British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Gateway to the Earth

An application of machine learning to geomagnetic index prediction: Aiding human space weather forecasting

Laurence Billingham, Gemma Kelly

British Geological Survey Geomagnetism Team

A40: #IUGG-4329

powered by scikit learn

© NERC All rights reserved

BGS forecasts

BGS Global Geomagnetic Activity Forecast for Met Office

Forecast period (noon-to-noon GMT)	Forecast Global Activity level	
	Average	Max
23 JUN-24 JUN	STORM G2	STORM G3
24 JUN-25 JUN	ACTIVE	STORM G3
25 JUN-26 JUN	STORM G1	STORM G3

For more information about the forecast and activity categories see
www.geomag.bgs.ac.uk/education/activitylevels.html

Activity during last 24 hours

	Global			Local (UK)		
Date	Average	Max	At time (UT)	Average	Max	At time (UT)
22 JUN-23 JUN	STORM G3	STORM G4	18:00-21:00	STORM G3	STORM G4	18:00-21:00

Additional Comments

An hour-long period of strong southward IMF around 19:00 UT on 22 JUN produced a peak in

- Daily geomagnetic forecast
  - next 3 days
  - human forecasters based on intuition from experience
- online: http://tinyurl.com/BGSSwForc

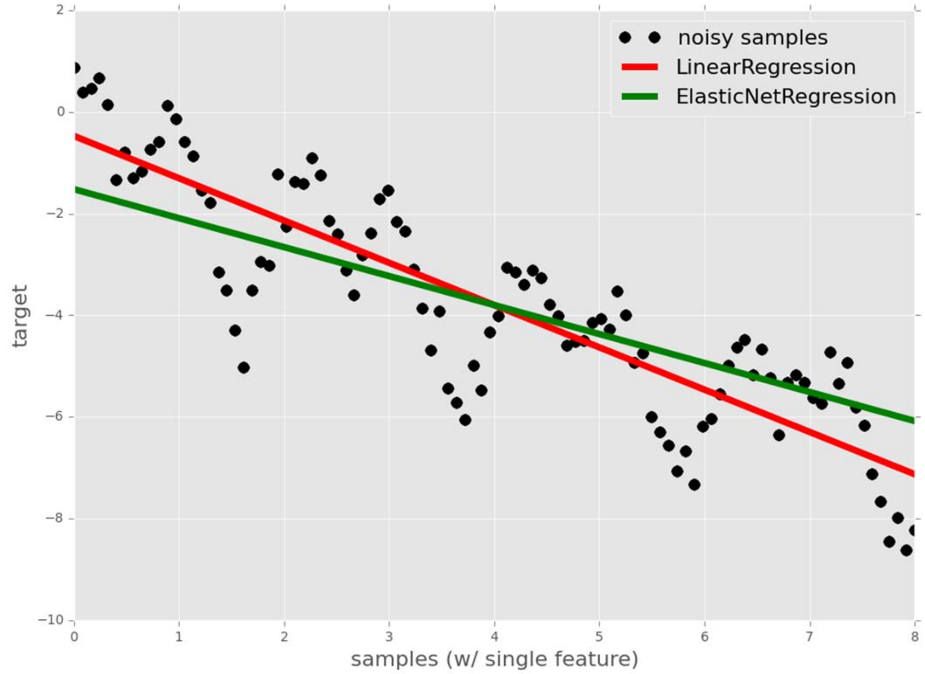
- humans need to do other tasks (and sleep, eat, …)
- have algorithmic tools to help human forecasters and alert them
- we show scoping study for new tool to help human forecasters
  - automatic prediction of 3 hourly a_p

© NERC All rights reserved

Algorithms: some intuition

- 2 broad classes
  - regression $\in \mathbb{R}$
  - classification: storm, quiet

- Generalized linear regression
  - global model
  - least squares unstable to noise and outliers, non-unique
  - introduce regularization
  - penalize coefficients
  - prefer lower gradients
  - set some coefficients to 0

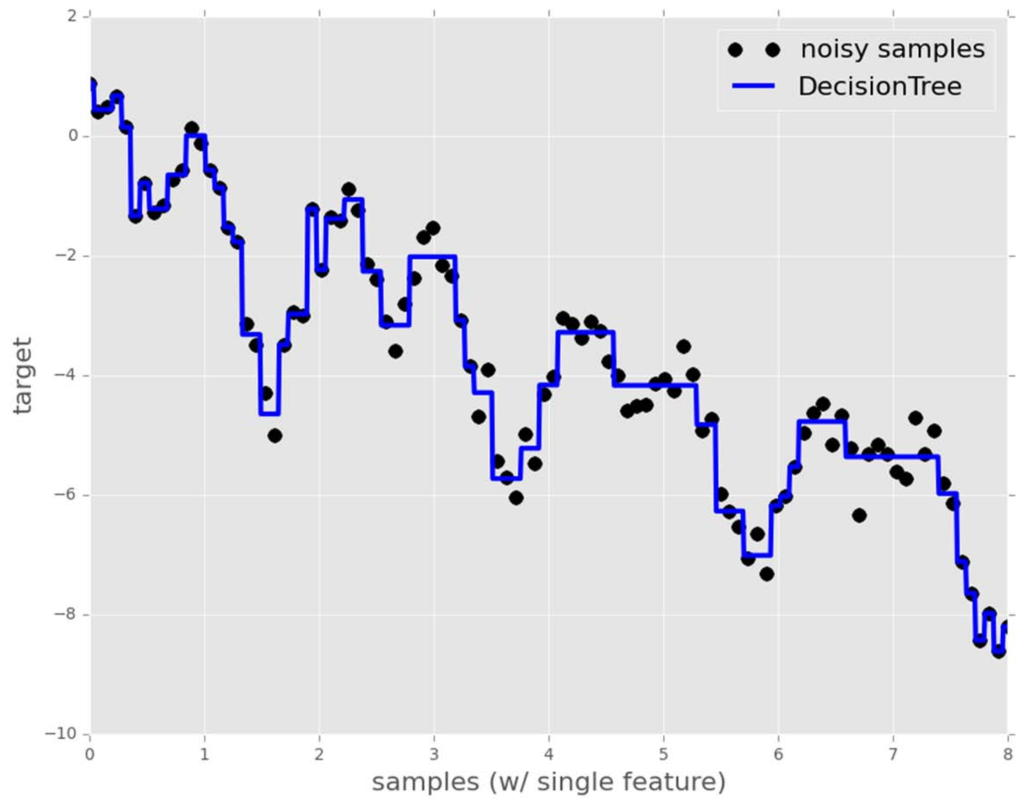


$$\min_{\mathbf{w}} \left(\boxed{||\mathbf{X}\mathbf{w} - \vec{y}||_2^2} + \boxed{\alpha\rho||\mathbf{w}||_1} + \boxed{\frac{\alpha(1-\rho)}{2}\,||\mathbf{w}||_2^2} \right)$$

© NERC All rights reserved

Algorithms: some intuition

- Tree methods
  - fit piecewise constant model
  - split on information criterion
  - local
  - non linear
  - easy to overfit: fit the noise

