Machine learning in official statistics: two case studies

Joep Burger

Data Science for Better Decisions, 17 December 2019

Tradition

Questionnaire + probability sample + design-based estimators

Pros

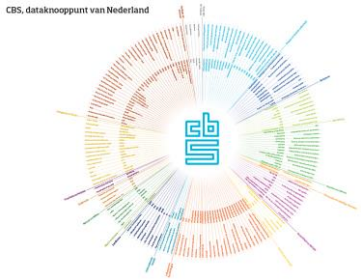
- In control
- Unbiased estimates
- Known variance

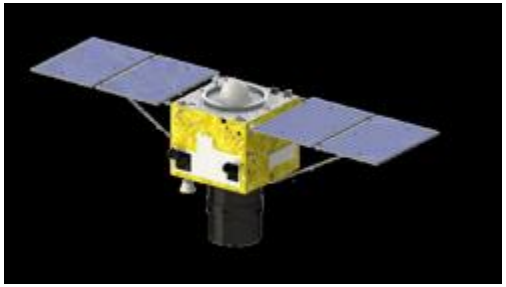
Cons

- Expensive
- Slow
- High-level
- In control?

Alternative

- Automatically generated digital data
  - Dutch System of Social statistical Datasets
  - Sensors, satellites, internet, …
- IT infrastructure
- AI





Research questions

1. Can we predict instead of ask about future behavior?

→ shorter questionnaires, reduce survey fatigue

2. Can we learn statistical information from images?

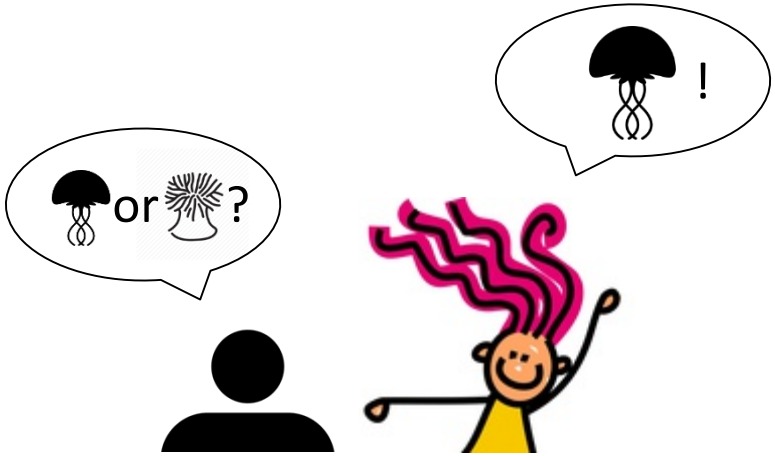
→ higher precision, higher frequency

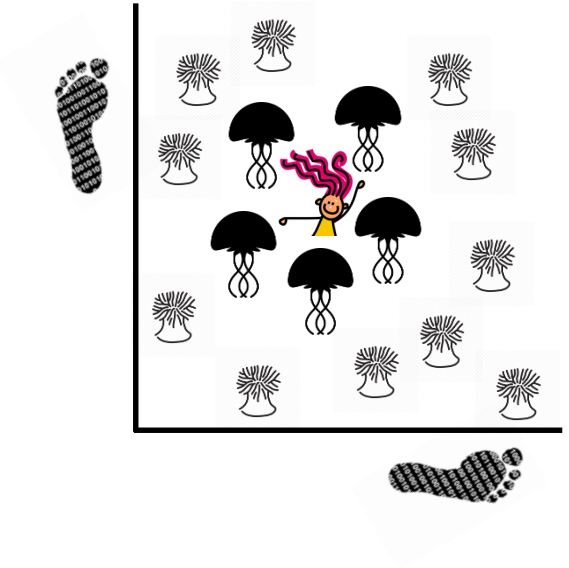
Case study 1: Dutch Housing Survey

- Three-yearly survey about housing conditions and needs among 18+
- ‘Do you want to move within 2 years?’
- Compare with actual moving behavior

by Aki Inomata

Ask or predict





Ask

Predict

Data

- Registered events 1995–2016
- 100k sample + 63k HS2015 respondents
- Features on persons, households, dwellings, regions
  - Time-independent
  - Status $j - u$
  - Time since latest change
  - Number of changes
- Binary target variable {moves,stays}

