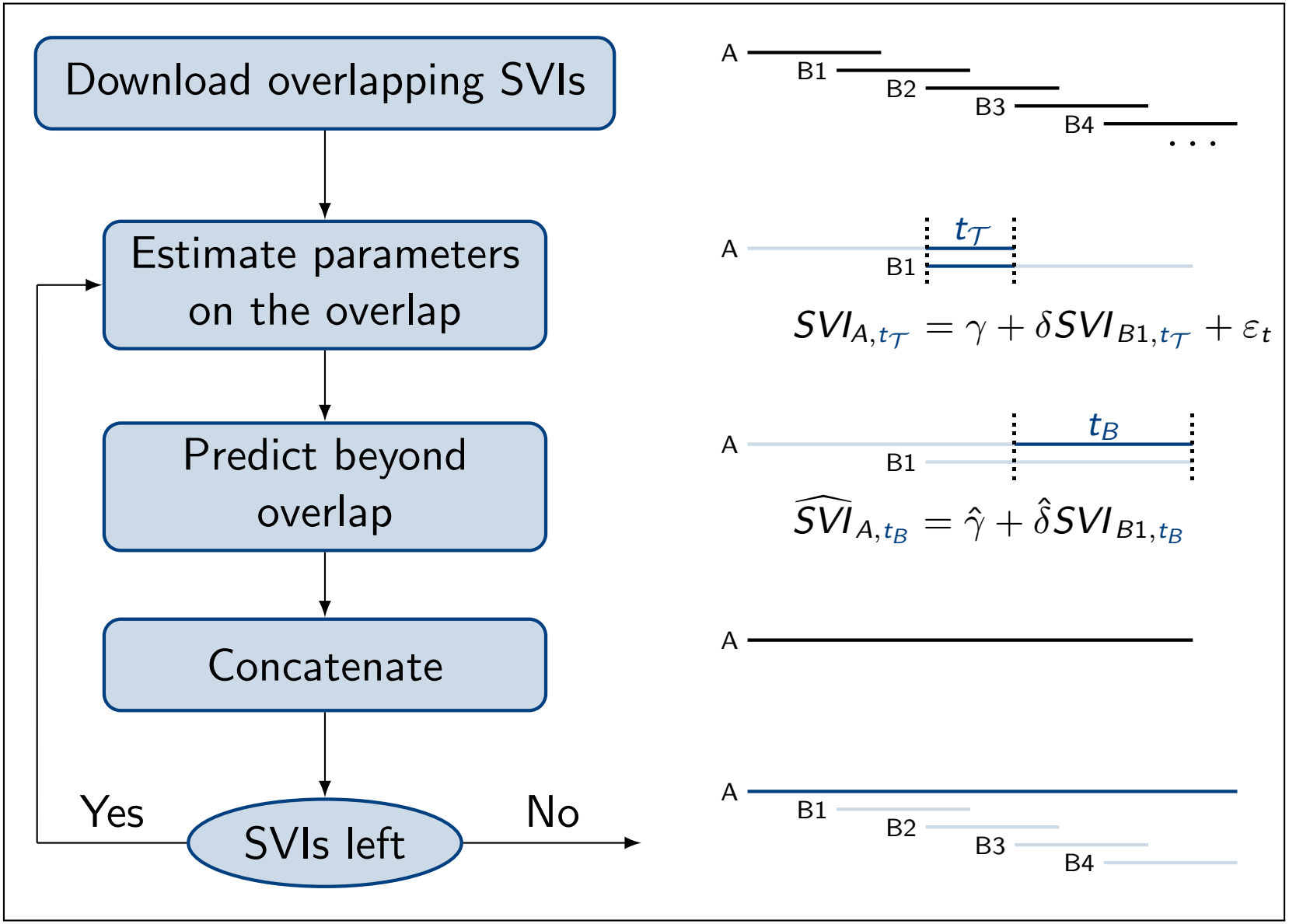
$$SVI_{j,i,t|\mathcal{M},\mathcal{T}} = 100 \cdot \text{round}\left(\frac{s_{j,i,t} - L}{\max_{\substack{m\in\mathcal{M}\\ t\in\mathcal{T}}}(s_{m,i,t}) - L}\right)$$

$j \in \mathcal{M}$	the set of all topics searched during $\mathcal{T}$
$t \in \mathcal{T}$	the time frame used when downloading (e.g. 90 days)
i	the region
$s_{j,i,t}$	the search propensity for topic j in i during t
L	unknown threshold (only Google knows)

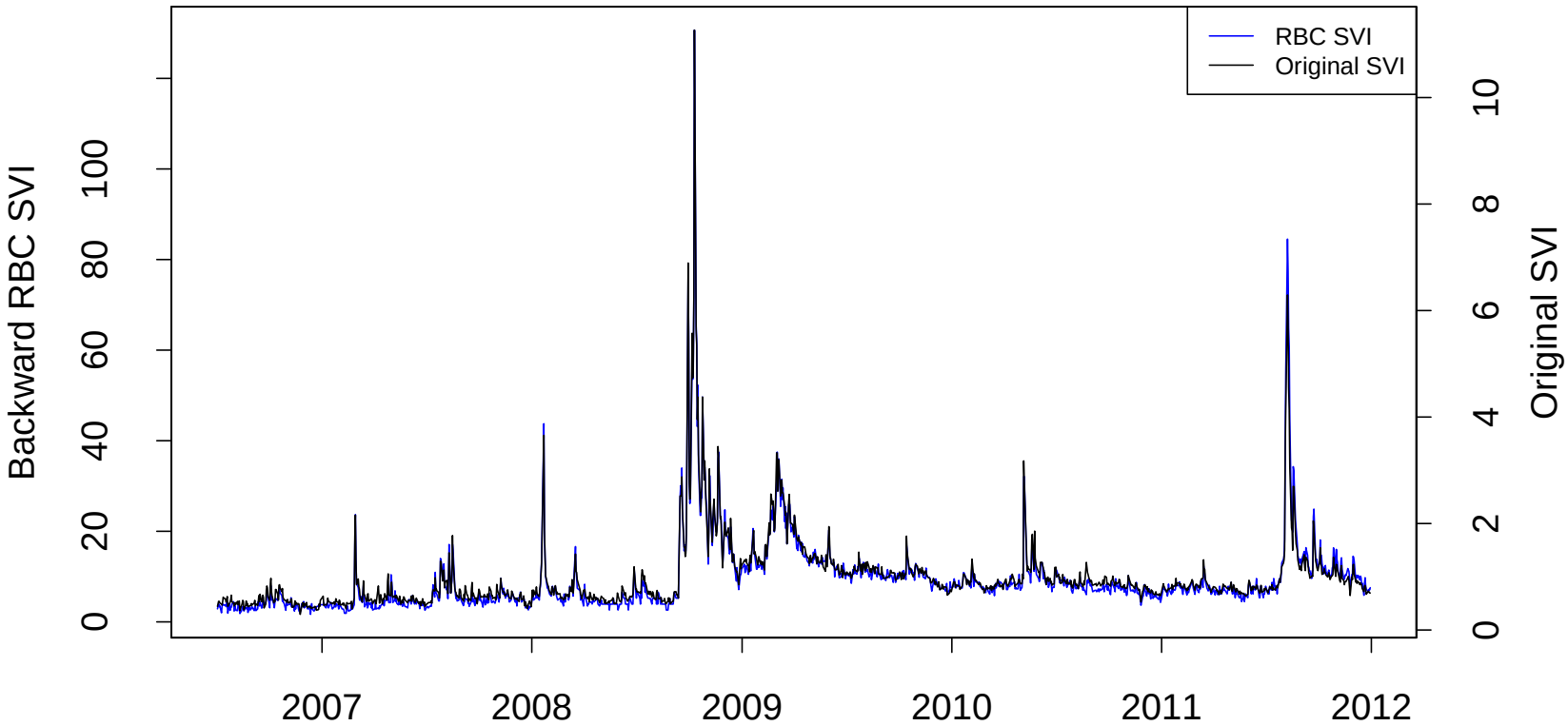
$\Rightarrow$ **This is an affine linear transformation (in expectations).**

$$SVI_{j,i,t|\mathcal{M},\mathcal{T}} = \alpha_{\mathcal{M},i,\mathcal{T}} + \beta_{\mathcal{M},i,\mathcal{T}} s_{j,i,t} + \nu_{j,i,t},$$

Regression Based Concatenation



This works.



Correlation:

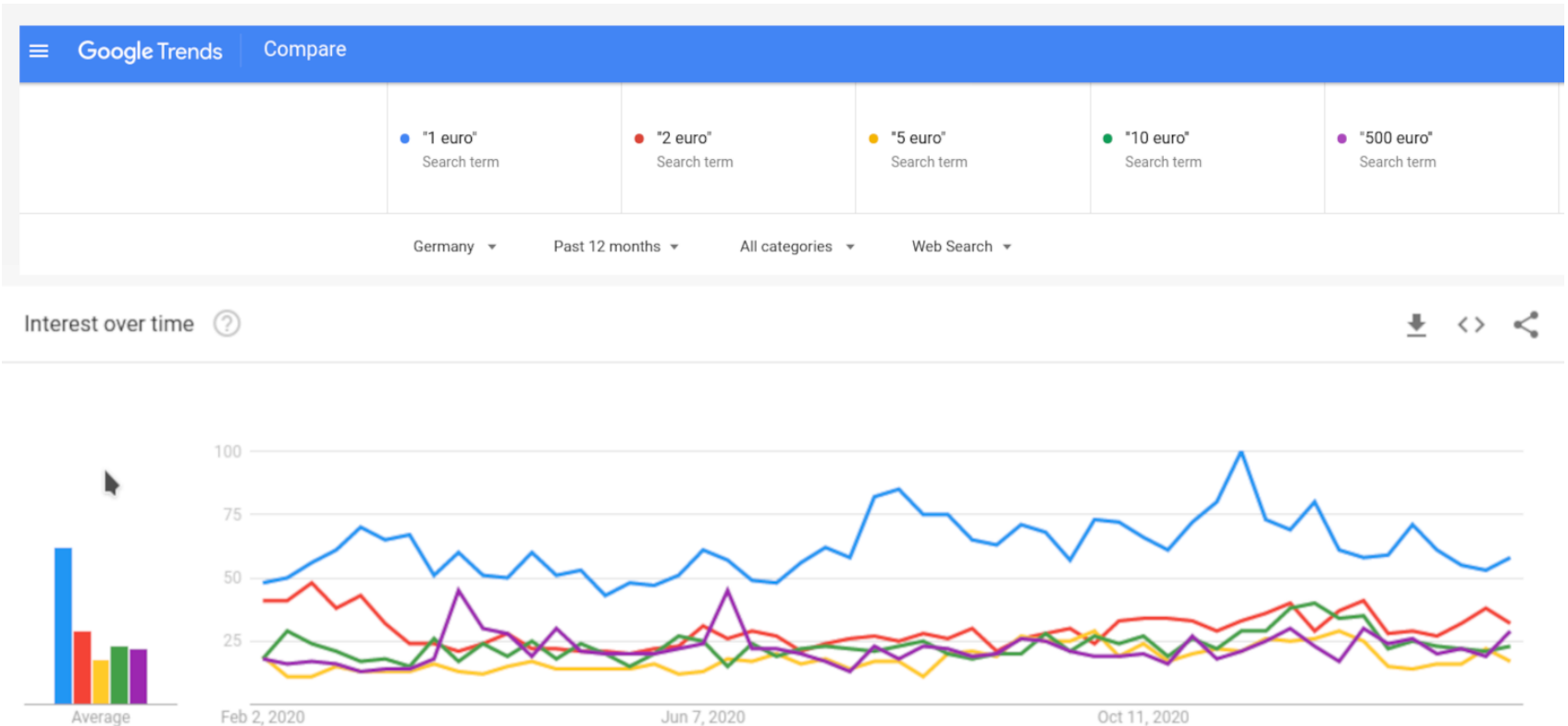
	levels	returns
RBC/original	0.9854	0.7294
Naiv/original	0.4036	0.6496

Use Query Comparison

WIth the ratio of (comparable) SVIs

$$R_{j,i,t} = \frac{s_{j,i,t} - L}{s_{j_1,i,t} - L} \approx \frac{s_{j,i,t}}{s_{j_1,i,t}},$$

we can also make SVIs coherent across queries.



Construct the Index of Prices Searched Online

Calculate the distribution function over a set of prices

$$F(p) = \text{Prob}(P \leq p) \approx \frac{\sum_{s=1}^{p} SVI_{j_s,i,t|\mathcal{P},\mathcal{T}}}{\sum_{s \in \mathcal{P}} SVI_{j_s,i,t|\mathcal{M}_P,\mathcal{T}}},$$

approximate the probability density function via

$$f(p) \approx \frac{\Delta F(p)}{\Delta p},$$

and calculate the expected value, the Index of Prices Searched Online (IPSO).

$$\text{IPSO} = \mathbb{E}[P] \approx \sum_{p \in \mathcal{P}} p \cdot f(p).$$

Forecasting US Inflation and Consumption?

In-sample fit is weak:

Series	S	p	Granger IPSO$\rightarrow$	Granger $\rightarrow$IPSO	Contemp.	*RMSE*
US Inflation	0	2	**0.030**	**0.000**	0.086	0.309
	12	1	0.080	0.338	0.644	0.272
US Consumption	0	1	0.886	0.815	0.322	0.280
	12	1	0.875	0.324	**0.039**	0.281
US BEIR	0	2	0.306	0.427	**0.036**	0.030

where S is the seasonality component and p the lag length.

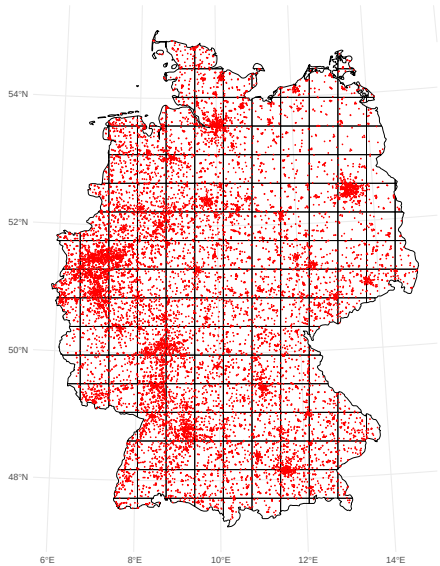
Forecasting US Inflation and Consumption?

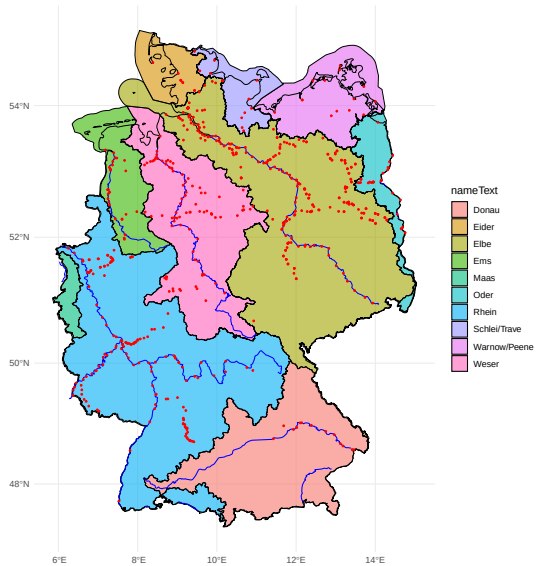
Compared to AR(p)-benchmark, *RMSPE* reduction out-of-sample is sizable

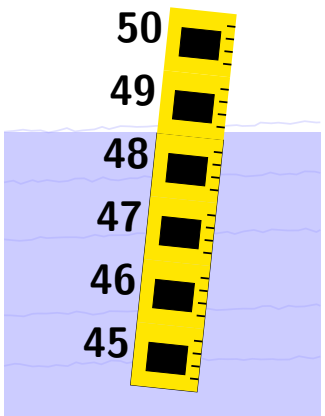
		S	with IPSO	without IPSO	Difference
US Inflation	*RMSPE*	0	0.172	0.277	**-0.104**
		12	0.137	0.196	**-0.059**
	R^2_{MZ}	0	26.95	25.83	**1.12**
		12	54.89	54.05	**0.84**
US Consumption	*RMSPE*	0	0.180	0.364	**-0.184**
		12	0.189	0.266	**-0.078**
	R^2_{MZ}	0	1.12	5.96	-4.84
		12	1.76	1.45	**0.31**
US BEIR	*RMSPE*	0	0.0205	0.0459	**-0.0253**
	R^2_{MZ}	0	2.7696	4.7620	-1.9924

where S is the seasonality component and p the lag length.

Do Water Levels Affect Fuel Prices? Evidence from High-Frequency Data on German Waterways (WIP with K. Kuck)







The Data

Integrated:

- Tankerkönig – Fuel price data since June 2014 for 17.592 gas stations (all price changes)
- pegelonline – Wasserstraßen und Schiffahrtsbundesamt – Water level data since the year 2000 (high-frequency)
- River basin GIS data Environmental Systems Research Institute, Inc. (ESRI)
- Shape files for borders Natural Earth Data

Planned:

- Deutsche Wetterdienst (DWD)

$\Rightarrow$ In total $\sim$ 160 GB data.

Why rivers may matter for fuel prices?