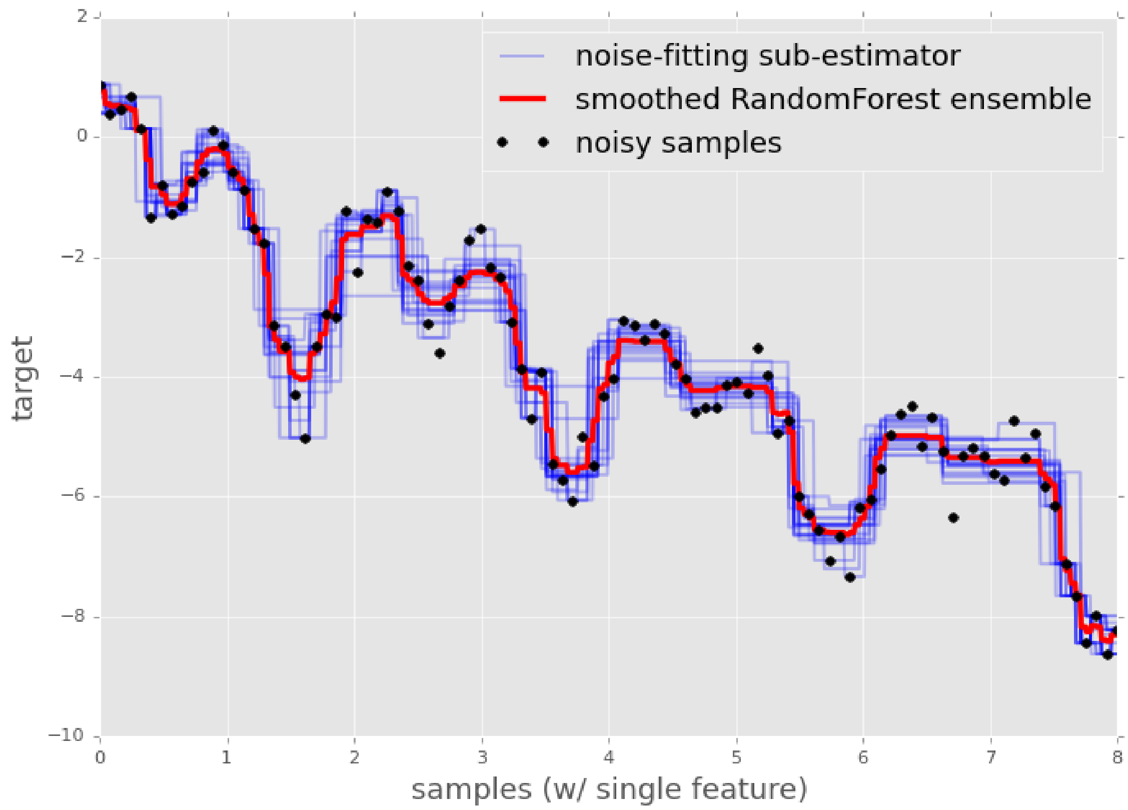


© NERC All rights reserved

Algorithms: some intuition

- Tree methods
  - easy to overfit: fit the noise

- Ensemble methods
  - fit ‘forest’ of overfitting trees
  - smooth out overfit
  - more robust to noise

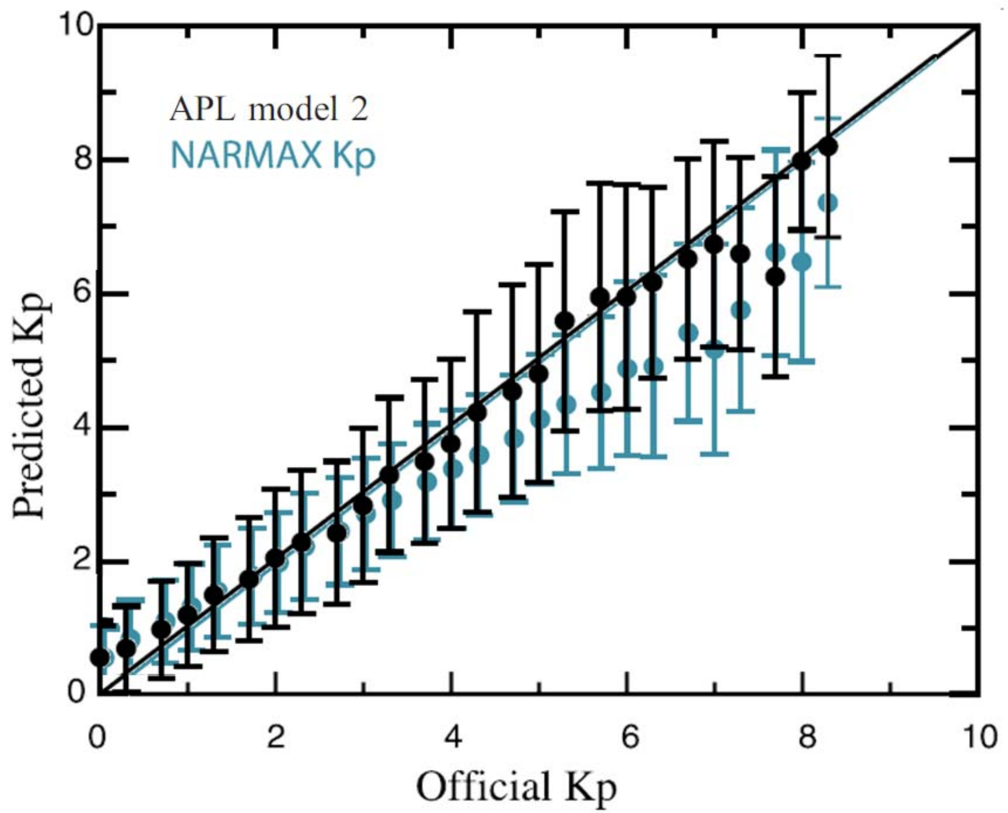


© NERC All rights reserved

Dataset

‘**targets**’ to predict

- Predict values for
  1. next 3-hour interval
  2. next-but-one interval
  3. 24 hour running mean
  4. 24 hour running maximum
  - cf. Wing, Bala Reiff: 1-6 hours ahead
- some techniques can model all 4 in once pass



after Wing 2005

© NERC All rights reserved

Dataset

- Pick set of 'targets' to predict
- all different 'views' on the same thing: activity at an 'average' sub-auroral observatory
  - ap
    - 'feels like' $\in \mathbb{R}$
  - Kp
    - can be made to 'feel like' $\in \mathbb{R}$:

      2, 2.333, 2.6667
    - really a category:

- NOAA G scale
  - categorical
  - hide unwanted detail during quiet times
  - humans forecasters and customers use

- choice affects
  - definition of success
  - ease of comparison vs previous studies
  - which algorithms work
    - e.g. differentiable loss for regression $\in \mathbb{R}^n$