	1995	1996	1997	1998	…	2012	2013	2014	2015	2016
Optimize and Train		$j-17$	$j-16$	$j-15$	…	$j-1$	j	$j+1$		
Generalize				$j-17$	…	$j-3$	$j-2$	$j-1$	j	$j+1$

Methods

- Generalized linear model
  - Logit link
  - Main effects without (1) or with (2) interactions
  - Regularization: none (GLM), L1 (LAS), L2 (RDG)
- Random forest
  - $\sqrt{m}$ features per node
  - At least 20 observations per leaf
  - 500 trees
- Survival

Control

– Housing Survey

‘Do you want to move within 2 years?’	SVY1
1 Definitely not	0
2 Possibly, maybe	0.5
3 Would like to, cannot find anything	0.5
4 Definitely	1
5 I have already found other housing/dwelling	1
6 Don’t know	0.5

Quality measures

Probability [0,1]

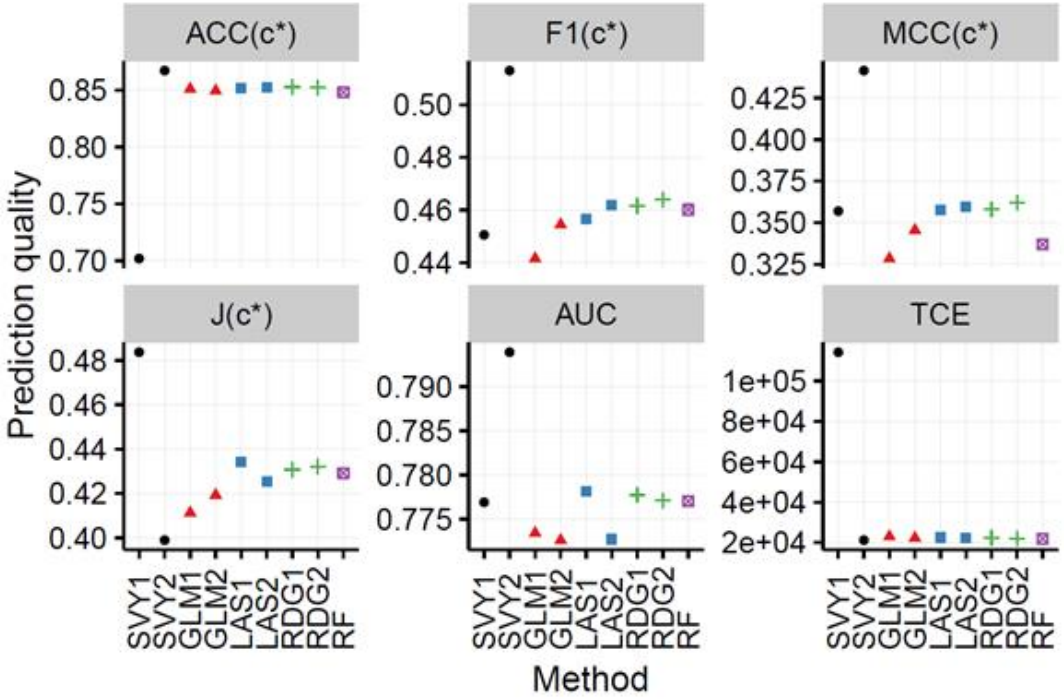
- Total cross entropy

Binary {moves,stays}

- Accuracy
- F1
- MCC
- J
- AUC

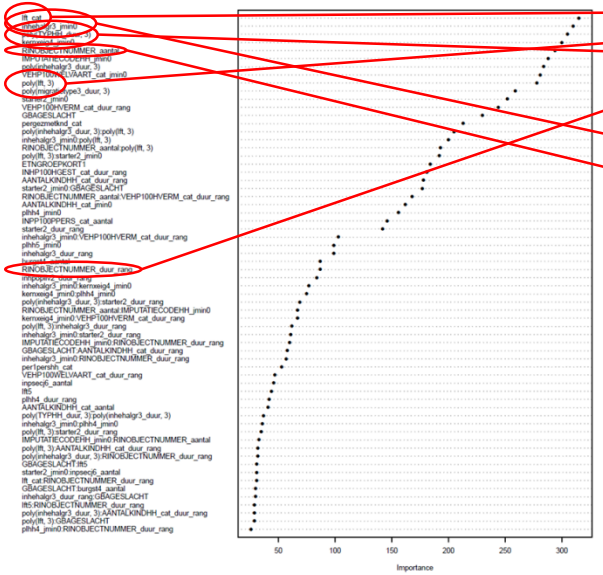
		Predicted		
		Moves	Stays	Total
Actual	Moves	TP	FN	AP
	Stays	FP	TN	AN
	Total	PP	PN	N

Compare HS2015 respondents

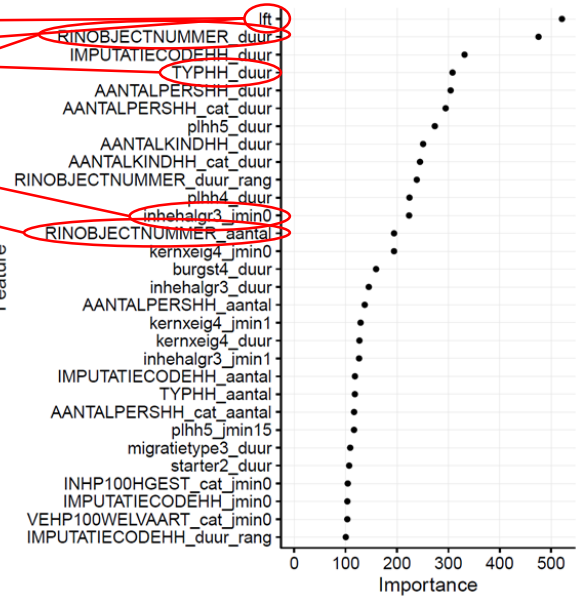


Interpretability

RDG2



Random forest



Conclusions

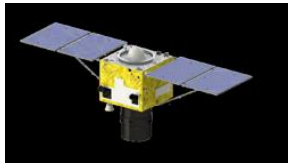
- Models predict actual moving behavior about equally well as respondents themselves
- Only subtle differences between methods
- Best method depends on quality measure

- Not limited to sample units
- Applicable to other target variables
- Useful for sampling design and weighting