- Major rivers (e.g., Rhine, Elbe) are critical corridors for transporting fuels by barge.
- 38.1 mio tons of "mineral oil and coke-oven products" had been transported in 2019 by ship
- Approximate estimates: 100 mio tons via pipelines and 80–120 mio tons via road tankers

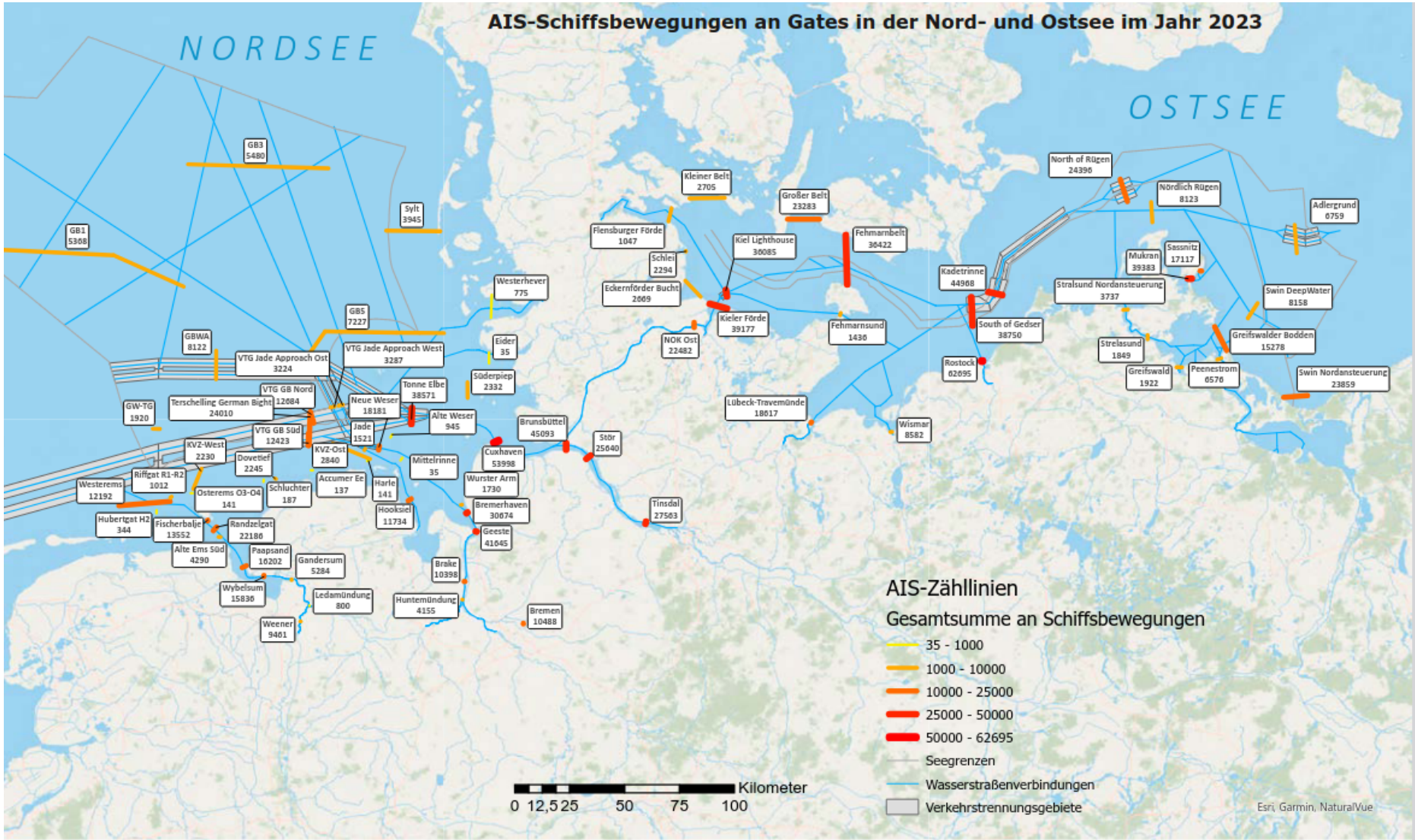
$\approx$ 20% of oil products are transported via ships

$+$ **Large Volumes, Low Cost:** bulk transport of fuels, lower costs compared to rail or road.

$+$ **Strategic Locations:** Refineries and tank farms often near rivers.

BUT: **Vulnerability to Water Levels:** Low or high water can disrupt fuel deliveries, local shortages or price spikes.

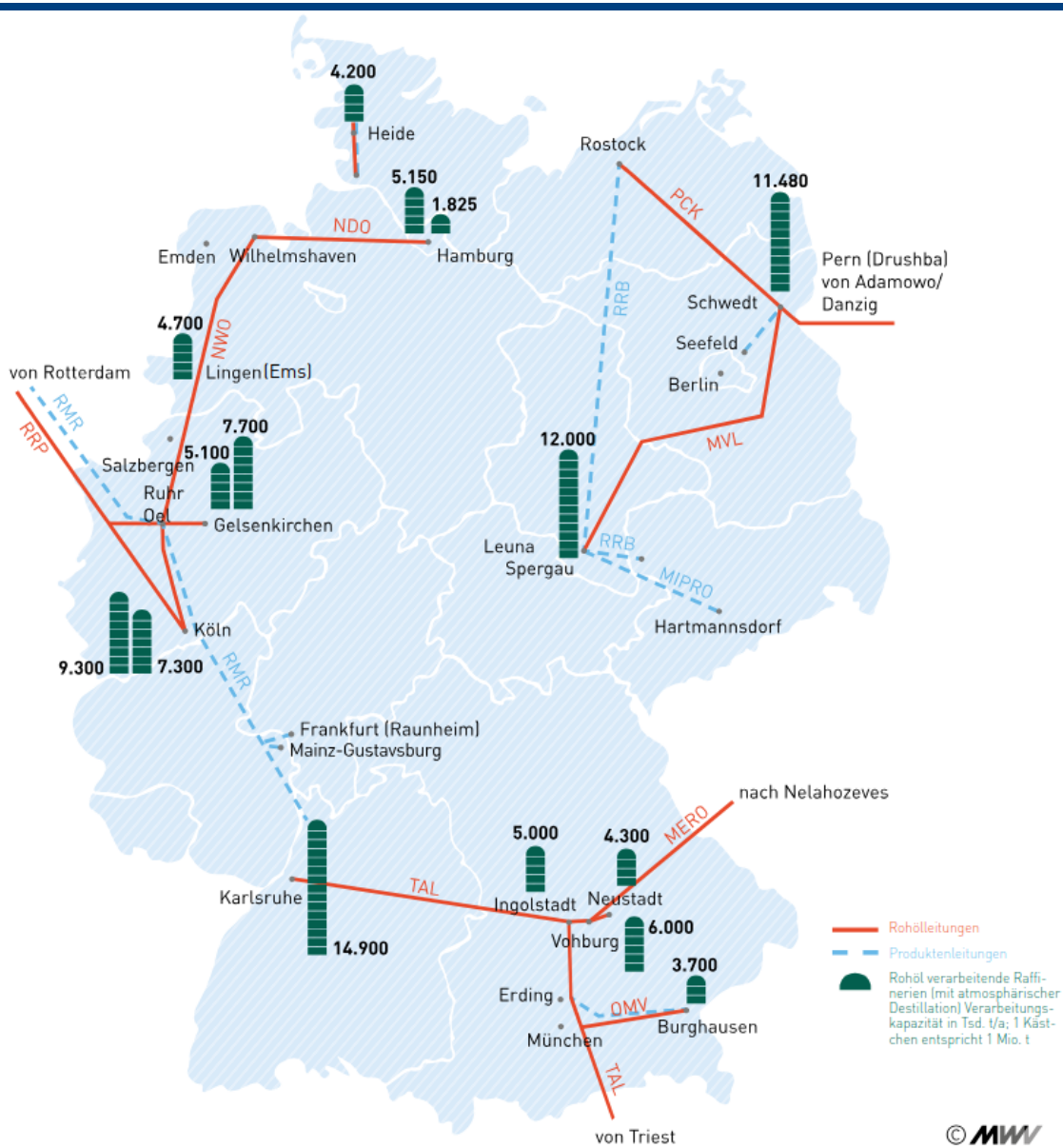
Ports



Kartographie: Fachstelle für Geodäsie und Geoinformatik der WSV (FGeoWSV), Dienstort Aurich, Juli 2024

Quelle der Seegrenzen, Wasserstraßenverbindungen und Verkehrstrennungsgebiete: Bundesamt für Seeschifffahrt und Hydrographie (BSH)

Pipelines


4.200
Heide
5.150
1.825
NDO
Emden
Wilhelmshaven
Hamburg
4.700
NWO
Lingen (Ems)
von Rotterdam
RMR
RRP
7.700
5.100
Salzbergen
Ruhr Oel
Gelsenkirchen
9.300
7.300
Köln
RMR
Frankfurt (Raunheim)
Mainz-Gustavsburg
Karlsruhe
14.900
TAL
5.000
Ingolstadt
4.300
Neustadt
MERO
nach Nelahozeves
6.000
Vohburg
3.700
Erding
OMV
Burghausen
München
TAL
von Triest
Rostock
PCK
11.480
Pern (Drushba)
von Adamowo/
Danzig
Schwedt
Seefeld
Berlin
RRB
12.000
MVL
Leuna
Spergau
RRB
MIPRO
Hartmannsdorf
Rohölleitungen
Produktenleitungen
Rohöl verarbeitende Raffinerien (mit atmosphärischer Destillation) Verarbeitungskapazität in Tsd. t/a; 1 Kästchen entspricht 1 Mio. t
© MW

The Model

$$y_{it} = \alpha + \mu_i + \lambda_t + \mathbf{x}'_{tf}\boldsymbol{\beta} + \mathbf{w}'_{it}\boldsymbol{\delta} + \rho\bar{y}_{i,t,t-14,t-44} + \varepsilon_{it},$$

y_{it} daily average fuel price (diesel, e5, or e10) for grid cell i

$\mathbf{x}_{tf}$ daily river basin f specific water level variables
- standardised average level $\bar{z}_{tf}$ at t, $t+3$ and $t+7$
- propensity $\underline{s}_{tf}$ of high-water level measurements (NW) at t, $t+3$ and $t+7$
- propensity $\bar{s}_{tf}$ of low-water level measurements (HW) at t, $t+3$ and $t+7$
- propensity $\bar{s}^*_{tf}$ of non-navigable level measurements (HSW) at t, $t+3$ and $t+7$

$\mathbf{w}_t$ set of interactions between all variables in $\mathbf{x}_{tf}$

$\bar{y}_{i,t}$ 30-day moving average starting at lag 14 ranging to lag 44

μ_i fixed effect for grid cell i

λ_t fixed effect for month t

Approximate Marginal Effects

The marginal effects can be approximated contemporaneously (with water level information at t) or with foresight (with water level information at $t+3$ or $t+7$) or with both.

$$\mathbb{E}[y_{it} \mid \text{Normal } f] - \mathbb{E}[y_{it} \mid \text{Low } f] \approx (\beta_{0f} + \delta_{0fi}) \left(\bar{z}_{tf|\text{Normal}} - \bar{z}_{tf|\text{Low}}\right) + (\beta_{2f} + \delta_{2fi}) \, \text{avg}_t(\underline{s}_{tf}),$$

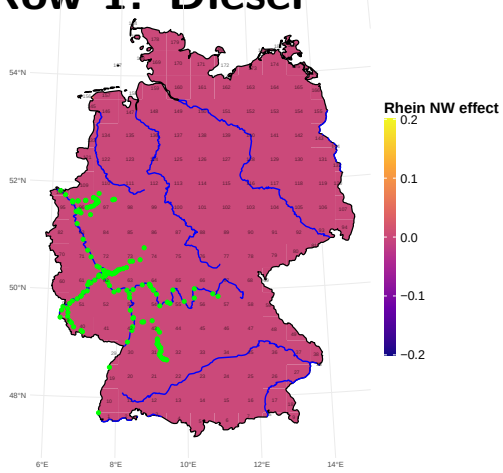
$$\mathbb{E}[y_{it} \mid \text{High } f] - \mathbb{E}[y_{it} \mid \text{Normal } f] \approx (\beta_{0f} + \delta_{0fi}) \left(\bar{z}_{tf|\text{High}} - \bar{z}_{tf|\text{Normal}}\right) + (\beta_{1f} + \delta_{1fi}) \, \text{avg}_t(\bar{s}_{tf}),$$

$$\mathbb{E}[y_{it} \mid \text{NSB } f] - \mathbb{E}[y_{it} \mid \text{Normal } f] \approx (\beta_{0f} + \delta_{0fi}) \left(\bar{z}_{tf|\text{NSB}} - \bar{z}_{tf|\text{Normal}}\right) + (\beta_{2f} + \delta_{2fi}) \, \text{avg}_t(\bar{s}_{tf}) + (\beta_{3f} + \delta_{3fi}) \, \text{avg}_t(\bar{s}^*_{tf}).$$