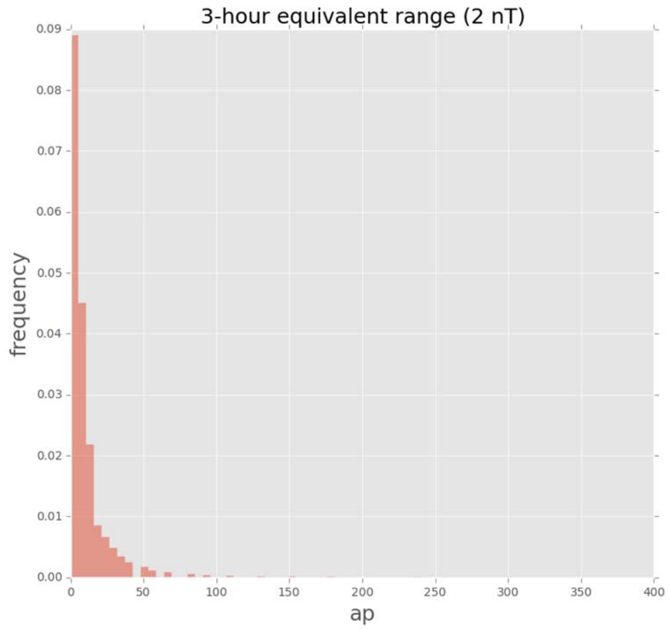
ap	Kp	NOAA Category
	< 3+	G0
18	3+	
22	4-	
27	4o	
32	4+	
39	5-	G1
48	5o	
56	5+	
67	6-	G2
80	6o	
94	6+	
111	7-	G3
132	7o	
154	7+	
179	8-	G4
207	8o	
236	8+	
300	9-	
400	9o	G5

© NERC All rights reserved

Dataset

- **select** samples
  - storms **rare** but important
  - **balance** dataset otherwise storms look like noise
  - storm rarity limits dataset size

- **split**
  - training set (**w**)
  - parameter-fitting set (α,ρ)
  - test set
- each balanced storm/quiet

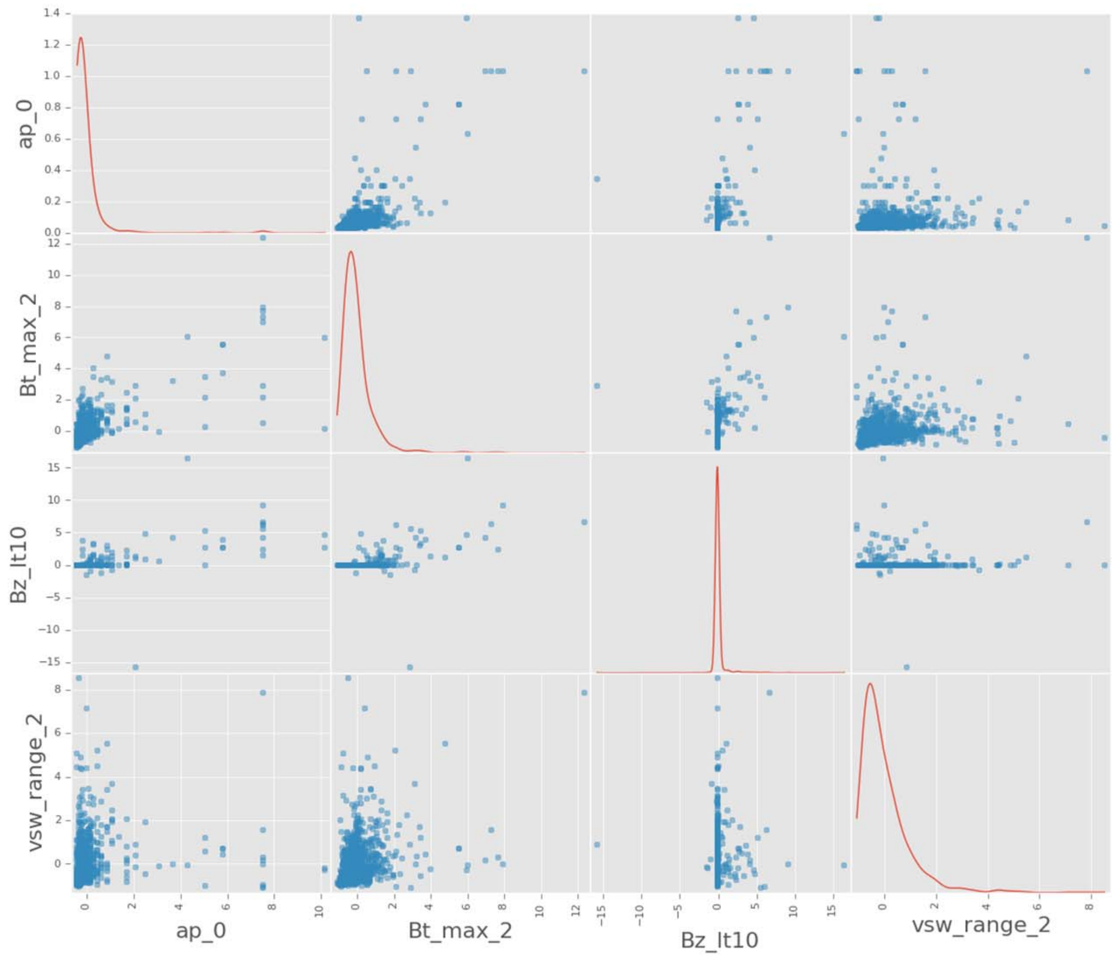


- **scale** the training set
  - $x_{sample} \rightarrow (x - \bar{x})/s$
- same scaling to test, parameter-fitting sets
- scaling before split would 'leak' information

© NERC All rights reserved

Dataset

ACE and ground magnetometer data from 1998 to 2015



features:

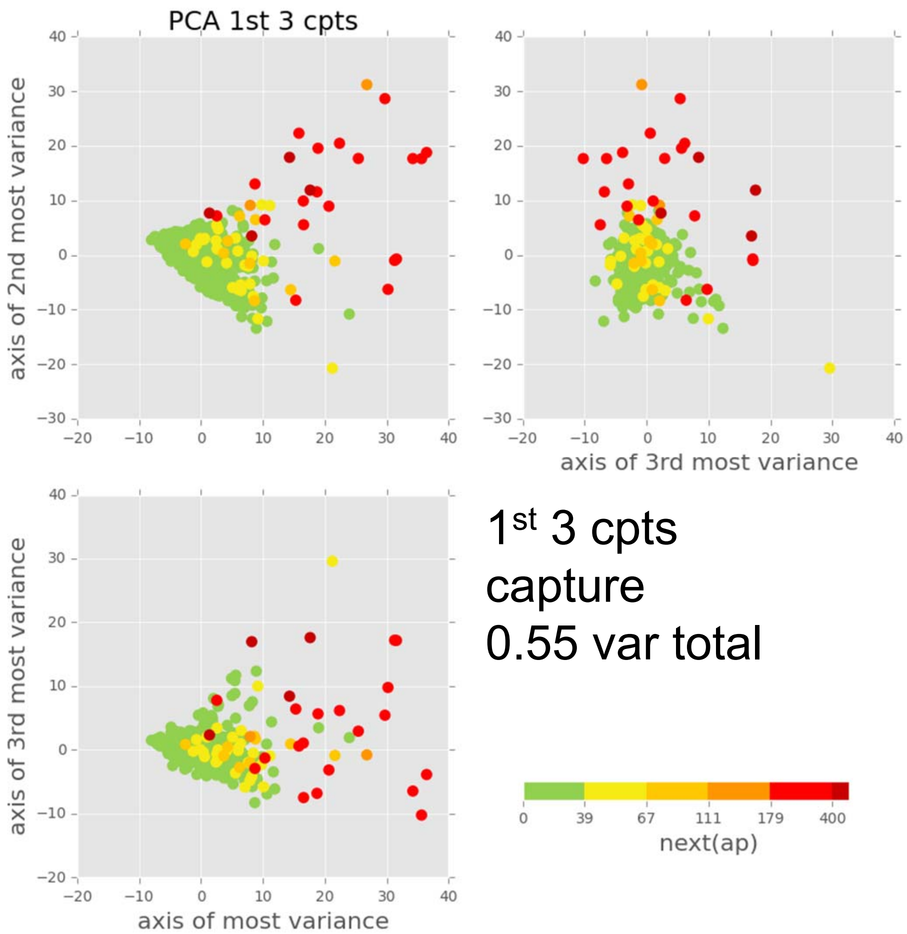
- $max(v_{SW})|_{last\ 5\ hours}$
- $max(ap)|_{27\ days\ \ ago}$
- $range(|B_{IMF}|)|_{4\ to\ 5\ hrs\ ago}$
- $time(B_{z,\ IMF} < 10\mathrm{nT})|_{last\ 24\ hrs}$
- $\ldots$
  - ~100 features in all

© NERC All rights reserved

Principal component analysis and friends

- have unwanted $x_i \propto x_j$
  - shocks: $\Delta v_{SW} \propto \Delta|B_{IMF}|$
  - MHD flux freezing: $n_{proton} \propto |B_{IMF}|$
  - timeseries: autocorrelation

- keep 0.95 total variance
  - ~100 components -> ~30
  - axes are linear combinations ~100 coeffs but can be made sparse



1st 3 cpts capture 0.55 var total

© NERC All rights reserved

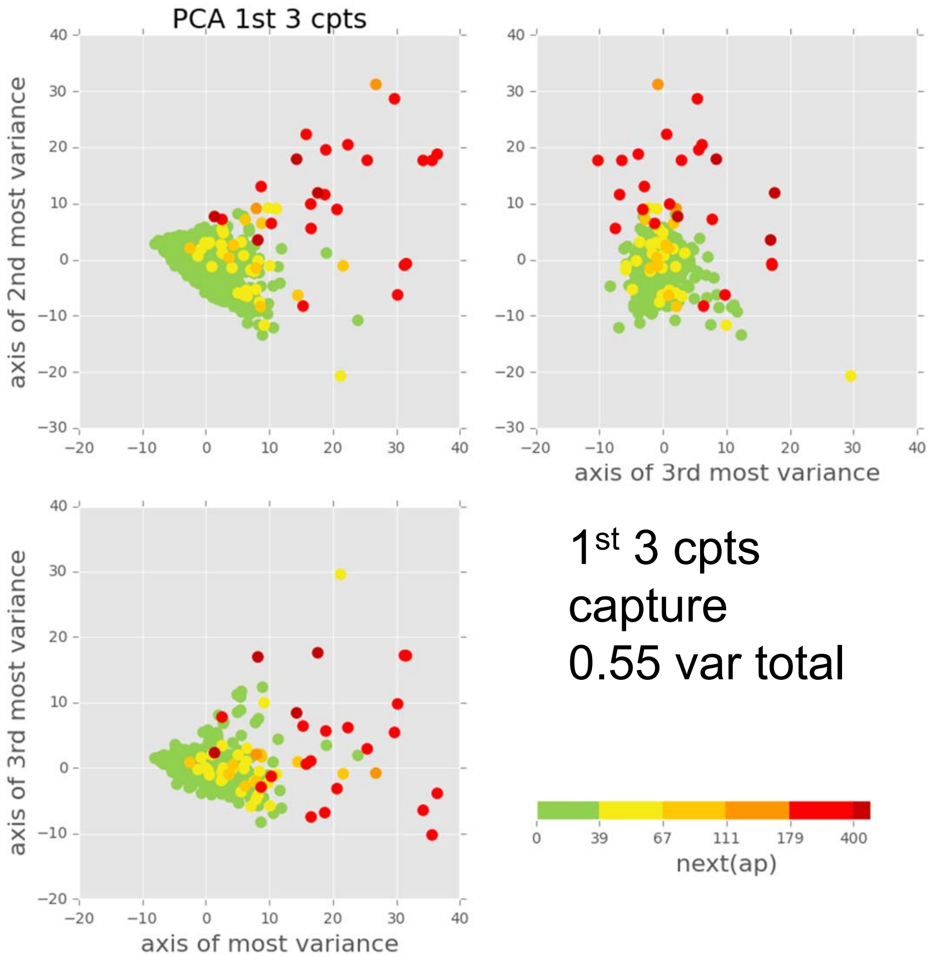
Principal component analysis and friends

- some separation between storm and quiet times

- most variance along

  1. $ap|_{last}$ / $|B_{IMF}|_{last\ few\ hrs}$
  2. $v_{SW}|_{last\ few\ hrs}$
  3. $ap|_{7\ to\ 10\ hrs\ ago}$
  4. $ap|_{\sim 28\ days\ ago}$

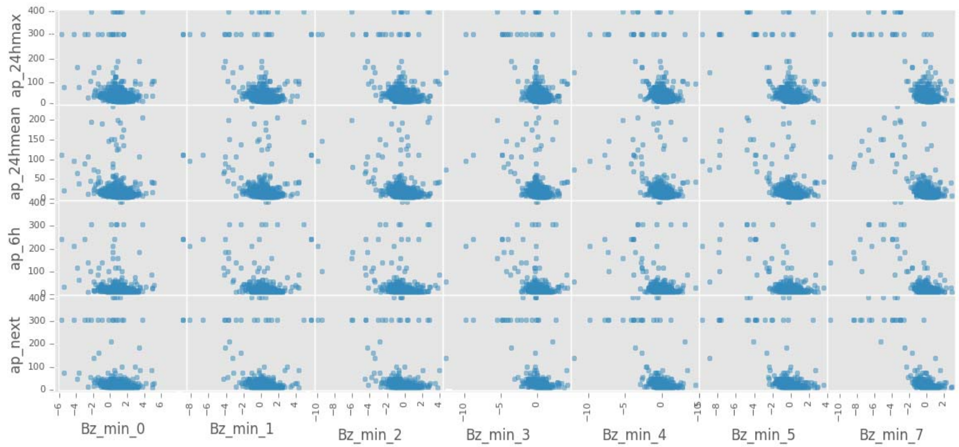
- surprising **lack of variance** along $B_{z,IMF}$ directions