Case study 2: imagery

- EU's Horizon2020 - MAKSWELL - WP3
- UNECE - ML in OS - WP1 - imagery

Input data



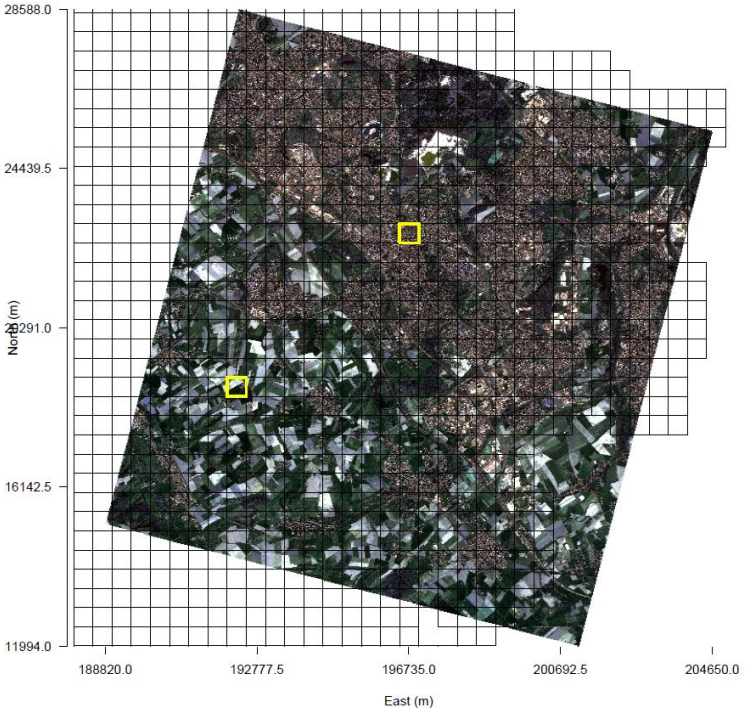
- Superview satellite images
- Pre-processed by Netherlands Space Office
- www.satellietdataportaal.nl
- 2019
- Resolution 0.5 m
- 4 channels (RGBI)
- 16-bit color palette [0,65535]
- Coordinate Reference System EPSG:28992 (Amersfoort / RD)

Output data

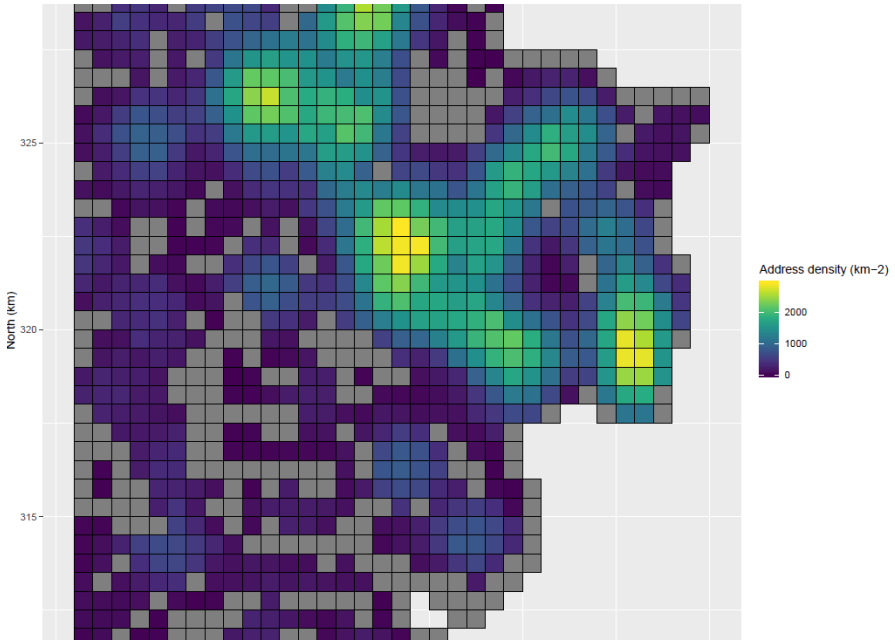
- Square statistics
- https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische%20data/kaart-van-500-meter-bij-500-meter-met-statistieken
- 500 m × 500 m (25 ha)
- Persons (e.g. origin, social-security benefits)
- Households (e.g. single-person, size)
- Dwellings (e.g. density, energy consumption)
- Coordinate Reference System EPSG:28992 (Amersfoort / RD)