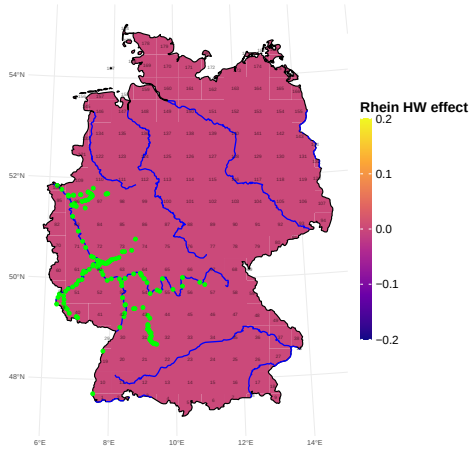
(Preliminary) Results

Rhine – Contemporaneous Extreme Water Event – 5% sig. level

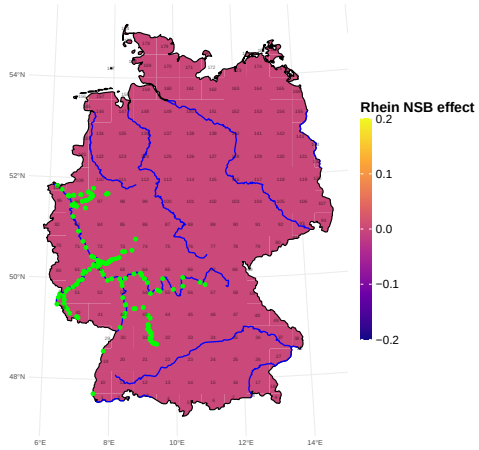
Row 1: Diesel



low water

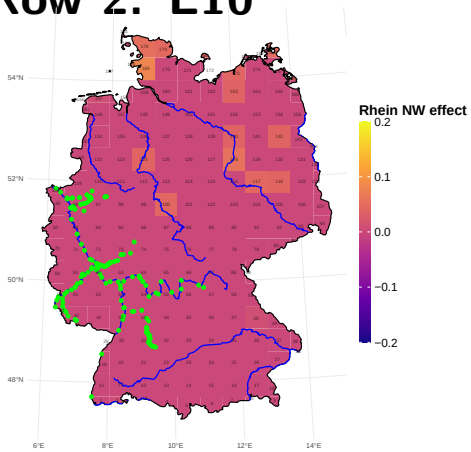


high water

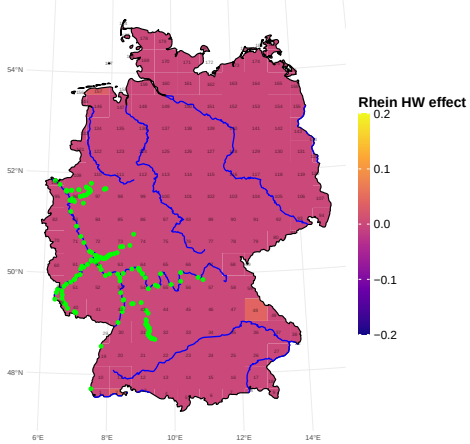


non-navigable

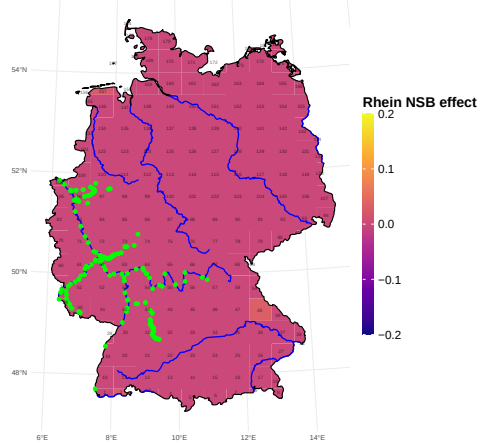
Row 2: E10



low water



high water

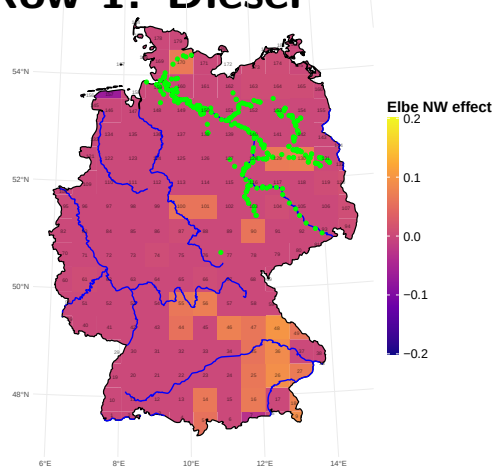


non-navigable

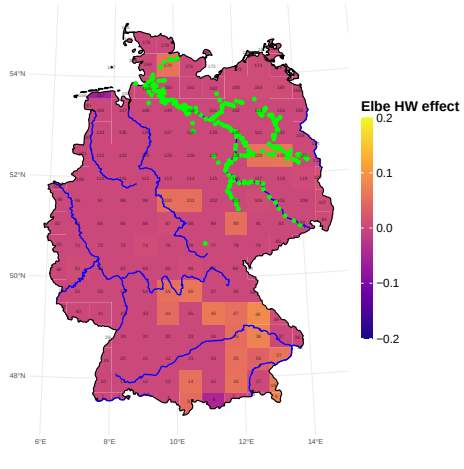
(Preliminary) Results

Elbe – Contemporaneous Extreme Water Event – 5% sig. level

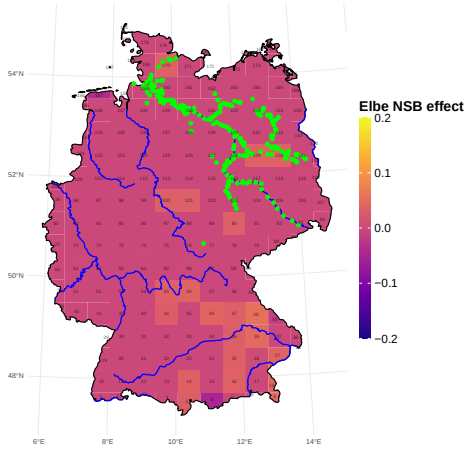
Row 1: Diesel



low water

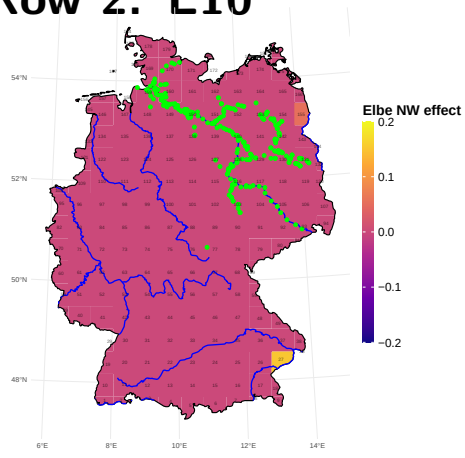


high water

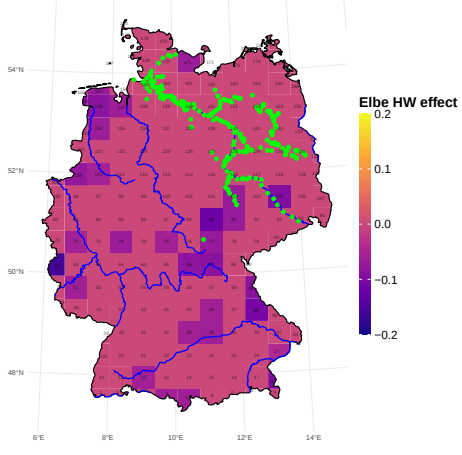


non-navigable

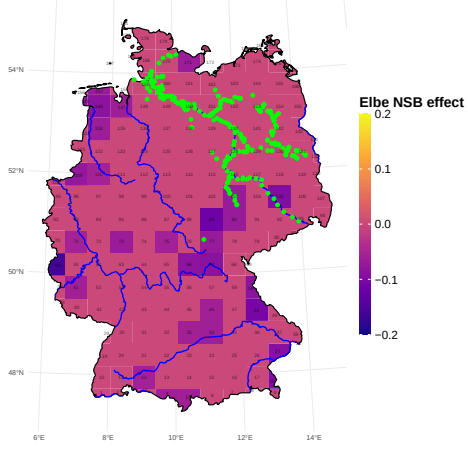
Row 2: E10



low water



high water

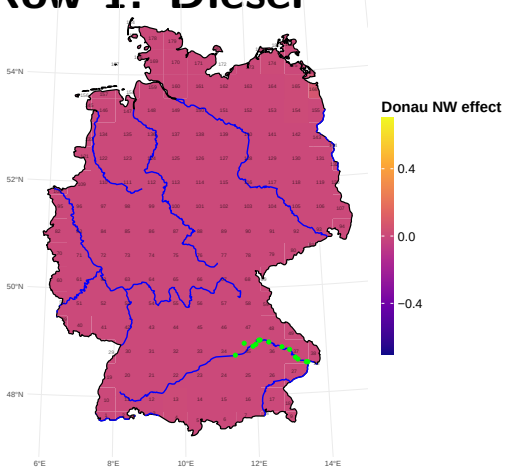


non-navigable

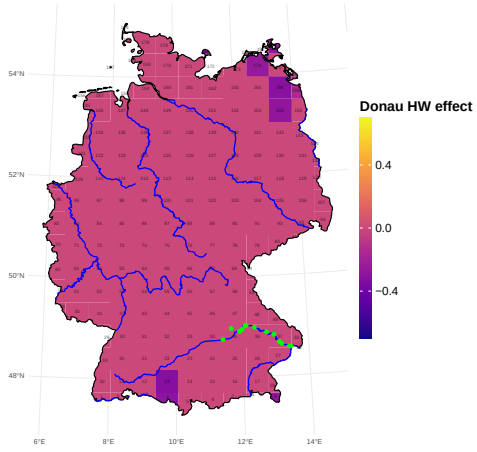
(Preliminary) Results

Donau – Contemporaneous Extreme Water Event – 5% sig. level

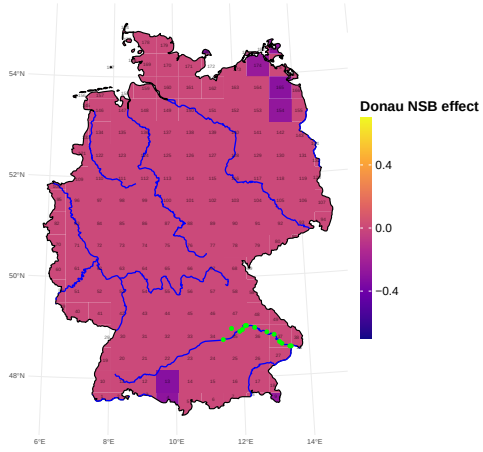
Row 1: Diesel



low water

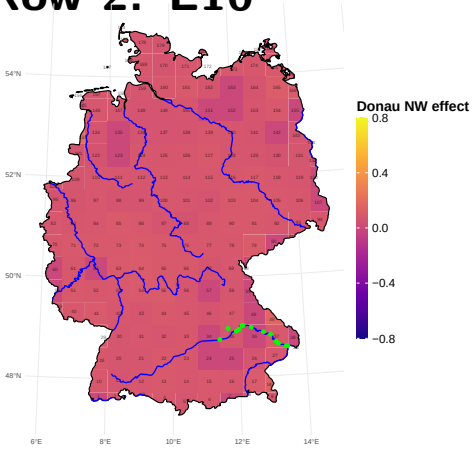


high water

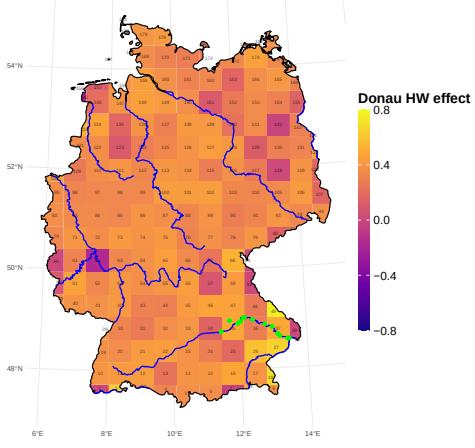


non-navigable

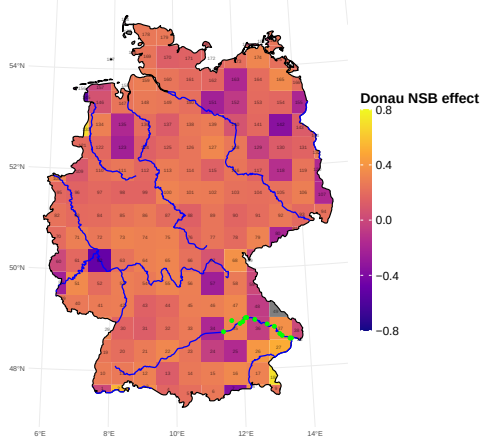
Row 2: E10



low water



high water

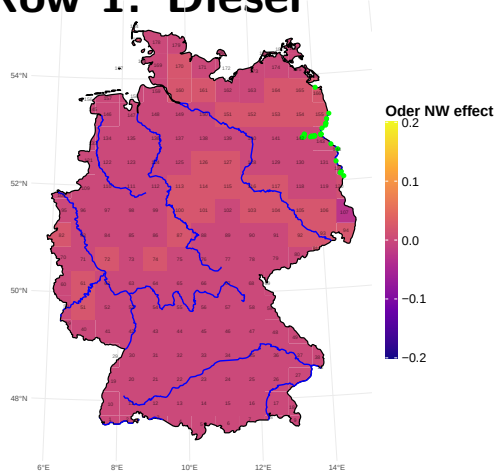


non-navigable

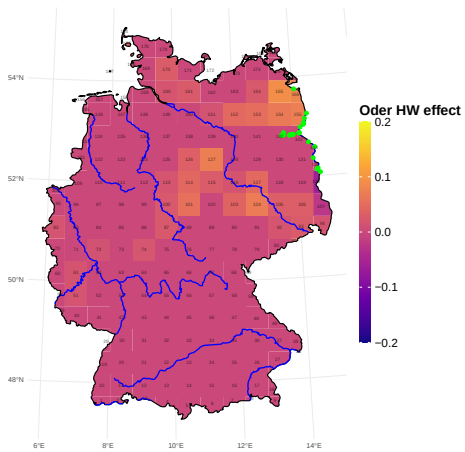
(Preliminary) Results

Oder – Contemporaneous Extreme Water Event – 5% sig. level

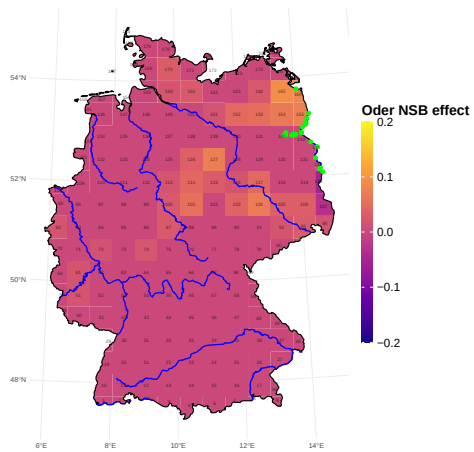
Row 1: Diesel



low water

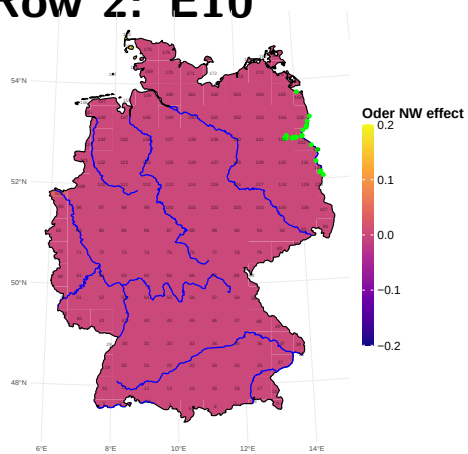


high water

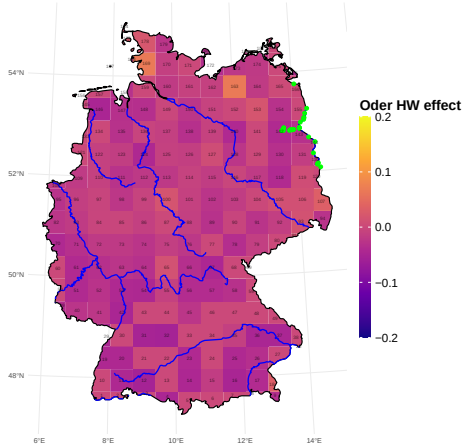


non-navigable

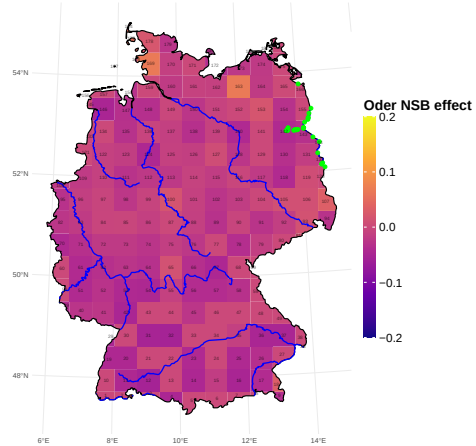
Row 2: E10



low water



high water

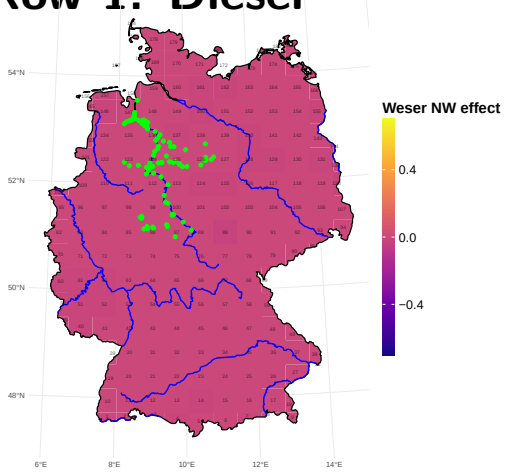


non-navigable

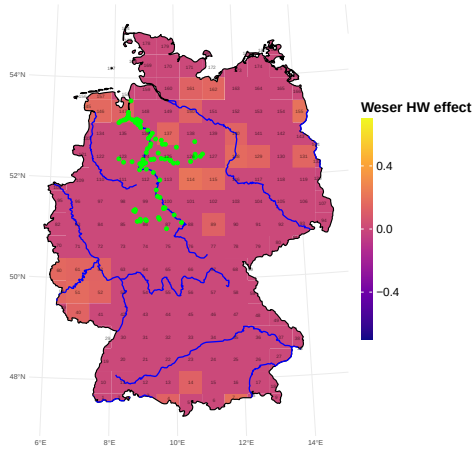
(Preliminary) Results

Weser – Contemporaneous Extreme Water Event – 5% sig. level

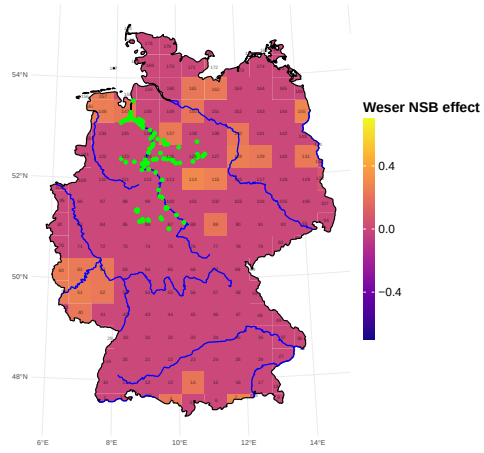
Row 1: Diesel



low water

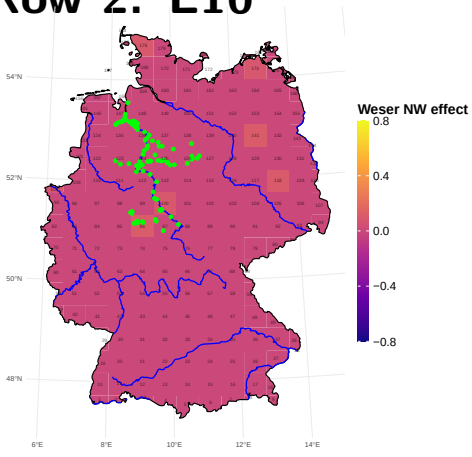


high water

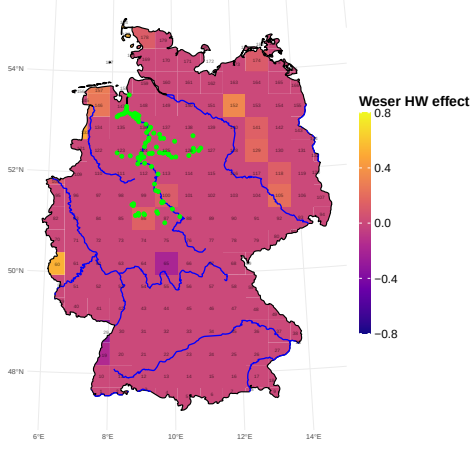


non-navigable

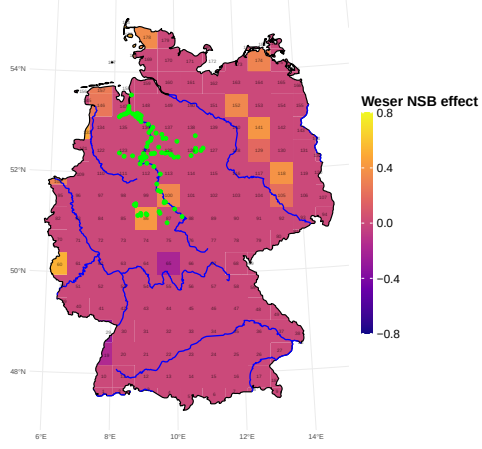
Row 2: E10



low water



high water

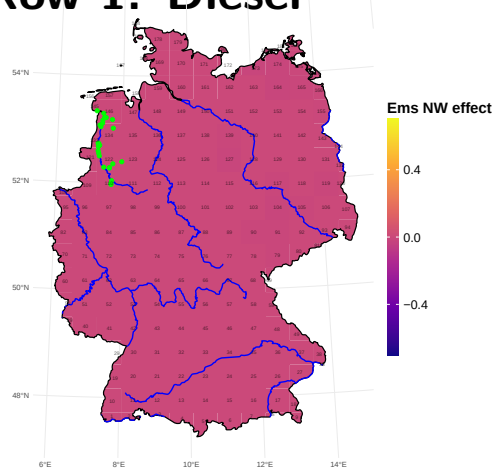


non-navigable

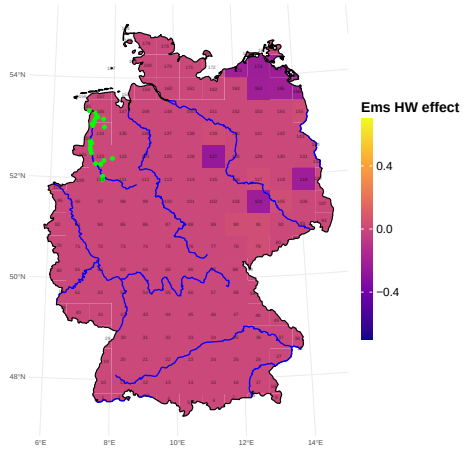
(Preliminary) Results

Ems – Contemporaneous Extreme Water Event – 5% sig. level

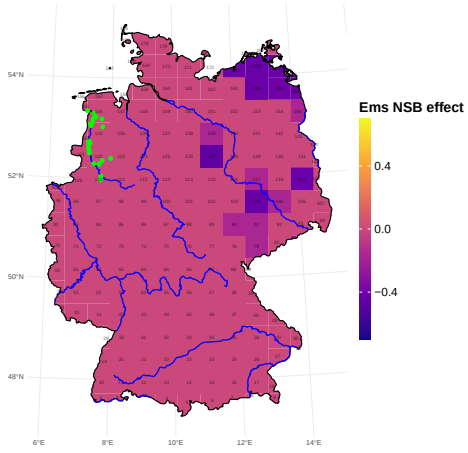
Row 1: Diesel



low water

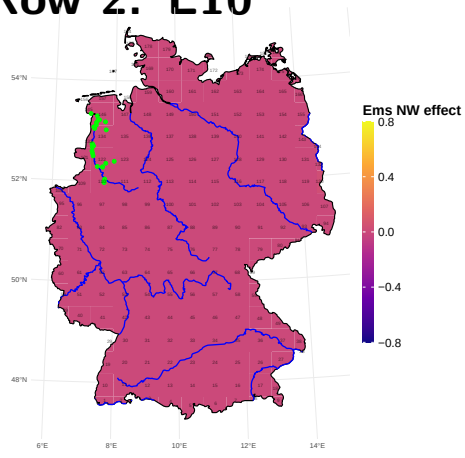


high water

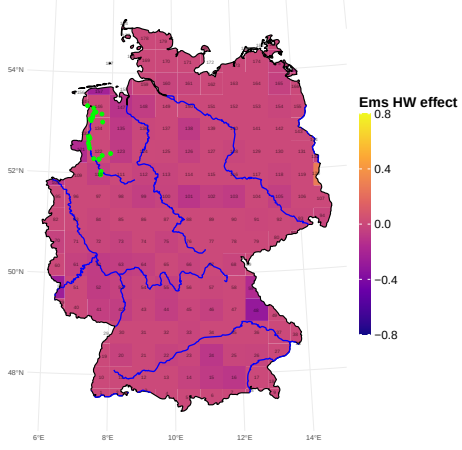


non-navigable

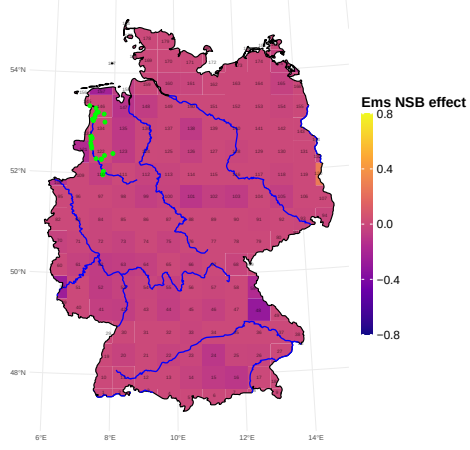