© NERC All rights reserved

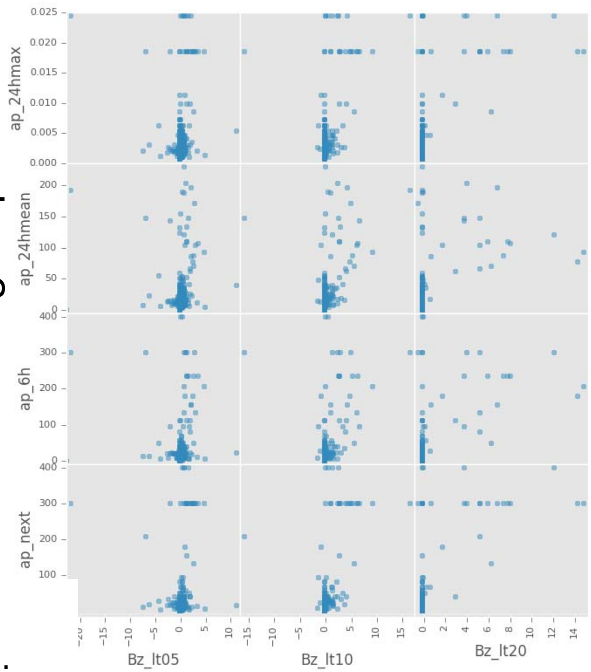
Lack of Bz variance

target ap



min(Bz)| few hours ago

target ap

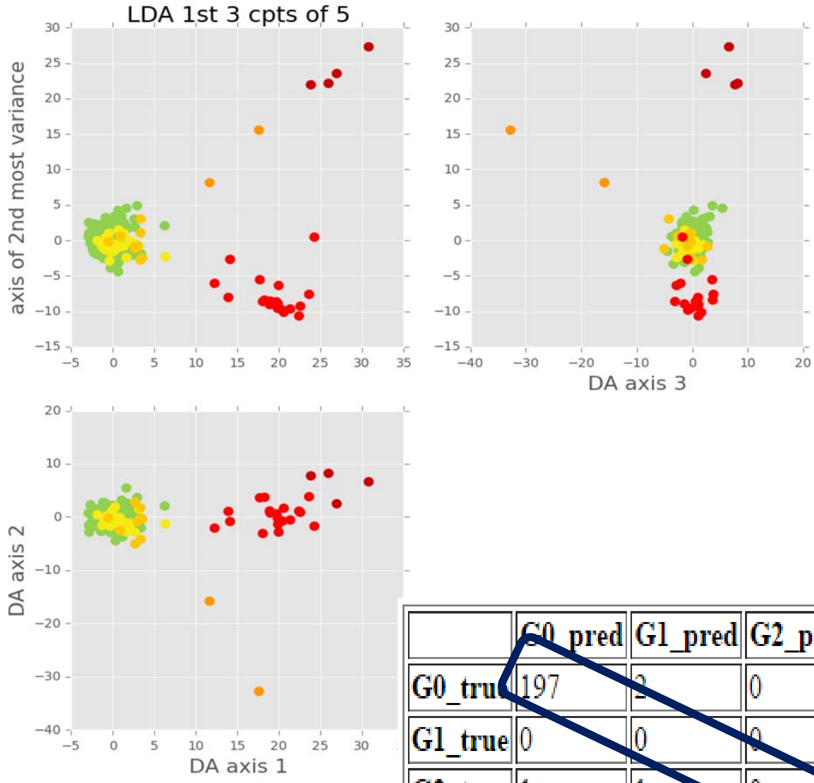


time Bz below threshold

- neither values of Bz nor threshold times correlate well with ap
  - lack of variance $\not\Rightarrow$ no causality but algorithms have less of a 'lever' in Bz
  - Kp and derivatives are perhaps not best parameters for space weather [see Kelly et al. talk in A18 on 2015-06-27 09:45]
  - non-linear dimensionality reduction (kernel PCA, Isomap) results similar

Classification > regression

- other decompositions that separate G levels
- different algorithms
- stratified train, parameter fit splitting
- easier cross validation



	G0_pred	G1_pred	G2_pred	G3_pred	G4_pred	G5_pre
G0_true	197	2	0	0	0	0
G1_true	0	0	0	0	0	0
G2_true	1	1	0	0	0	0
G3_true	0	1	0	0	1	0
G4_true	0	0	0	0	2	0
G5_true	0	0	0	0	0	1