Processing

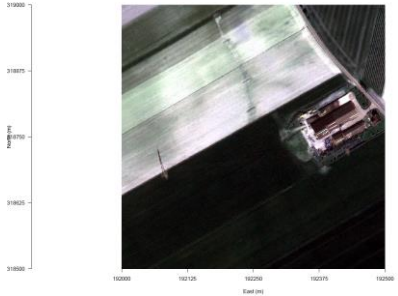
image + 25-ha grid



square statistics

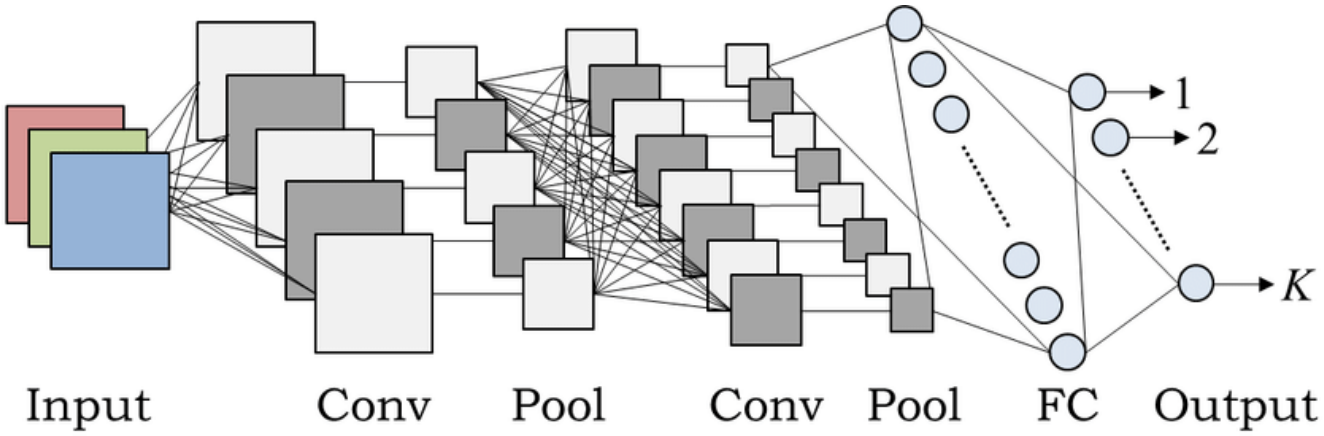


Intermediate result

ID	$X_{n\times1000\times1000\times4}$	$Y_{n\times1}$
E1965N3225R500		5 (‘very urban’)
E1920N3185R500		1 (‘rural’)
⋮	⋮	⋮

Convolutional neural network

– Multi-class classification



Hidaka & Kurita 2017

To do

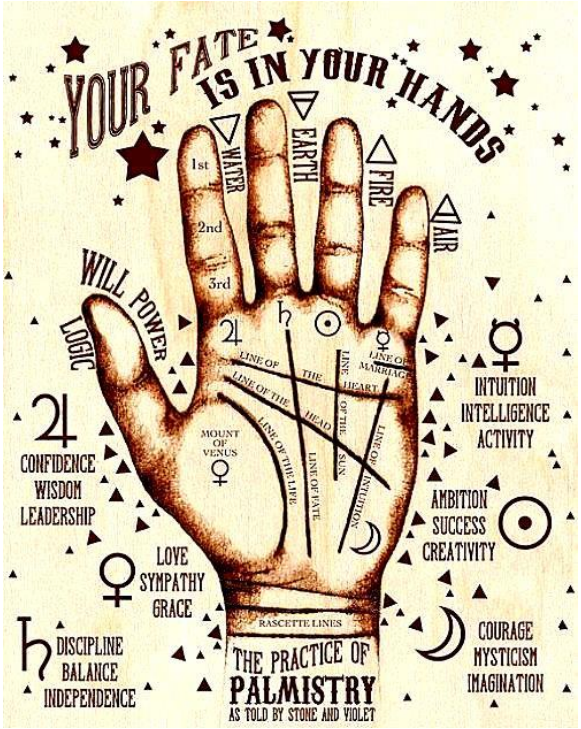
- Scale up/out
- Data augmentation and normalization
- Find pretrained CNN
- Optimize CNN architecture and hyperparameters
- Other grid (n 25-ha or 25n 1-ha)
- Other input (aerial, Landsat, street view)
- Other output (geotagged income, input features)