Row 2: E10



low water



high water

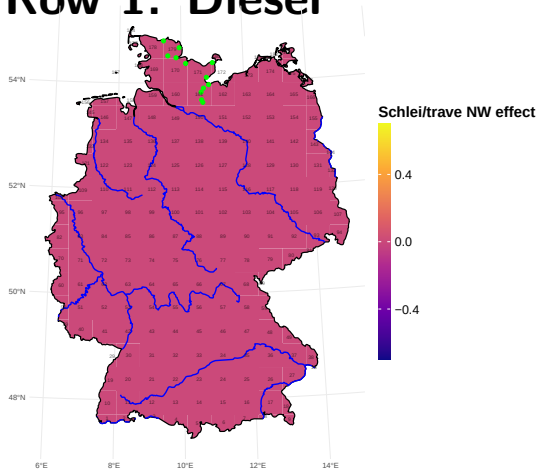


non-navigable

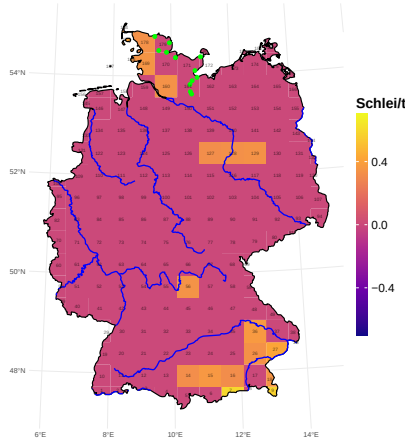
(Preliminary) Results

Schlei/Trave – Contemporaneous Extreme Water Event – 5% sig. level

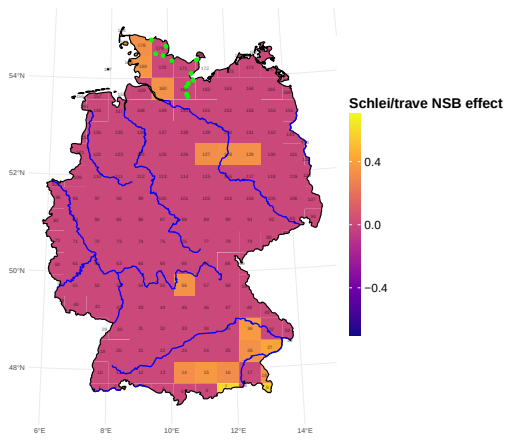
Row 1: Diesel



low water

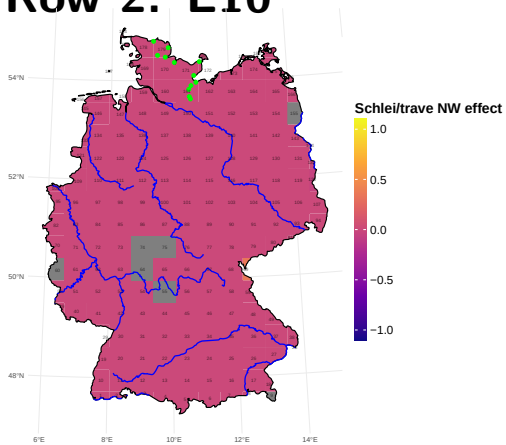


high water

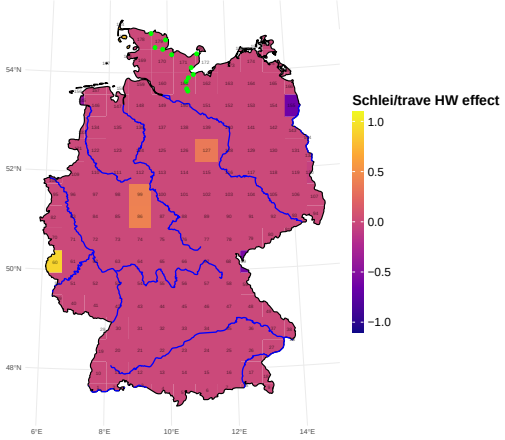


non-navigable

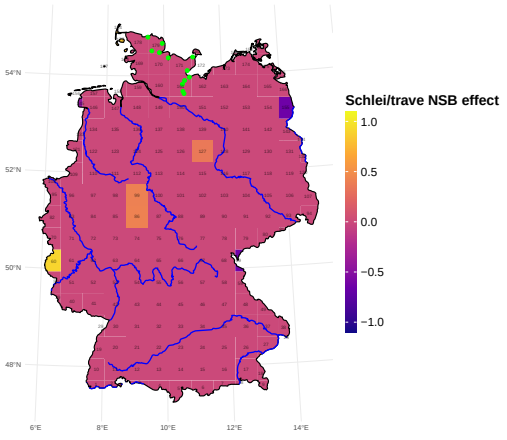
Row 2: E10



low water



high water

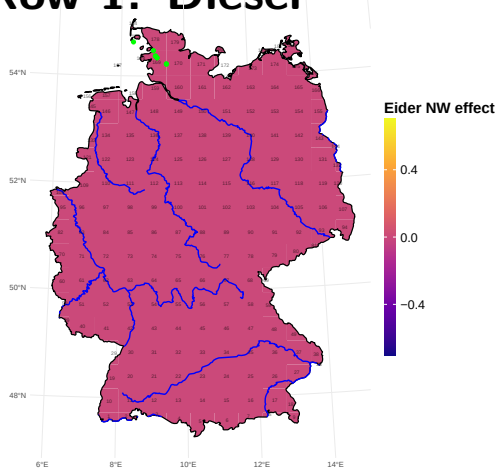


non-navigable

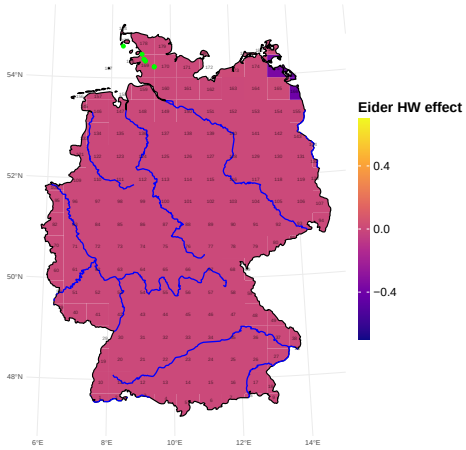
(Preliminary) Results

Eider – Contemporaneous Extreme Water Event – 5% sig. level

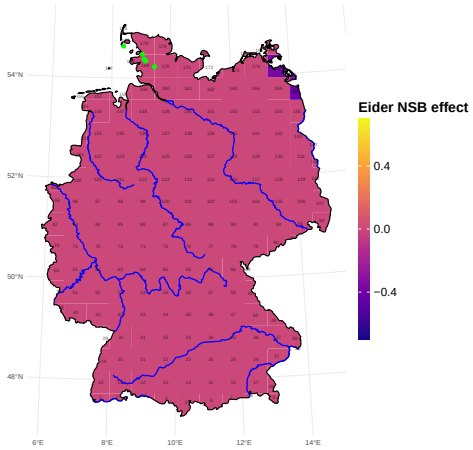
Row 1: Diesel



low water

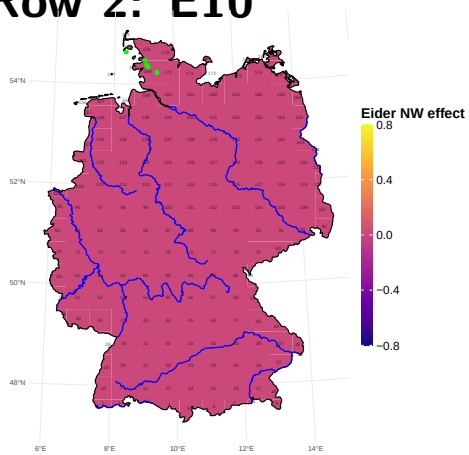


high water

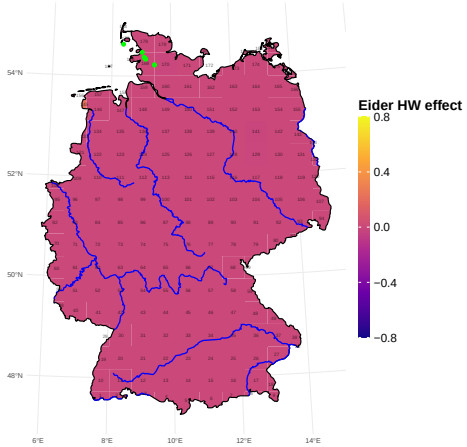


non-navigable

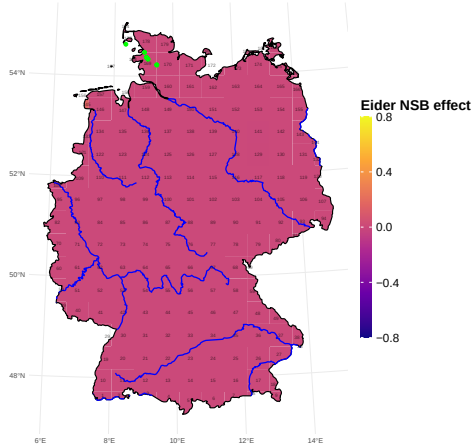
Row 2: E10



low water



high water

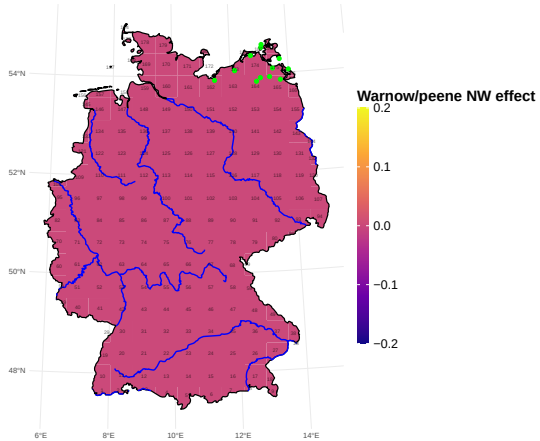


non-navigable

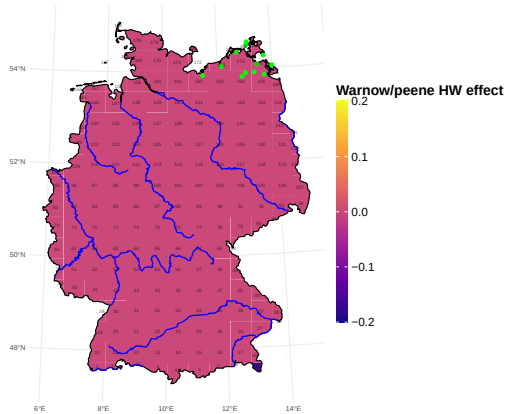
(Preliminary) Results

Warnow/Peene – Contemporaneous Extreme Water Event – 5% sig. level

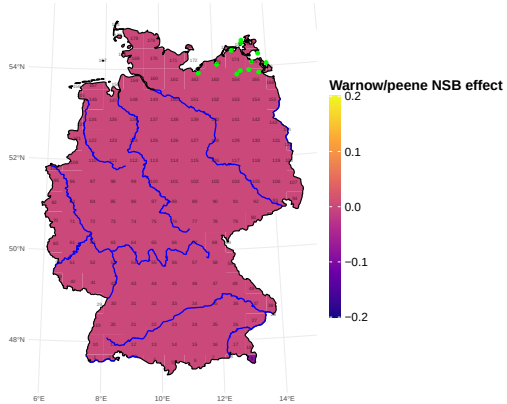
Row 1: Diesel



low water

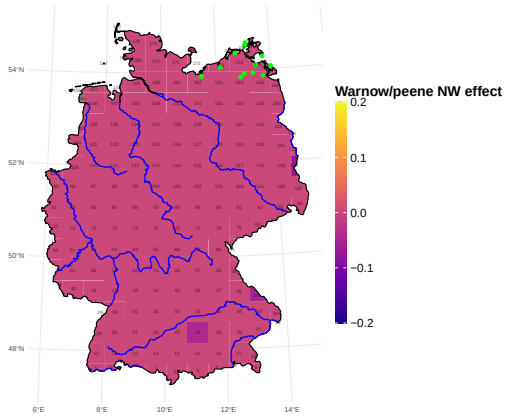


high water

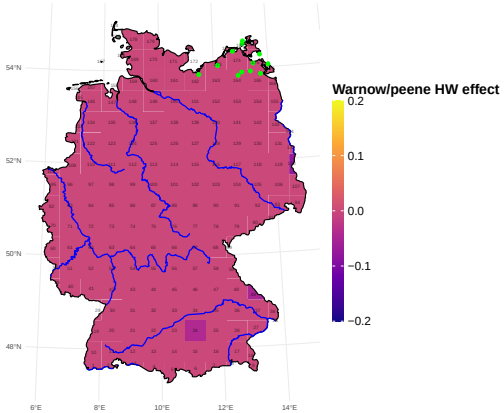


non-navigable

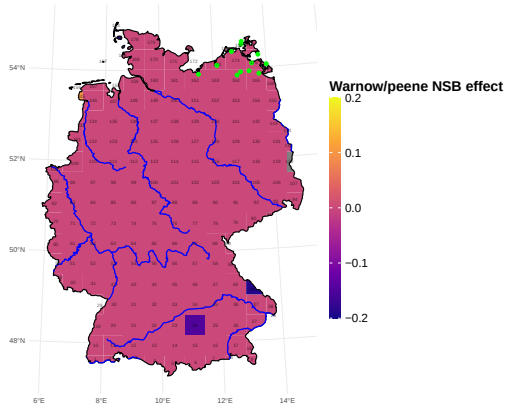
Row 2: E10



low water



high water



non-navigable

Open Questions & Future Work

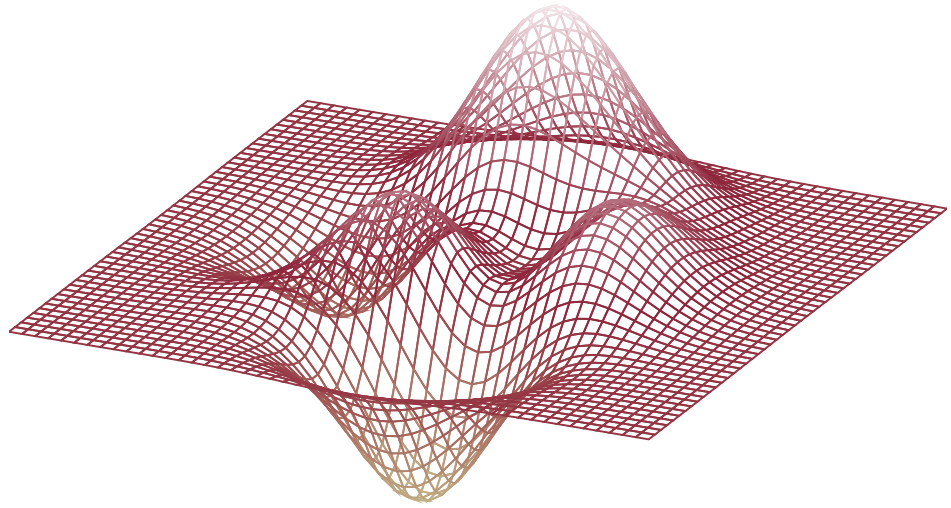
Did inland water transportation lose importance?

- Low water levels (e.g. 2018, 2022): hinder large barge operations, forcing cargo to road/rail
  $\Rightarrow$ freight quantities or historical AIS data
- Aging fleet & infrastructure: old vessels, insufficient channel depths and bridge underpasses reduce competitiveness
- Road transport growth: nowadays trucks carry far more freight; IWT share of tonnage is only $\sim 10 - 12\%$ overall
- Policy & modal shift lacking: Despite environmental advantages, limited investment has slowed the shift from road to IWT

How much does a 7 day draught / flood cost at the gas station?

Conclusion on New Data & Big Data

- Data collection becomes fast.
- Capability of handling big data is key.
- Standard packages like `fixest` have too high resource requirements.
- GPU ressources are not commonly used in traditional Econometrics – possible interesting avenue: Fixed effects without the inverse / neural net set up – wider data input possible
- Non-standard methodological contributions become more interesting.