best so far

RandomForestClassifier

G next score 0.93 pm 0.06

G max 24 h 0.85 pm 0.14

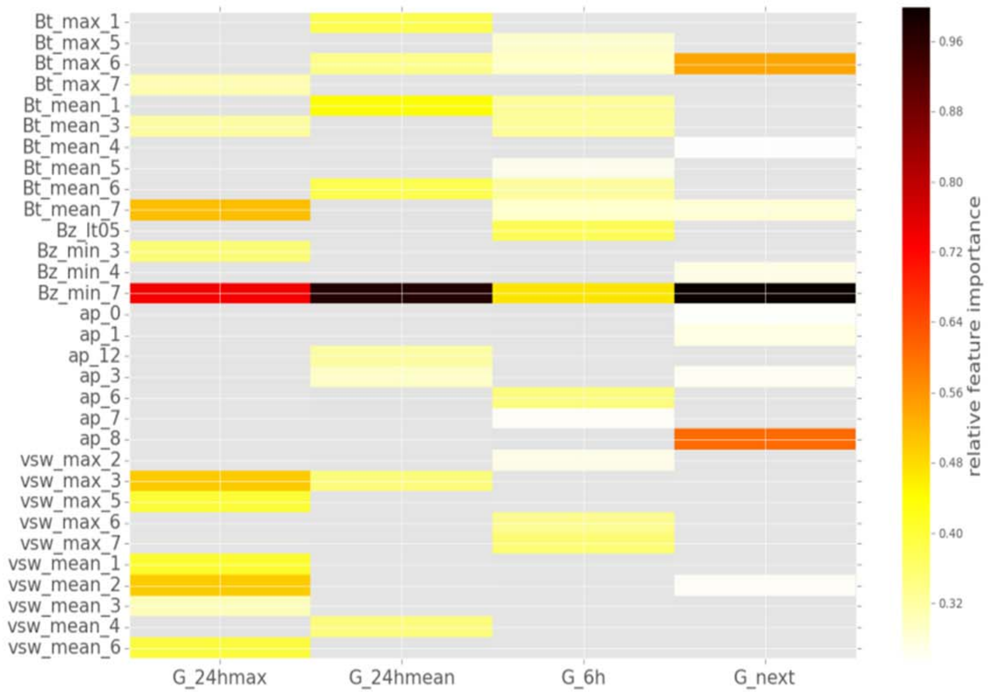
© NERC All rights reserved

Most informative features

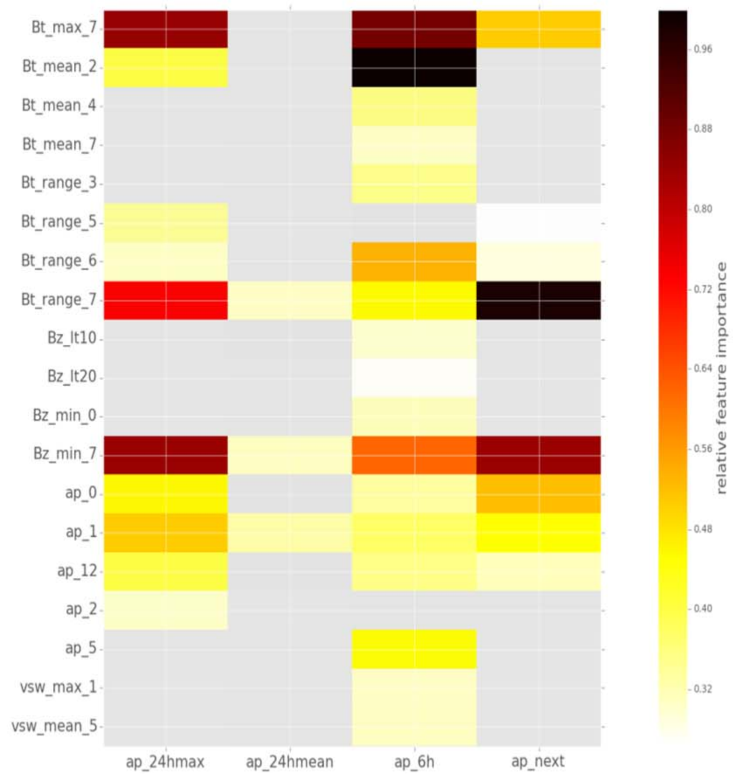
feature importances > 0.25 * most important ∀ targets

- hard to do this with neural nets

RandomForest Classifier



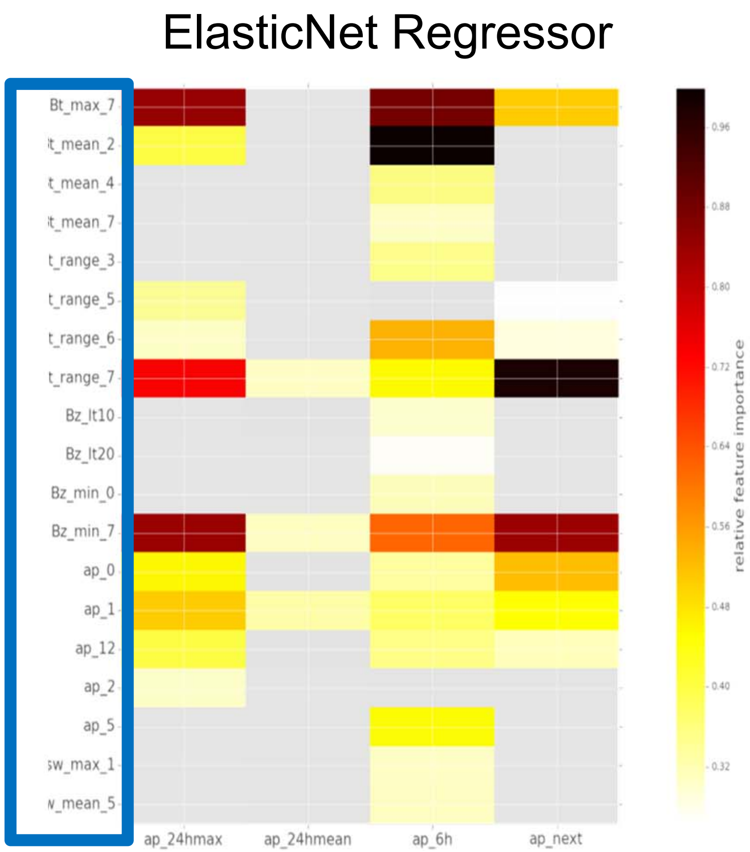
ElasticNet Regressor



Most informative features

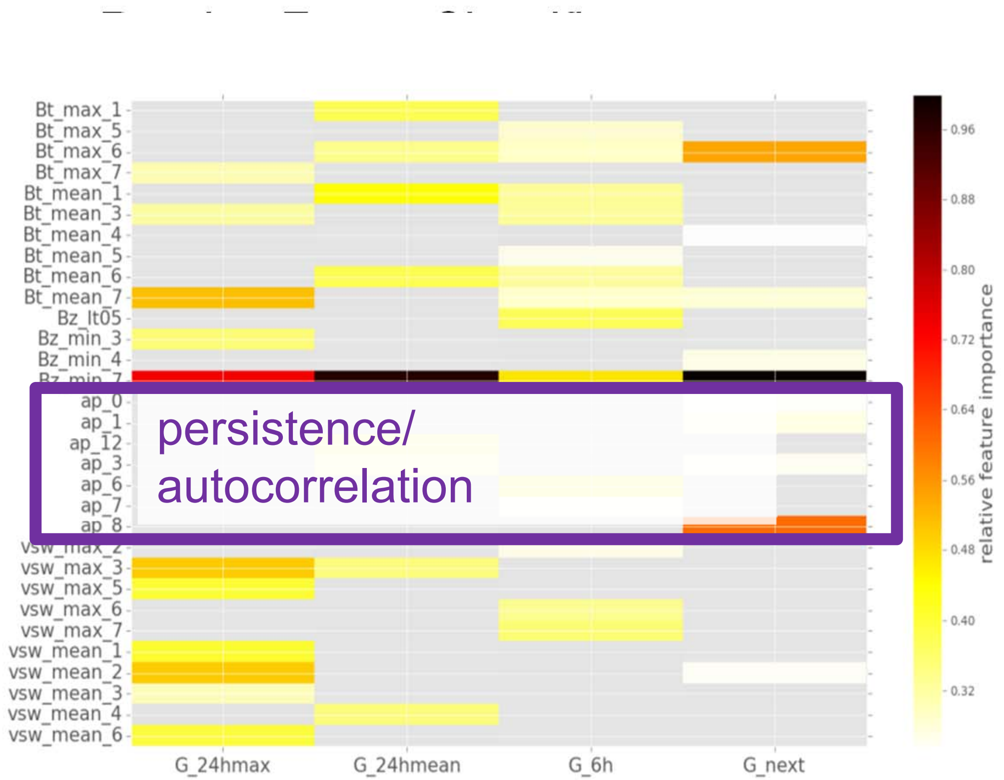
feature importances > 0.25 * most important ∀ targets