Data science for better decisions

- Explain or predict
- Process or output phase
- Define quality
- Crossover statistics - computer science

Acknowledgments

Bart Buelens
Tim de Jong
Yvonne Gootzen
Jacqueline van Beuningen
Saskia Janssen-Jansen
Jan van den Brakel
Marc Ponsen
Lyana Curier
Abel Coronado
Jimena Juarez



Discussion

- Operationalization
  - ‘Do you want to move?’
  - ‘Are you planning to move?’
- Generalization error

Future

- Predict quality by subpopulation
- Multinomial target variable
  - Type of dwelling
  - Location
- More features, including mobile phone data
- Multilevel models
- Train with complete population
- Survival analysis

Features

Person
- Age
- Gender
- Origin
- Migration status
- Marital status
- Position in household
- Socio-economic status
- First-time buyer
- Personal income

Household
- Household composition
- Household type
- Indicator imputed relationship
- Main source of household income
- Number of children living at home
- Household size
- Household income
- Household standard of living
- Household assets

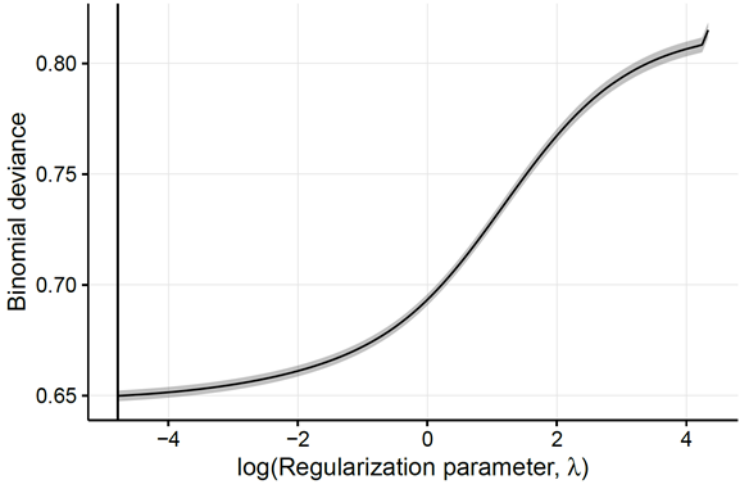
Dwelling
- Dwelling ID
- Ownership

Neighborhood
- Level of urbanization
- Proportion 65+
- Proportion born abroad
- Proportion imputed relationship
- Proportion with children
- Proportion single hh
- Proportion low income
- Proportion privately-owned
- Duration of residence

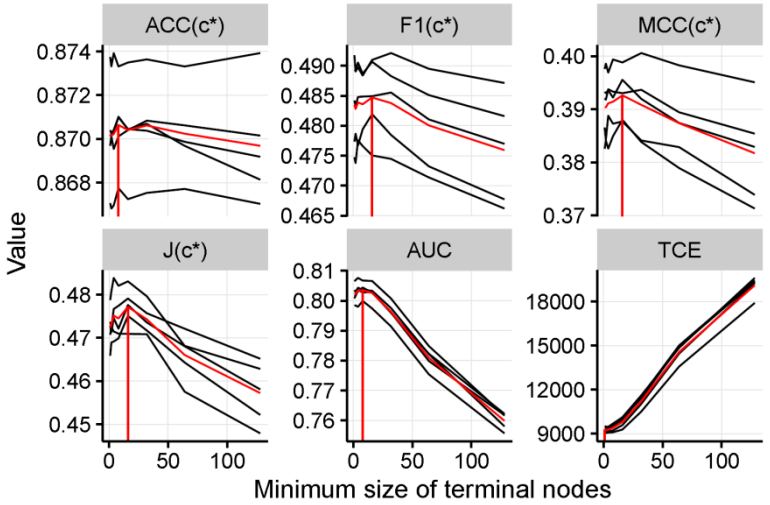
Province
- Number of job vacancies
- Number of dwellings for sale
- Price index privately-owned dwellings
- YoY change in price index