Transfer Entropy and Conditional Density Estimation

with Smoothed Quantile Regressions

Bleher (2024a,b), Bleher, Koch and Dimpfl (2024)

Mutual Information and Transfer Entropy

Mutual Information

$$I(X,Y,Z) = \iiint_{\mathbb{R}^3} f_{X,Y,Z}(x,y,z) \log\left(\frac{f_{X,Y,Z}(x,y,z)}{f_X(x)f_Y(y)f_Z(z)}\right) dzdxdy$$

$$= \mathbb{E}\left[\log\left(\frac{f_{X,Y,Z}(x,y,z)}{f_X(x)f_Y(y)f_Z(z)}\right)\right]$$

Transfer Entropy Information

$$T_{X\to Y} = \mathbb{E}\left[\log\left(\frac{f_{Y_t|X_{t-1},Y_{t-1}}(y_t \mid x_{t-1}, y_{t-1})}{f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1})}\right)\right]$$

Toy Example

Simulated Vector Error Correction Model (VECM)

$$y_{1,t} = y_{1,t-1} + \varepsilon_{1,t}$$
$$y_{2,t} = y_{2,t-1} + \alpha\left(y_{1,t-1} - y_{2,t-1}\right) + \varepsilon_{2,t}$$

Parameter

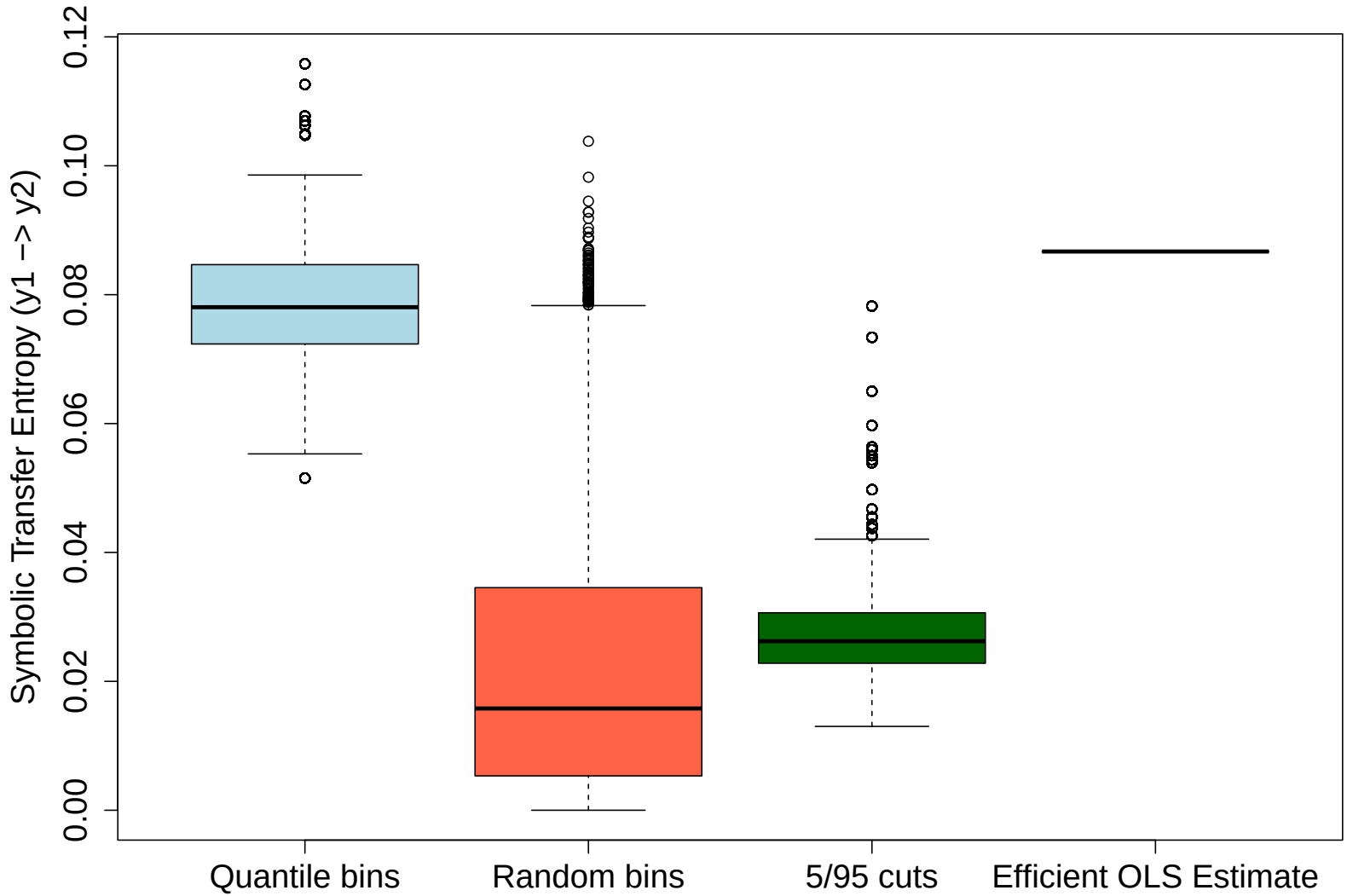
- i.i.d. $\varepsilon_{1,t}, \varepsilon_{2,t} \sim \mathcal{N}(0,1)$, $n = 2000$
- error correction coefficient: $\alpha = 0.5$
- shifted to non-zero values

Remarks

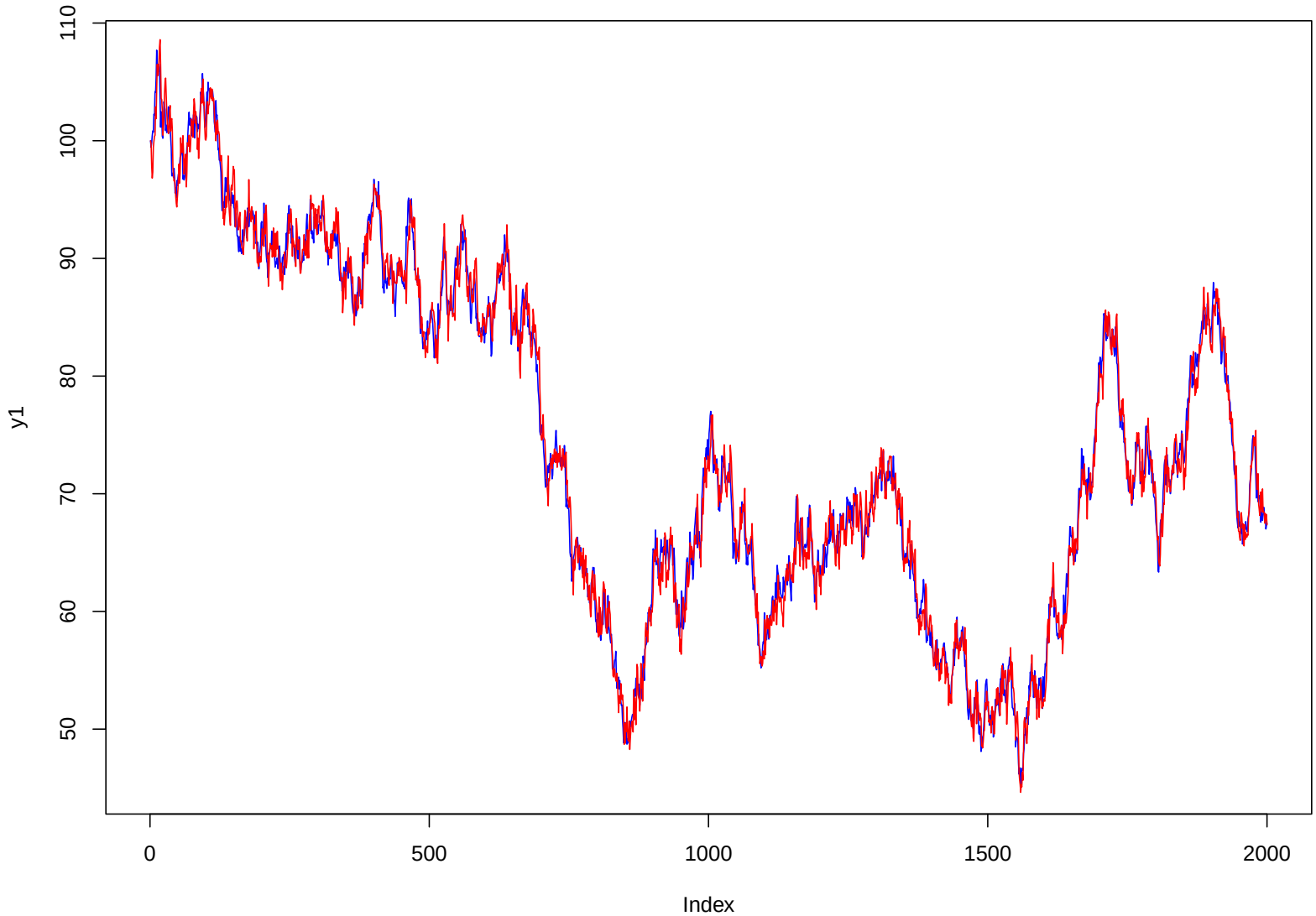
- One-sided causality $y_1 \to y_2$, i.e., $T_{Y_1 \to Y_2} \neq 0$
- Estimation via **Symbolic Transfer Entropy** (e.g., Behrendt, Dimpfl, Peter and Zimmermann 2018)
- Theoretical value here:

$$T_{Y_1 \to Y_2} = \frac{1}{2} \log \left(\frac{\mathrm{Var}\left(\Delta y_{2,t} \mid \Delta y_{2,t-1}\right)}{\mathrm{Var}\left(\Delta y_{2,t} \mid \Delta y_{2,t-1}, \Delta y_{1,t-1}\right)} \right) \approx 0.08$$

Problem 1: Distribution of Symbols Governs TE



Problem 2: Number of Symbols vs. Sample Size



Motivation

Calculate Transfer Entropy without throwing away information i.e. without discretization

(as done by Papana, Kyrtsou, Kugiumtzis, Diks et al. (2013), Behrendt et al. (2018))

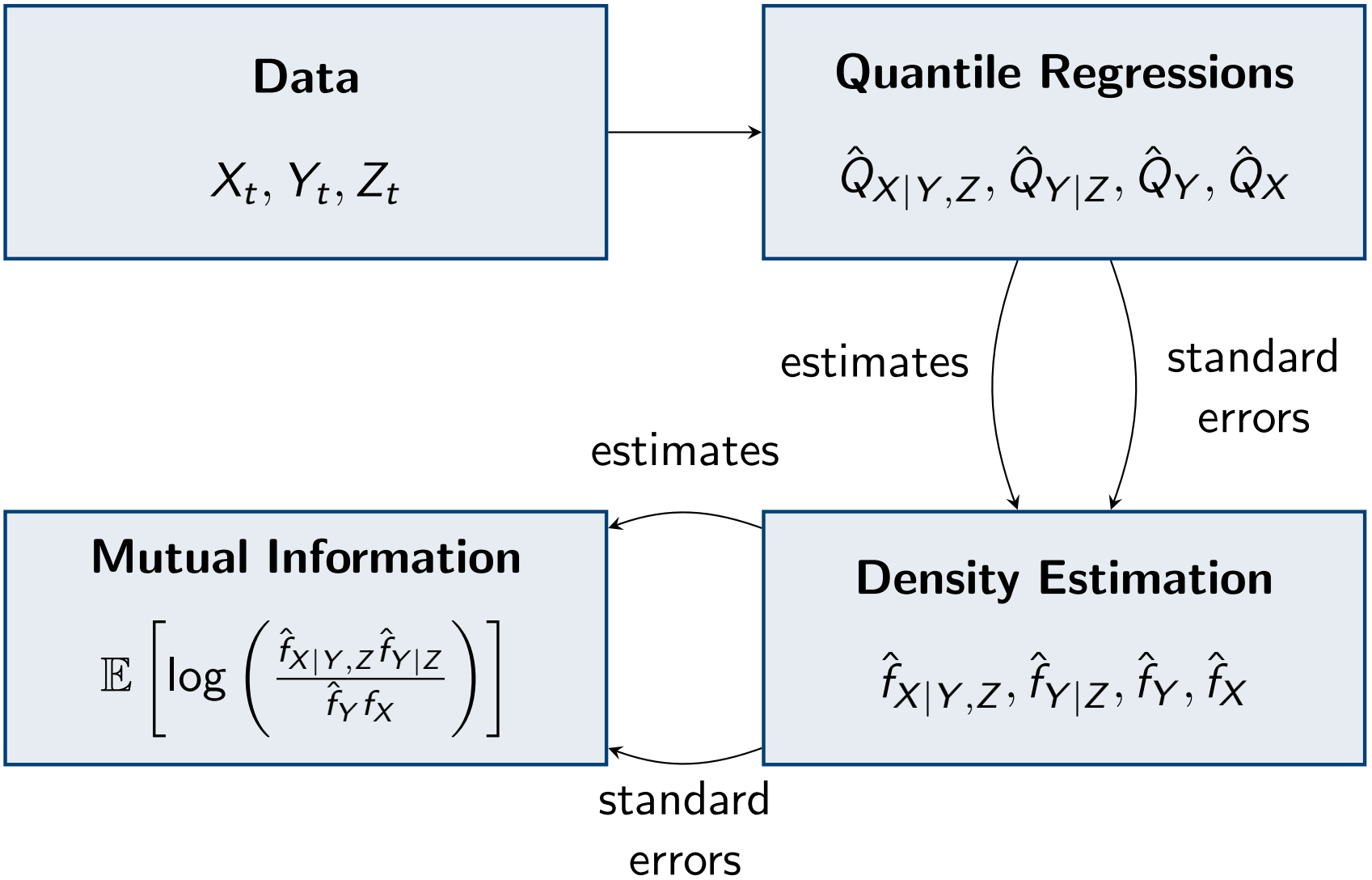
Measure information flow from data

- ... using as much information as possible.
- ... using as few assumptions as necessary.
- ... keeping computational cost low.
- ... craft a testable framework.

Interesting for:

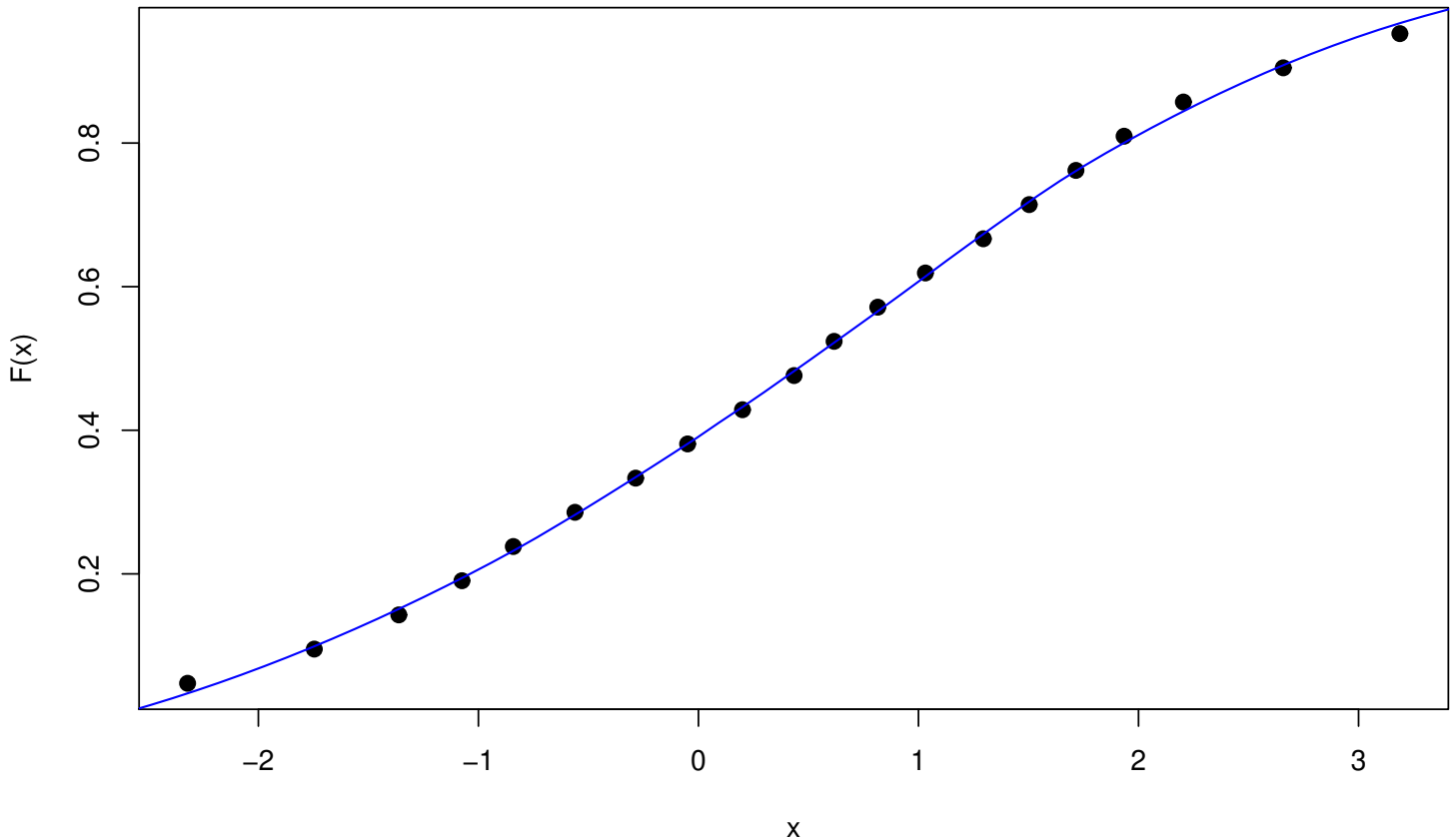
- variable selection (e.g., non-linear equivalent to stepwise regression)
- causality analysis (e.g., non-linear equivalent to Granger Causility in time series)
- conditional density estimation

The Project


Data
X_t, Y_t, Z_t
Quantile Regressions
$\hat{Q}_{X|Y,Z}, \hat{Q}_{Y|Z}, \hat{Q}_Y, \hat{Q}_X$
estimates
standard errors
estimates
Mutual Information
$\mathbb{E}\left[\log\left(\frac{\hat{f}_{X|Y,Z}\hat{f}_{Y|Z}}{\hat{f}_Y f_X}\right)\right]$
Density Estimation
$\hat{f}_{X|Y,Z}, \hat{f}_{Y|Z}, \hat{f}_Y, \hat{f}_X$
standard errors

Density estimation

Figure: Estimated and smoothed quantiles



Insert extra points at upper and lower tails when smoothing

Quantile regression

$$F_j = \min_{\boldsymbol{\theta} \in \mathbb{R}} \sum_{i=1}^{N} \rho_{\tau_j}(y_i - \mathbf{x}_i' \boldsymbol{\theta}_j)$$

where $\rho_{\tau_j}(u) = u\,(\tau_j - \mathbb{1}(u < 0))$ with $\mathbb{1}(\cdot)$ as the indicator function or in this case essentially a Heaviside step function.