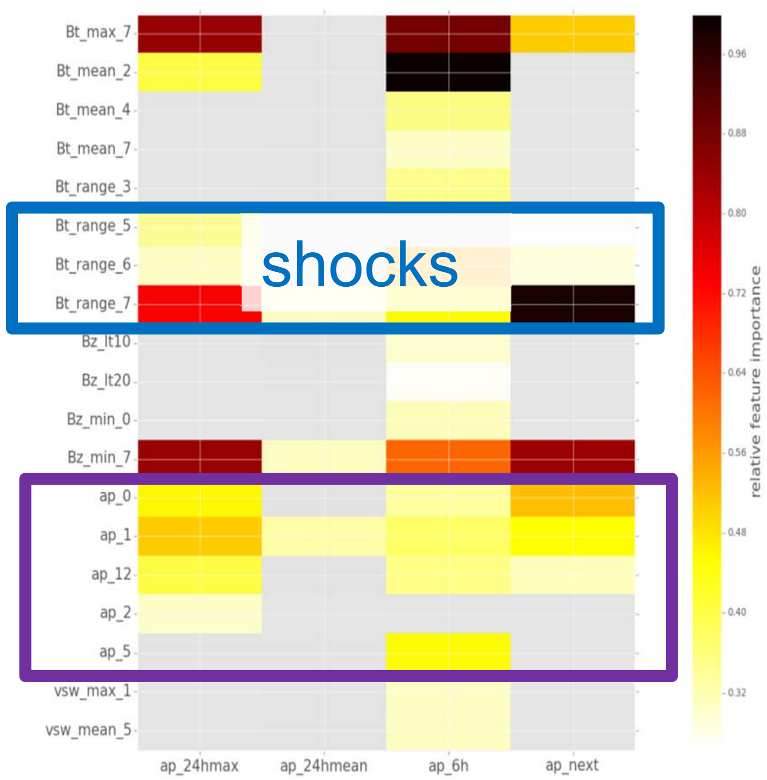
non-linear, local models more diverse



Most informative features

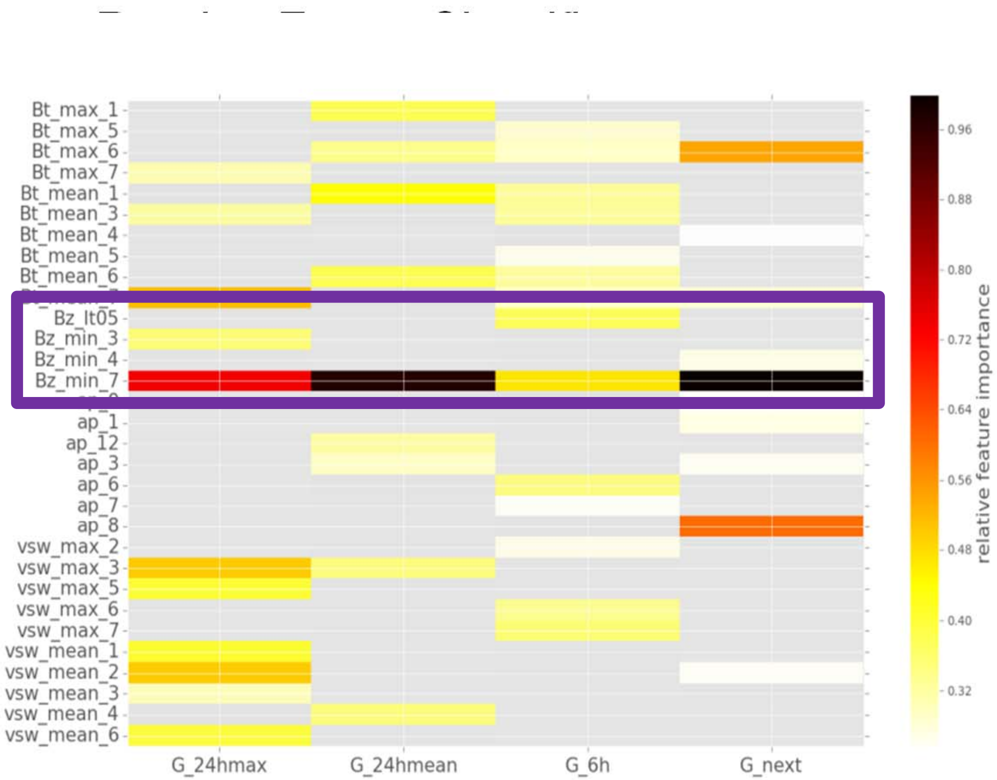
feature importances > 0.25 * most important ∀ targets