Optimization

Ridge



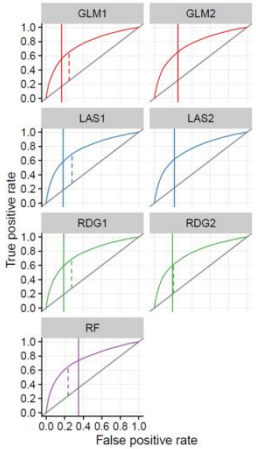
Random forest



Quality measures

Probability [0,1]

- $TCE = \sum_{i=1}^{N} CE_i$
- $CE_i = -\sum_{k=1}^{2} p_{ik} \log \hat{p}_{ik} = -\left(y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)\right)$

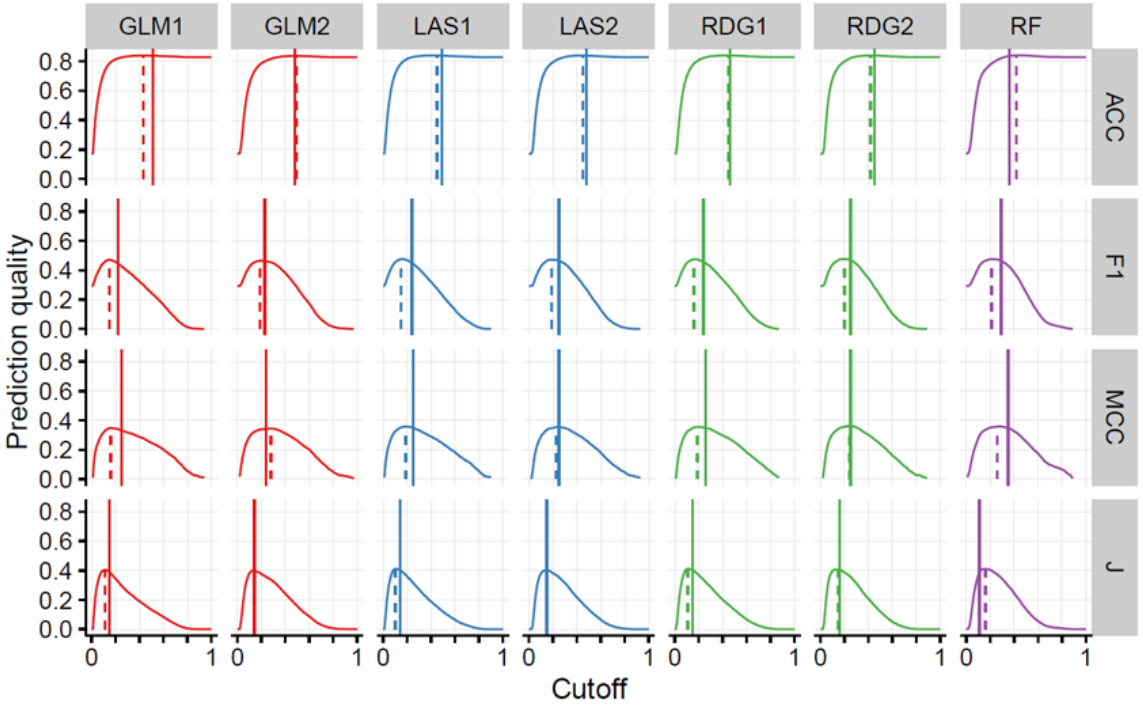


		Predicted		
		Moves	Stays	Total
Actual	Moves	TP	FN	AP
	Stays	FP	TN	AN
	Total	PP	PN	N