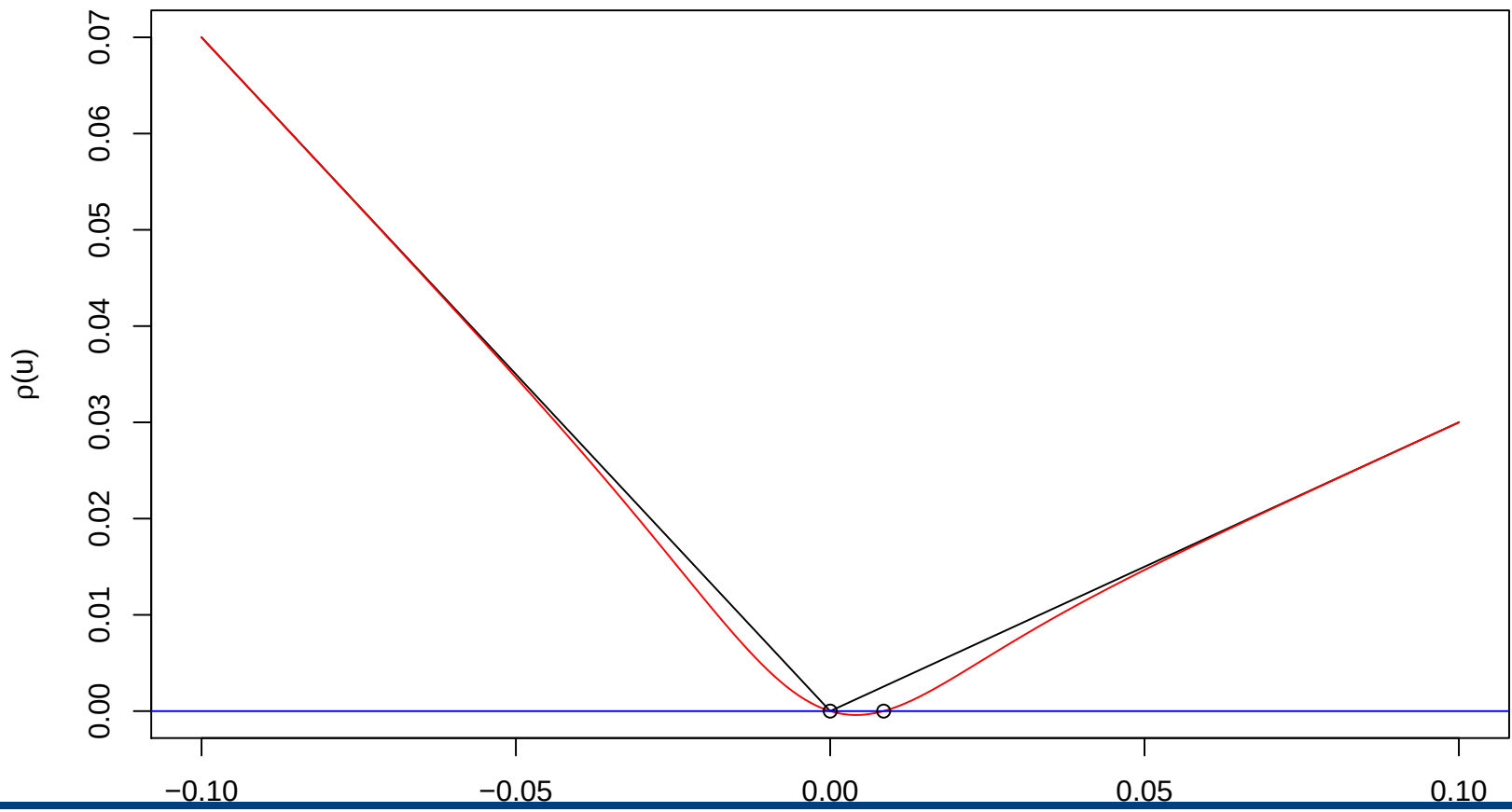
Fitted quantiles can be calculated then by

$$\mathbf{Q}_{\mathbf{y}}(\tau_j \mid \mathbf{X}) = \mathbf{X}' \hat{\boldsymbol{\theta}}(\tau_j)$$

Problem: Joint standard error of several QR on same data?
Solution: GMM asymptotics

Crafting GMM Standard Errors

Substitute the Heaviside step function $\mathbb{1}(u < 0)$ with the sigmoid function $\mathbf{1}(u) = (1 + e^u)^{-1}$ (sigmoid function)



Crafting GMM Standard Errors

GMM objective function:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbf{g}_N(\boldsymbol{\theta})'\mathbf{W}\mathbf{g}_N(\boldsymbol{\theta})$$

First Order Conditions of QR:

$$\frac{\partial F_j}{\partial \theta_k} = \sum_{i=1}^{N} x_{ik}\left(\tau_j - \frac{1}{1+e^{Gu_i}}\right) - x_{ik}\,u_i\,G\frac{e^{Gu_i}}{(1+e^{Gu_i})^2} \stackrel{!}{=} 0$$

Collecting the empircal sampling errors of moment conditions

$$g_l(\tau_j, \boldsymbol{\theta}_j) = \mathbb{E}[X_l\,(\tau_j - \mathbf{1}(U<0)) - X_l\,U\,\mathbf{1}(U=0)]$$

Crafting GMM Standard Errors

$$\mathbf{g}_N(\boldsymbol{\theta}_j) = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} g_{1i}(\tau_j, \hat{\boldsymbol{\theta}}_j) \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} g_{Ki}(\tau_j, \hat{\boldsymbol{\theta}}_j) \end{pmatrix}$$

Joint asymptotic normality of quantile regression estimates

$$\sqrt{N}(\text{vec}(\hat{\boldsymbol{\theta}}) - \text{vec}(\boldsymbol{\theta})) \sim \mathcal{N}\left(0, \text{Avar}(\hat{\boldsymbol{\theta}})\right)$$

where

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) = (\hat{\mathbf{D}}'\mathbf{W}\hat{\mathbf{D}})^{-1}\,\hat{\mathbf{D}}'\mathbf{W}\,\hat{\mathbf{S}}\,\mathbf{W}\hat{\mathbf{D}}\,(\hat{\mathbf{D}}'\mathbf{W}\hat{\mathbf{D}})^{-1}$$

Smoothing

Locally Weighted Polynomial Regression

Given a point $P = (\hat{\boldsymbol{\theta}}\mathbf{x}_0, y_0)$

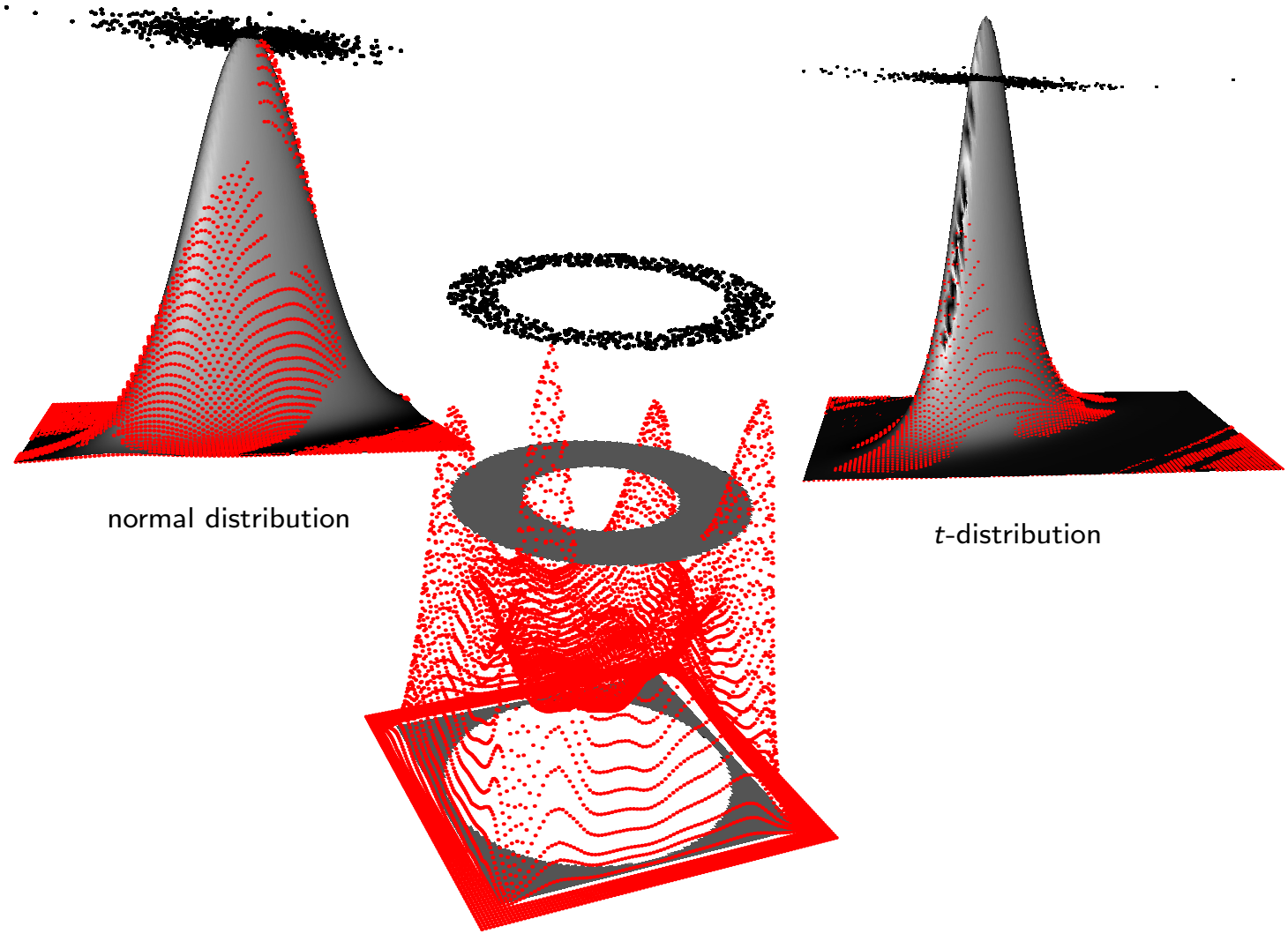
$$\hat{\gamma} = (\mathbf{Z}_P'\mathbf{W}_P\mathbf{Z}_P)^{-1}\mathbf{Z}_P'\mathbf{W}_P\boldsymbol{\tau}$$

where

- $\underset{Q\times p}{\mathbf{Z}_P} = \left(\boldsymbol{\iota}, \hat{\boldsymbol{\theta}}\mathbf{x}_0 - y_0, (\hat{\boldsymbol{\theta}}\mathbf{x}_0 - y_0)^2, \ldots, (\hat{\boldsymbol{\theta}}\mathbf{x}_0 - y_0)^p\right)$
- $\boldsymbol{\iota}$ as the column vector of ones
- $\mathbf{W}_P = \text{diag}\left(h^{-1}K(\frac{\hat{\boldsymbol{\theta}}\mathbf{x}_0 - y_0}{h})\right)$
- $K(\cdot)$ denotes the weight function
- h is the approximately optimal bandwidth

$\Rightarrow$ $\hat{\gamma}_1$ is a density estimate at point P

Smoothing



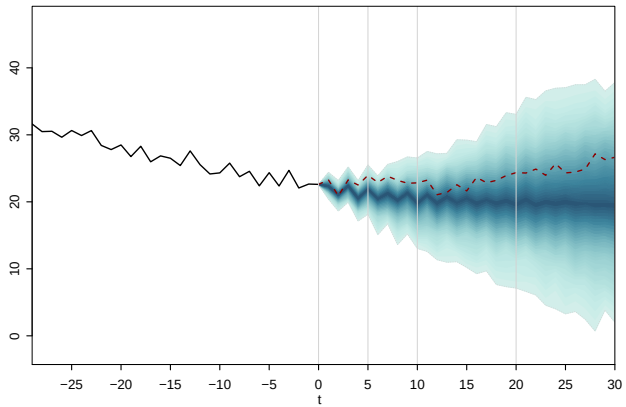
normal distribution

t-distribution

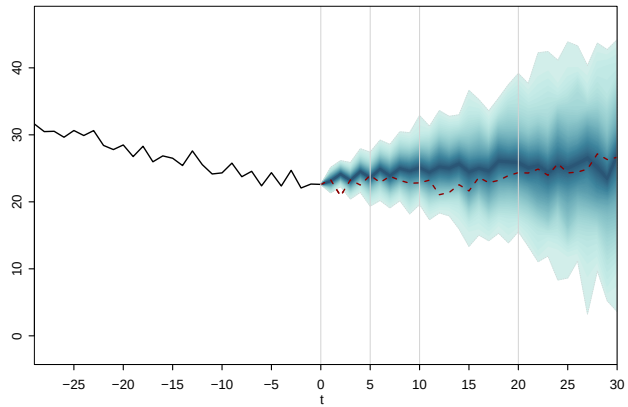
uniform distribution on a $2D$ shell

Application: CDE Forecasts

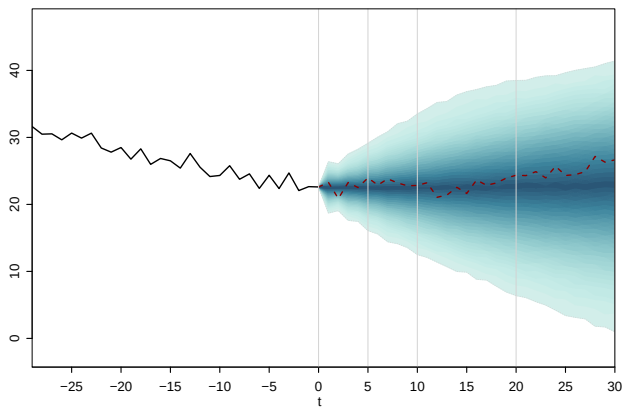
Fan Charts (correctly specified)



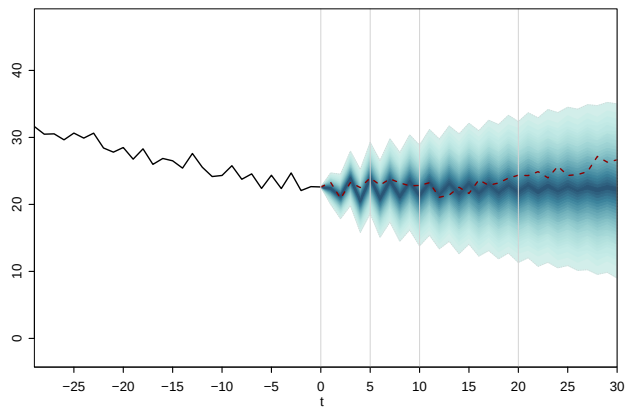
(a) LQP



(b) QVAR



(c) CKDE

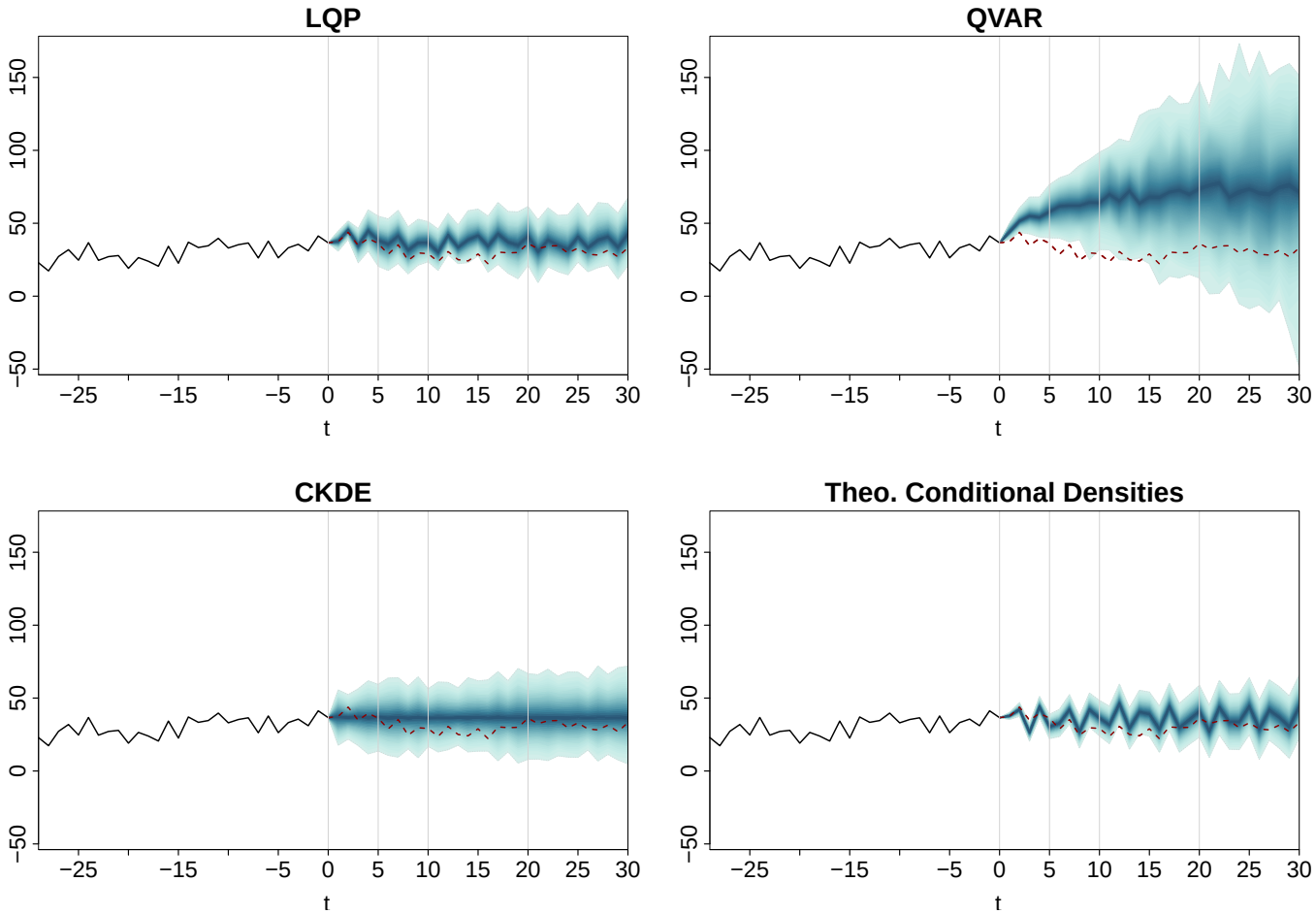


(d) Theoretical Distribution

Plots depict the I(1) of simulated VAR(p) processes and corresponding forecast.
red line: continuation of the simulated process up to $H = 30$.