Binary {moves,stays}

- $ACC = \frac{TP+TN}{N}$
- $TPR = \frac{TP}{AP}$ (sensitivity, recall)
- $FPR = \frac{FP}{AN}$ (complement of specificity)
- $J = TPR - FPR$
- $PPV = \frac{TP}{PP}$ (precision)
- $F1 = \left(\frac{TPR^{-1}+PPV^{-1}}{2}\right)^{-1}$
- $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{PP \times PN \times AP \times AN}}$

Prediction quality



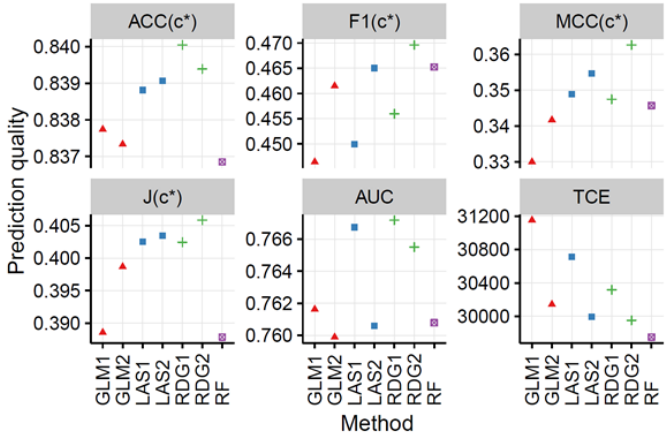
c*

Training (2013)

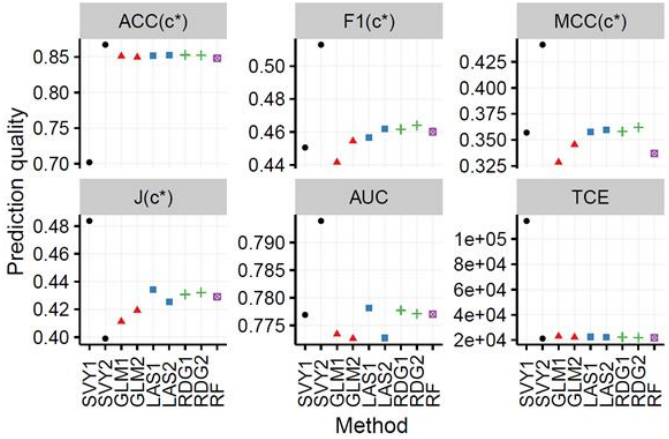
Generalization (2015)

Prediction quality (2)

Random sample

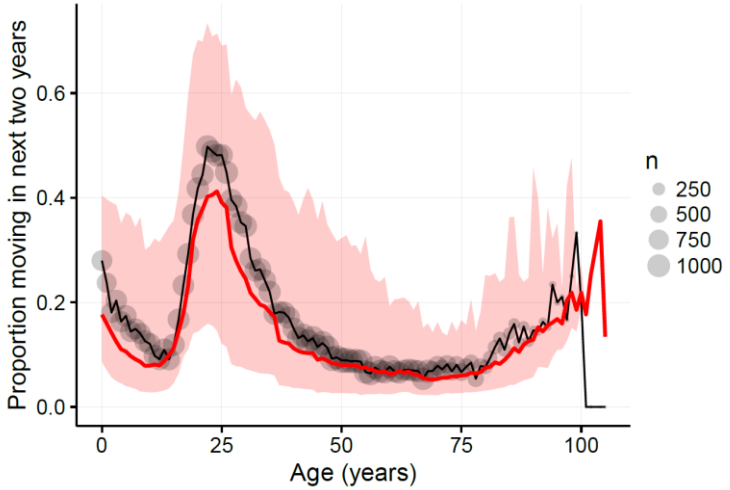


HS2015 respondents



Ridge regression with interactions (RDG2)

Measure	c*	TPR (%)	FPR (%)	PPV (%)
J	0.16	60	19	39
MCC	0.25	46	10	48



Missing data

- Surveys: nonresponse
- Registers: undercoverage
- Images: literally invisible statistics