Application: CDE Forecasts

Fan Charts (misspecified scenario)

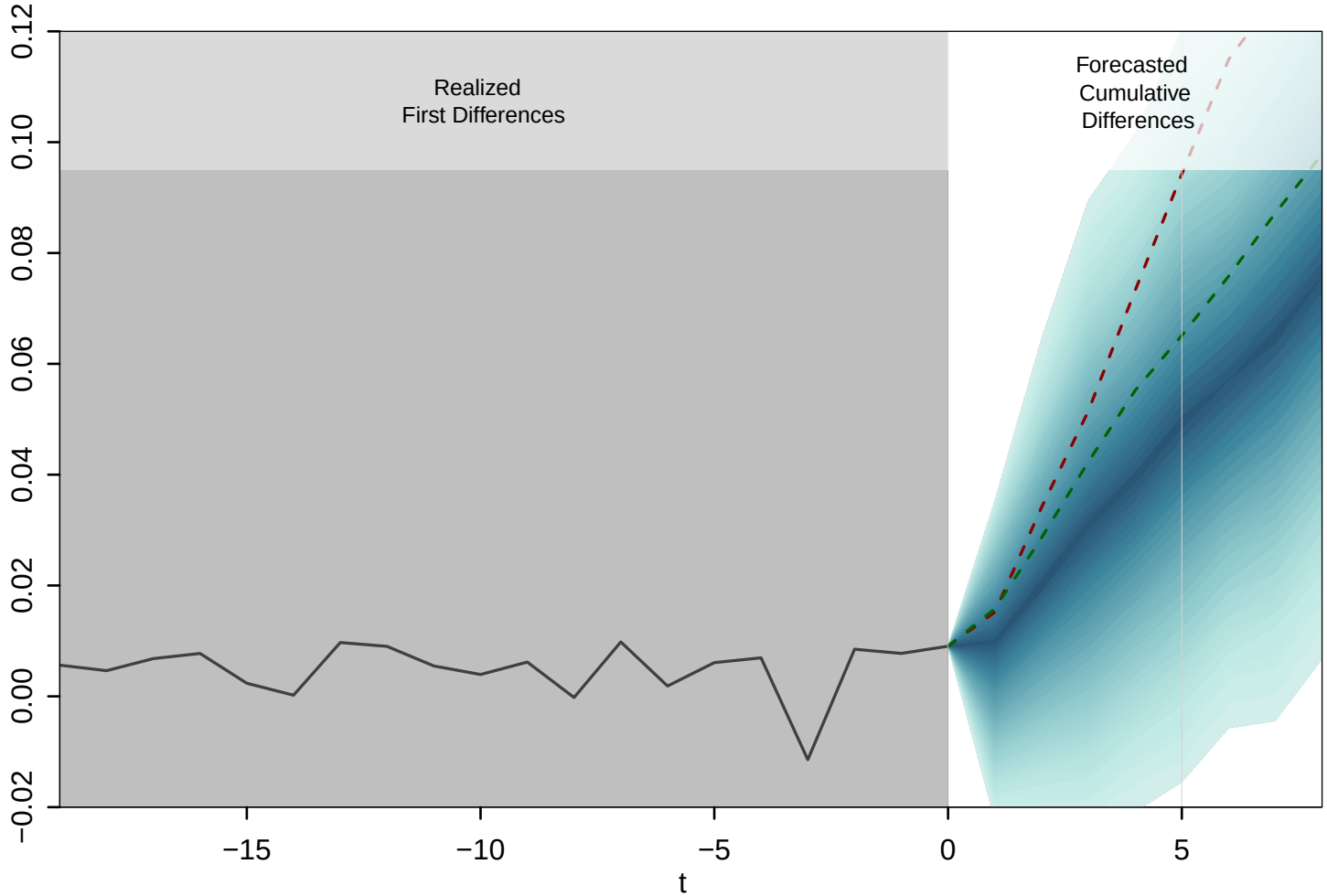


Plots depict the I(1) of simulated VAR(p) processes and corresponding forecast.

red line: projected continuation of the simulated process up to $H = 30$.

Application: US Inflation Forecasts

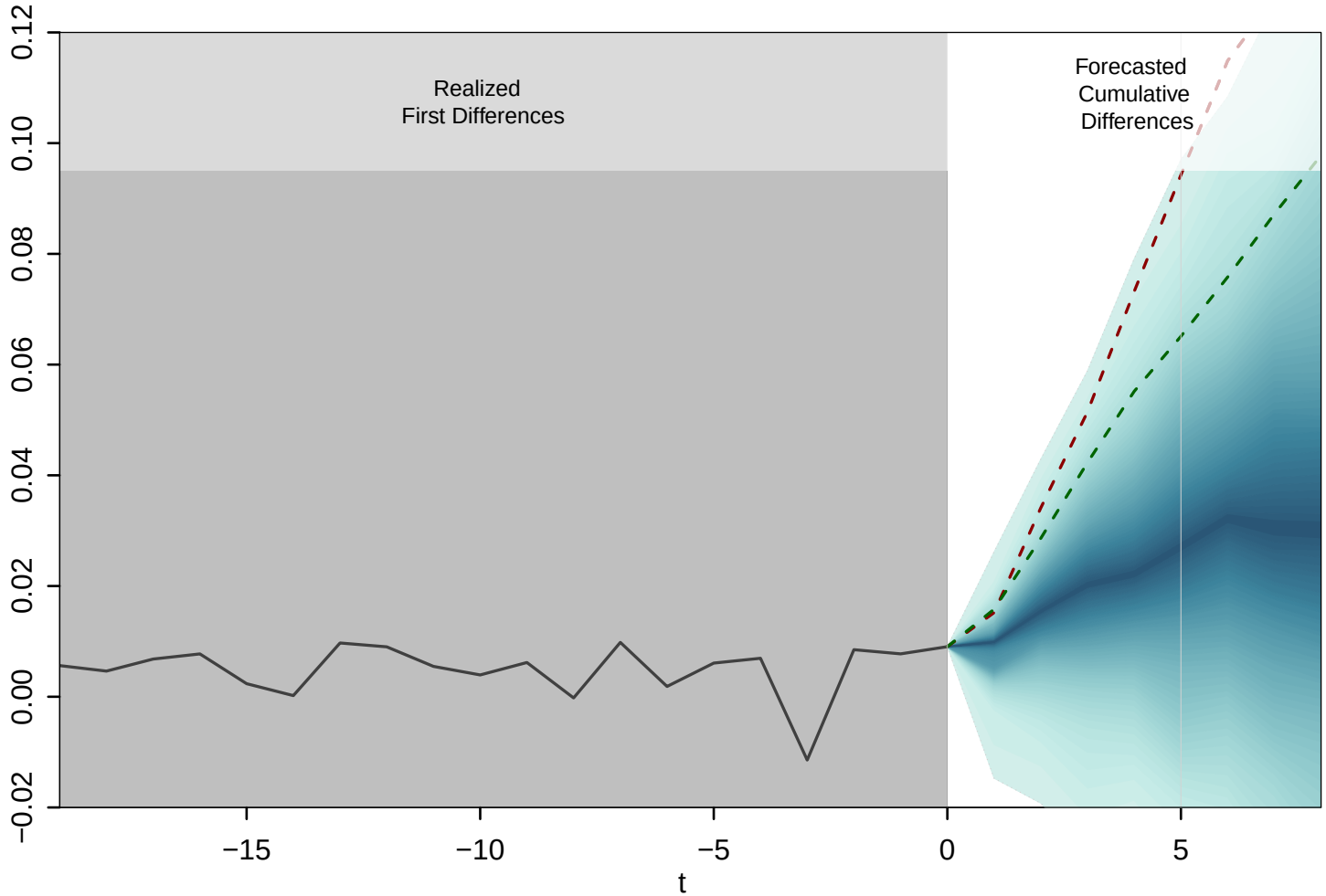
US Inflation Forecast as of 2020-Q4 (LQP)



The dashed red line illustrates the actual inflation log growth. The green line depicts the inflation log growth forecast of the benchmark model.

Application: US Inflation Forecasts

US Inflation Forecast as of 2020-Q4 (QVAR)



The dashed red line illustrates the actual inflation log growth. The green line depicts the inflation log growth forecast of the benchmark model.

Standard Errors of Densities

Via the Delta-Method

$$\lim_{N\to\infty} \sqrt{Qh^3} \frac{\left(\hat{f}(\hat{\boldsymbol{\theta}}) - f\right)}{\text{vec}(\hat{\boldsymbol{\theta}}) - \text{vec}(\boldsymbol{\theta})} \approx \left.\frac{\partial \hat{f}}{\partial \bar{\boldsymbol{\theta}}}\right|_{\bar{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}}.$$

follows

$$\text{Avar}\left(\hat{\gamma}_1\left(\boldsymbol{\theta}\right)\right) \approx \frac{1}{NQh^3} \text{vec}\left(\mathbf{H}\right)' \text{Avar}(\hat{\boldsymbol{\theta}}) \text{vec}\left(\mathbf{H}\right)$$

- $\mathbf{H}$ derivative of density estimate wrt $\hat{\boldsymbol{\theta}}$
  - ✓ estimation uncertainty from QR
  - ✓ estimation of bandwidth
  - ✓ ghost points

Standard Errors of Mutual Information

$$\hat{I}(X,Y,Z) = \mathbb{E}\left[\log\left(\frac{\hat{f}_{X,Y,Z}(x,y,z)}{\hat{f}_X(x)\hat{f}_Y(y)\hat{f}_Z(z)}\right)\right] = E[\hat{C}_i]$$

Again use the Delta Method for each contribution C_i

$$\lim_{N\to\infty}\sqrt{Q}\frac{\hat{C}_i(\hat{\boldsymbol{\theta}}) - C_i - C_i^*}{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}} = \lim_{N\to\infty}\frac{1}{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}}\sqrt{Q}\log\left(\frac{h_{X|Y,Z}^{-3/2}(\hat{f}_{X|Y,Z} - f_{X|Y,Z})h_{Y|Z}^{-3/2}(\hat{f}_{Y|Z} - f_{Y|Z})}{h_X^{-3/2}(\hat{f}_X - f_X)h_Y^{-3/2}(\hat{f}_Y - f_Y)}\right)$$

$$= \left.\frac{\partial\hat{C}_{x_i,y_i,z_i}}{\partial\bar{\theta}_{lm}}\right|_{\bar{\theta}_{lm}=\hat{\theta}_{lm}},$$

For each contribution (approximately)

$$\hat{C}_i\left(\hat{\boldsymbol{\theta}}\right) + C_i^* \sim \mathcal{N}\left(C_\mathbf{i}, \frac{1}{QN}\operatorname{vec}\left(\boldsymbol{\Upsilon}_\mathbf{i}\right)'\operatorname{Avar}(\hat{\boldsymbol{\theta}})\operatorname{vec}\left(\boldsymbol{\Upsilon}_\mathbf{i}\right)\right).$$

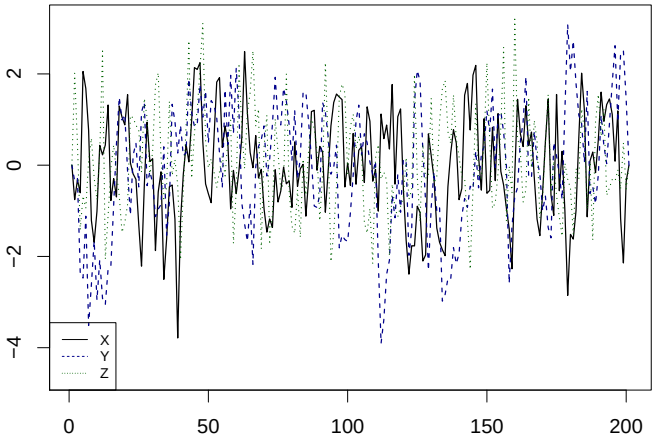
Standard Errors of Mutual Information

$$\text{Var}\left(\hat{I}_{X,Y,Z}\right) = \frac{1}{QN}\left[\frac{1}{N}\sum_{i=1}^{N}\text{vec}\left(\Upsilon_i\right)\right]' \text{Avar}(\hat{\theta})\left[\frac{1}{N}\sum_{j=1}^{N}\text{vec}\left(\Upsilon_j\right)\right]$$

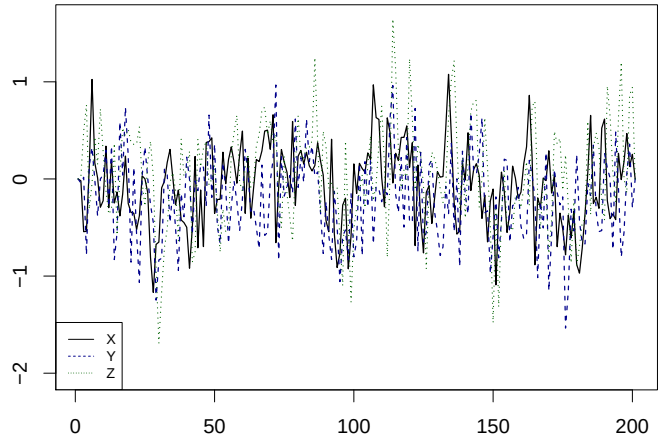
- Υ derivative of MI estimate wrt $\hat{\theta}$
  - ✓ estimation uncertainty from QR
  - ✓ estimation of density via local polynomial regressions
  - ✓ estimation of bandwidth
  - ✓ ghost points

Simulation Results – $TE_{X \to Z}$

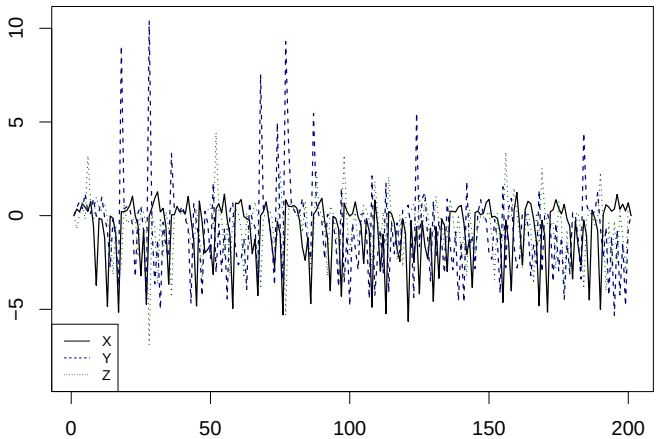
Simulated Time Series for TE estimation



Independent

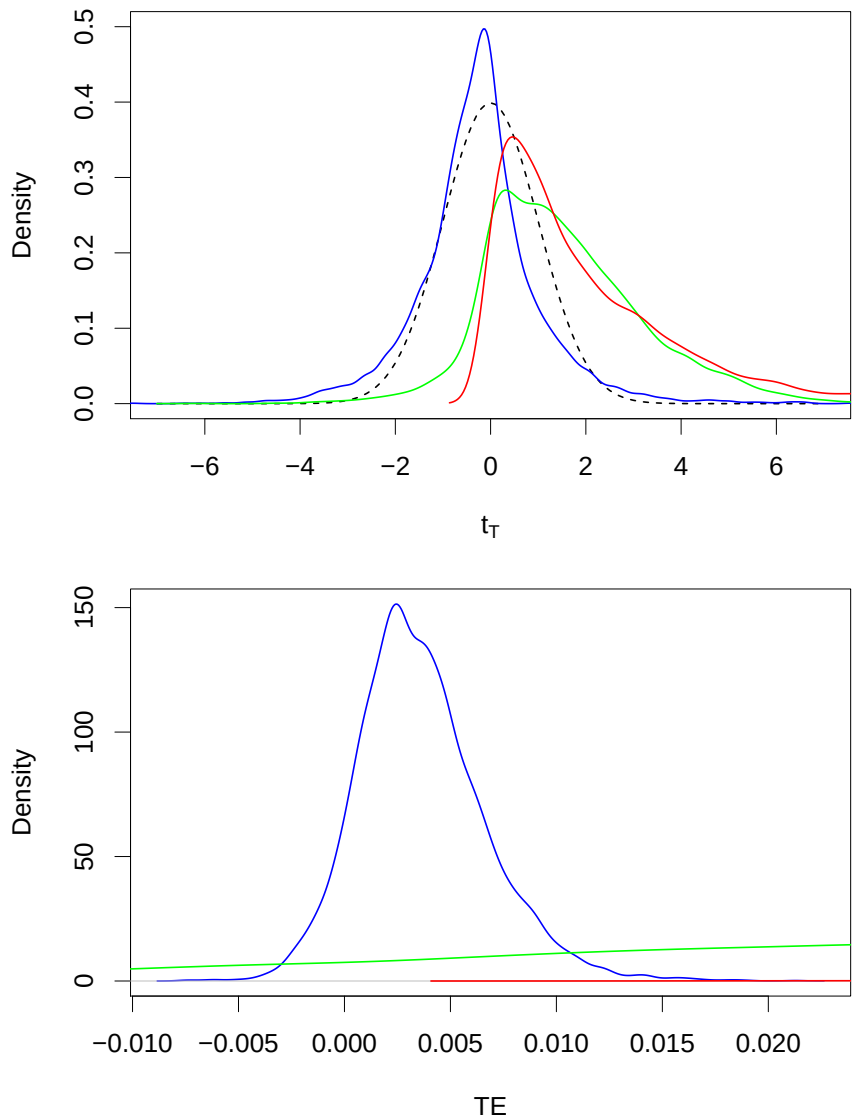


Linear



Non-Linear

Simulation Results – $TE_{X \to Z}$



Test Statistic

- standard normal
- independent
- linear
- heavily non-linear

Distribution of TE Values

Application – Transatlantic Information Flows

- Study by Dimpfl and Peter (2014)
- Transform one-minute intraday returns r_t of major blue-chip stock indices into symbols:
  - A if $r_t \leq r_{[0.05]}$
  - B if $r_{[0.05]} < r_t \leq r_{[0.95]}$
  - C if $r_t > r_{[0.95]}$
- Their result: significant symbolic transfer entropy in all directions.
- Knowing the previous minute's return category improves the prediction of the current one-minute return category

Application – Transatlantic Information Flows

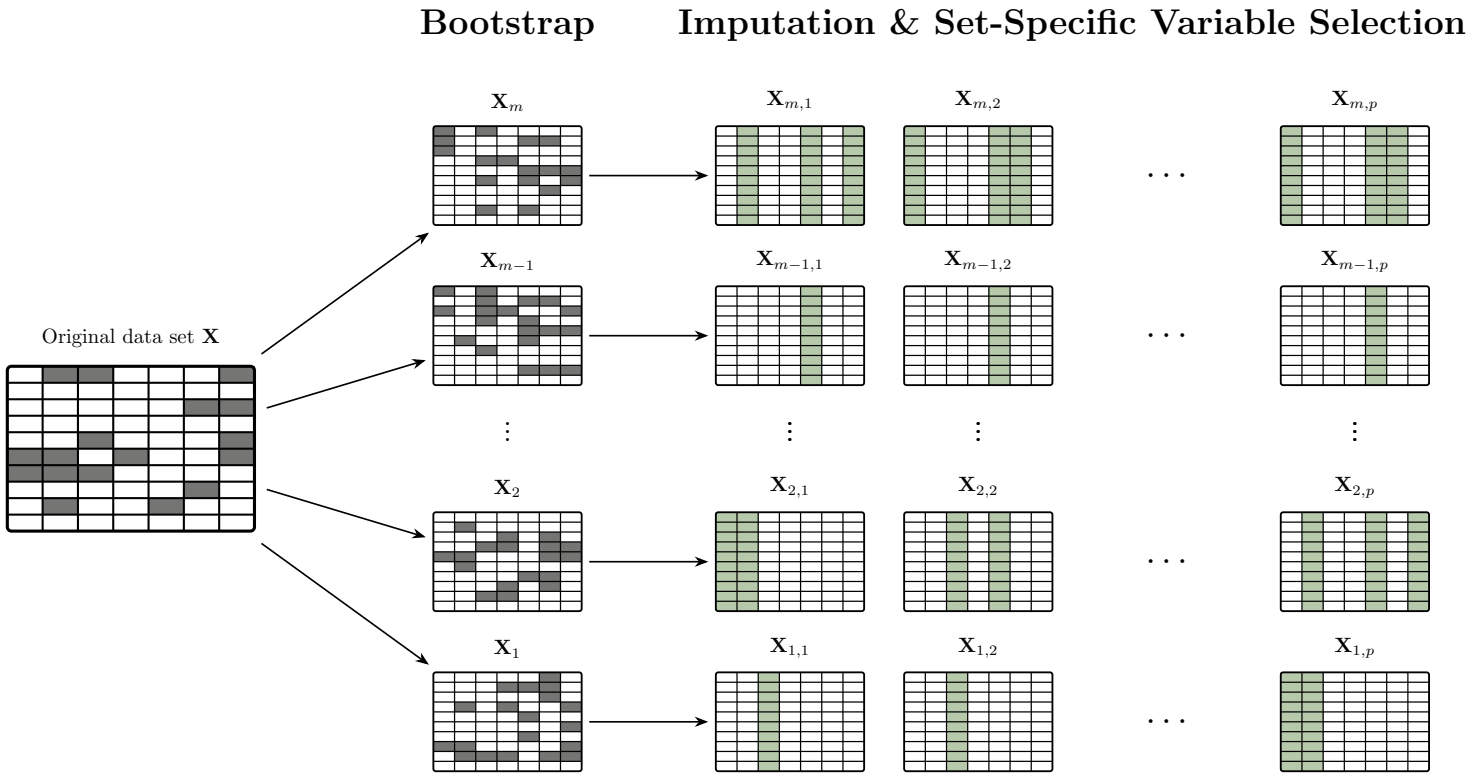
My results indicate predictive power of...

- the German market for the US market.
- the US market for the French market.

	$t_{T,\text{EU}\to\text{US}}$	$\hat{TE}_{\text{EU}\to\text{US}}$	$t_{T,\text{US}\to\text{EU}}$	$\hat{TE}_{\text{US}\to\text{EU}}$
DAX	**52.3082**	0.0411	1.4214	0.0081
CAC	-0.1721	-0.0003	**5.0460**	0.0150
FTSE	-1.8607	-0.0006	0.7634	0.0038

A Bayesian-Frequentist Approach to Post-Double Selection Regressions on Survey Data with Multiple Imputation

(WIP with C. Tarantola)



Bayes Factor

$$\text{Prob}(H_0 \mid -) = \frac{\text{Prob}(- \mid H_0) \cdot \text{Prob}\,(H_0)}{\text{Prob}\,(-)}$$

$$\text{Prob}(H_A \mid -) = \frac{\text{Prob}(- \mid H_A) \cdot \text{Prob}\,(H_A)}{\text{Prob}\,(-)}$$

Taking the ratio yields

$$\frac{\text{Prob}(H_0 \mid -)}{\text{Prob}(H_A \mid -)} = \frac{\text{Prob}(- \mid H_0)}{\text{Prob}(- \mid H_A)} \cdot \frac{\text{Prob}\,(H_0)}{\text{Prob}\,(H_A)}.$$

Generally, this can be written as

$$\text{(Posterior Odds)} = \text{(Bayes Factor)} \times \text{(Prior Odds)}.$$

Bayes Factor – Example

	+ CoViD	– No CoViD
H_0	TP	FN
H_A	FP	TN

Bayes Factor

$$BF = \frac{\text{Sensitivity}}{1 - \text{Specifity}}$$

$$\text{Sensitivity} := \frac{TP}{TP + FN}$$

$$\text{Specifity} := \frac{TN}{TP + TN}$$

Sensitivity $(\text{Ct} \leq 30) : 97.2\%$

Specificity $(\text{Ct} \leq 30) : 99.1\%$

Bayes Factor: $\frac{972}{9} = 108$

$$BF(+) \cdot \frac{\text{Prob}(CoViD)}{\text{Prob}(no\ CoViD)} = 108 \cdot \frac{1}{99} = \frac{108}{99}$$

$\Rightarrow$ Updated probability after 1 test: 52.17%

$\Rightarrow$ Updated probability after 2 test: 99.15%

Bayes Factor – Regression Context

Type I and II Error

	+ Not reject	− Reject
H_0	✓	α
H_A	β	✓

Bayes Factor

$$BF(-) = \frac{\text{Prob}(- \mid H_0)}{\text{Prob}(- \mid H_A)} = \frac{\alpha}{1-\beta} = \frac{\alpha}{\text{Power}} \text{ updates } \frac{\text{Prob}(H_0)}{\text{Prob}(H_A)}$$

Empirical Power (t-distribution)

With $t_{\text{crit}} = t_{1-\frac{\alpha}{2}, df}$ and non-centrality parameter $\hat{\lambda} = \frac{\hat{\beta}}{\text{s.e.}(\hat{\beta})}$

$$\widehat{\text{Power}} = 1 - F_t(t_{\text{crit}}; df, \hat{\lambda}) + F_t(-t_{\text{crit}}; df, \hat{\lambda})$$

Post Double Selection Framework

Simulation setup of Belloni, Chernozhukov and Hansen (2014):

$$y_i = D_i'\alpha_0 + \sum_{r=1}^{d}\sum_{l=1}^{g}\alpha_{r,l}D_{r,i}G_{r,i} + X_i'\beta + \sigma_y(D_i, G_i, X_i)\varepsilon_i$$

$$D_i = X_i'\gamma + \sigma_D(G_i, X_i)\nu_i$$

where

$$\sigma_y(D_i, G_i, X_i) = \sqrt{\frac{\left(1 + D_i'\alpha_0 + \sum_{r=1}^{d}\sum_{l=1}^{g}\alpha_{r,l}D_{r,i}G_{r,i} + X_i'\beta\right)^2}{\mathbb{E}\left[\left(1 + D_i'\alpha_0 + \sum_{r=1}^{d}\sum_{l=1}^{g}\alpha_{r,l}D_{r,i}G_{r,i} + X_i'\beta\right)^2\right]}}$$

$$\sigma_D(G_i, X_i) = \sqrt{\frac{\left(1 + X_i'\beta\right)^2}{\mathbb{E}\left[\left(1 + X_i'\beta\right)^2\right]}}$$

Post Double Selection Framework

Simulation Scenarios

Decision	**Scenario**	# Scenarios
Sample Size	$n = 100$ $n = 500$ $n = 1000$	3
Variance of DGP	homoscedastic heteroscedastic	2
R^2 of DGP	$R^2 = 20\%$ $R^2 = 60\%$	2
Missing Responses	missing completely at random (MCAR) missing at random (MAR) missing not at random (MNAR)	3
Degree of Missing Data	20% missing observations 40% missing observations 60% missing observations	3
Total		81 scenarios

Aggregate Bayes Factor

Calculation Algorithm

1. Use incomplete data set $\mathbf{X}$ to obtain M different bootstrapped and imputed data sets $\tilde{\mathbf{X}}_d$
2. Conduct double LASSO regressions on each $\tilde{\mathbf{X}}_d$.
3. Run post-double selection regression on each $\tilde{\mathbf{X}}_d$
4. For each coefficient j in the post-double selection regression, calculate the Bayes Factor

$$B_{dj} = BF(\text{reject} \mid \tilde{\mathbf{X}}_d)$$

5. Aggregate these Bayes-Factors

$$BF(\text{reject} \mid \mathbf{X}) = \left(\prod_{d=1}^{M} B_{dj} \right)^{1/M}$$

if β_j is not included in post-double selection in $\tilde{\mathbf{X}}_d$ set $B_{dj} = 1$.

Aggregate Bayes Factor

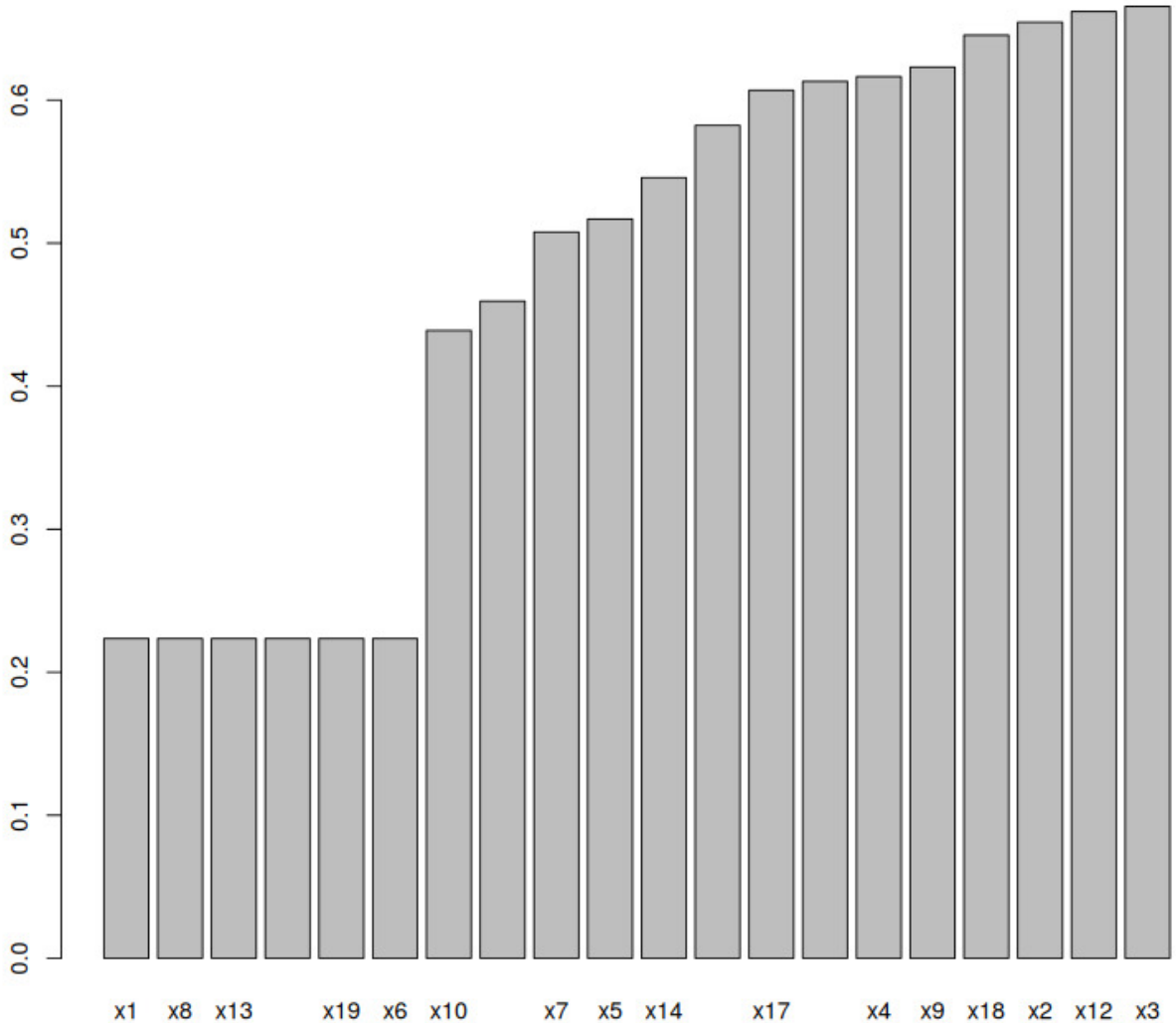
Decision Rule

Interpretation of Bayes factors
(adapted from Kass and Raftery (1995))

Bayes Factor (BF)	**Evidence Against H_0**
> 1	No evidence against H_0
1 to 1/3	Barely worth mentioning
!/3 to 1/10	Substantial
1/10 to 1/30	Strong
1/30 to 1/100	Very strong
$< 1/100$	Decisive

Aggregate Bayes Factor

Exemplary Simulation Results



Application: PIRLS data

Revisit:

- "Importance of students' social resources for their academic achievement and well-being in elementary school" (Schwerter, Stang-Rabrig, Kleinkorres, Bleher, Doebler and McElvany 2024)
- "Metropolitan, Urban, and Rural Regions: How Regional Differences Affect Elementary School Students in Germany" (Schwerter, Bleher, Doebler and McElvany 2025)

Conclusion Method Development

- Method development and code generation becomes faster with AI tools.
- Still heavy prerequisits are needed to use the tools effectively - understanding uncertainty, causality, sampling bias, etc. still requires robust statistical methodology.
- Efficient, robust and reliable methods are still worth doing research on.
- Specific problems, require specific solutions. AI often only provide the most frequent solutions.
- Methodological issues and new approaches for data analysis are still a fruitful field.

Overall Conclusion

- Research accelerates.
- Prerequisits will increase for interaction with AI and meaningful contributions.
- Mathematical and statistical reasoning provides reliability that AI cannot (yet).
- Computational infrastructure and resources become more important.
- The three circles of data science are becoming ever more important in research.

References

Behrendt, S., Dimpfl, T., Peter, F. and Zimmermann, D.: 2018, *RTransferEntropy: Measuring Information Flow Between Time Series with Shannon and Renyi Transfer Entropy*. R package version 0.2.7.

Belloni, A., Chernozhukov, V. and Hansen, C.: 2014, High-dimensional methods and inference on structural and treatment effects, *The Journal of Economic Perspectives* **28**(2), 29–50.

Bleher, J.: 2024a, Conditional density estimation and hypothesis testing through smoothed quantile regression, *Available at SSRN: http://dx.doi.org/10.2139/ssrn.5005311* .

Bleher, J.: 2024b, A quantile regression approach to calculating relative entropy measures. Available at SSRN: http://dx.doi.org/10.2139/ssrn.5005387.

Bleher, J. and Dimpfl, T.: 2022, Knitting multi-annual high-frequency google trends to predict inflation and consumption., *Econometrics and Statistics* **24**, 1–26.

References

Bleher, J., Koch, S. and Dimpfl, T.: 2024, Density forecasts with local quantile projection and quantile vector autoregression applied to inflation, *Available at SSRN: http://dx.doi.org/10.2139/ssrn.4855124* .

Dastgir, S., Demir, E., Downing, G., Gozgor, G. and Lau, C. K. M.: 2019, The causal relationship between bitcoin attention and bitcoin returns: Evidence from the copula-based granger causality test, *Finance Research Letters* **28**, 160–164.

Dimpfl, T. and Peter, F. J.: 2014, The impact of the financial crisis on transatlantic information flows: An intraday analysis, *Journal of International Financial Markets, Institutions and Money* **31**, 1–13.

Kass, R. E. and Raftery, A. E.: 1995, Bayes factors, *Journal of the American Statistical Association* **90**(430), 773–795.
URL: *http://www.jstor.org/stable/2291091*

Papana, A., Kyrtsou, C., Kugiumtzis, D., Diks, C. et al.: 2013, Partial symbolic transfer entropy, *University of Amsterdam* pp. 13–16.

References

Schwerter, J., Bleher, J., Doebler, P. and McElvany, N.: 2025, Metropolitan, urban, and rural regions: How regional differences affect elementary school students in germany, *AERA Open* **11**, 23328584251331453.

Schwerter, J., Stang-Rabrig, J., Kleinkorres, R., Bleher, J., Doebler, P. and McElvany, N.: 2024, Importance of students' social resources for their academic achievement and well-being in elementary school, *European Journal of Psychology of Education* **39**(4), 4515–4552.