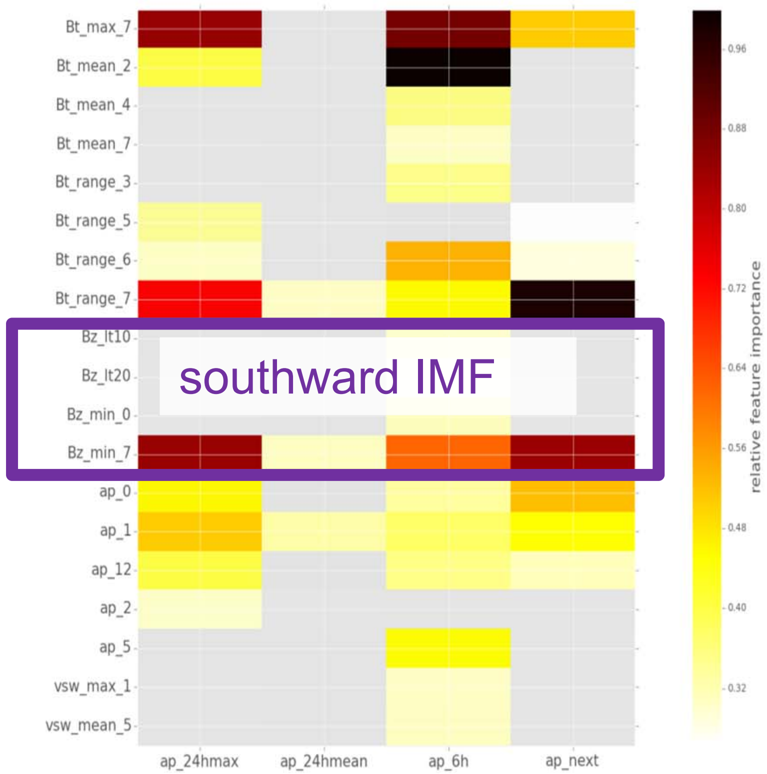




Most informative features

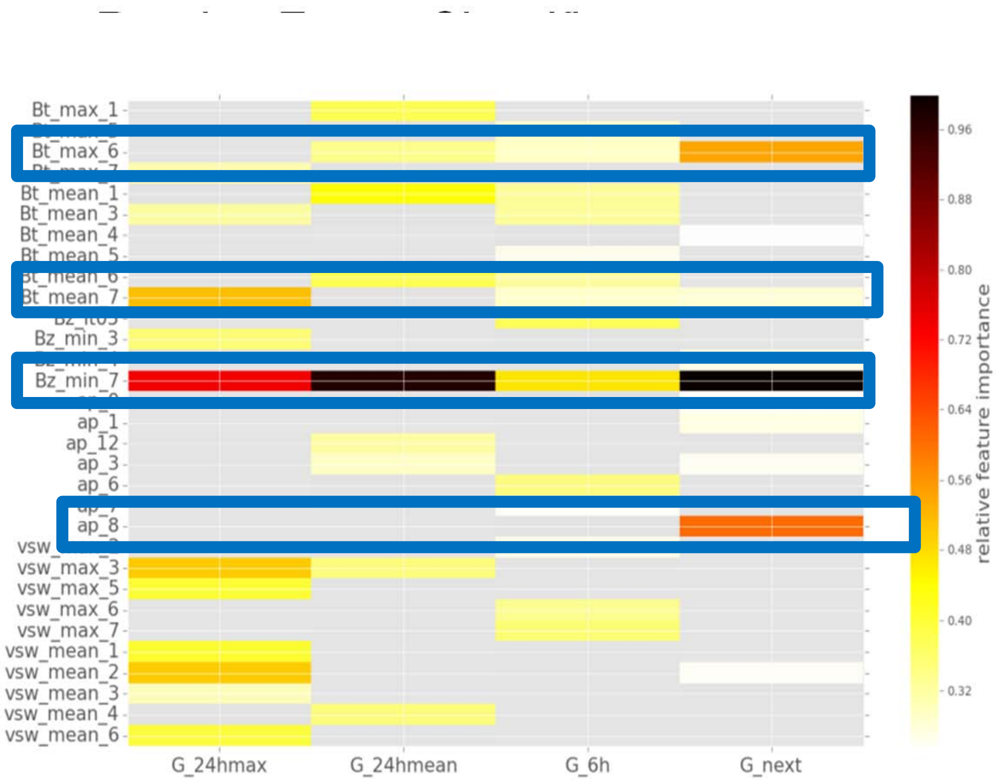
feature importances > 0.25 * most important ∀ targets

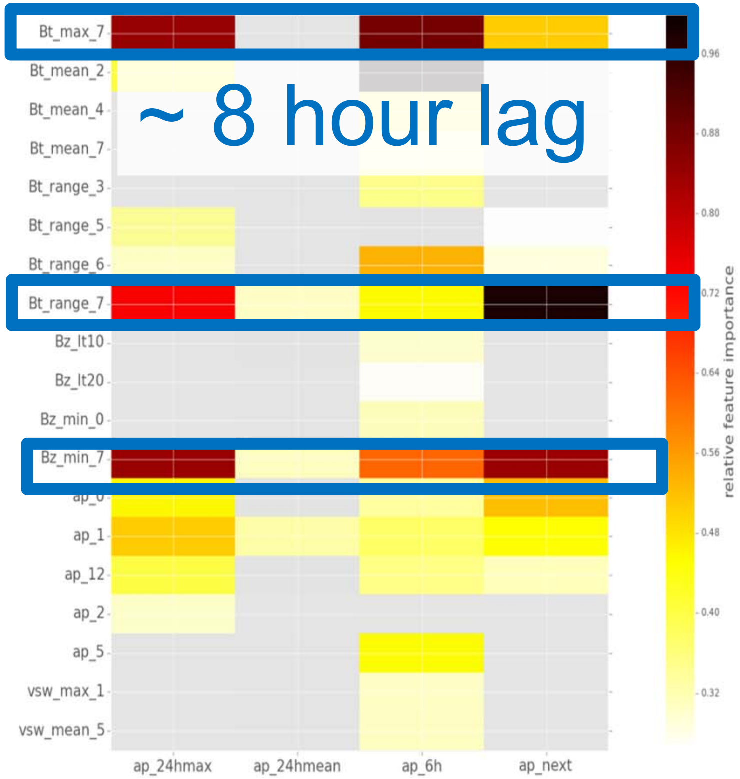




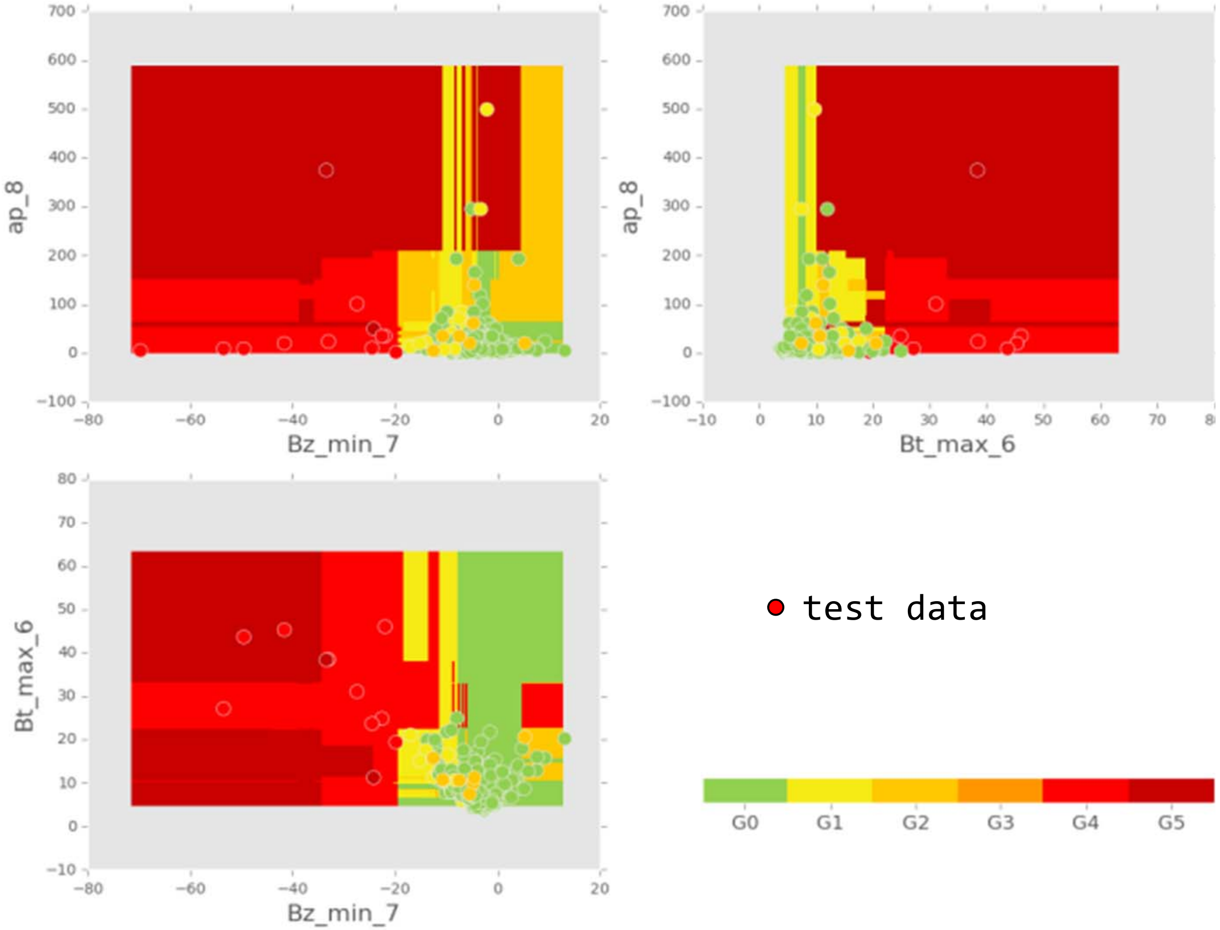
Most informative features

feature importances > 0.25 * most important ∀ targets





2D models

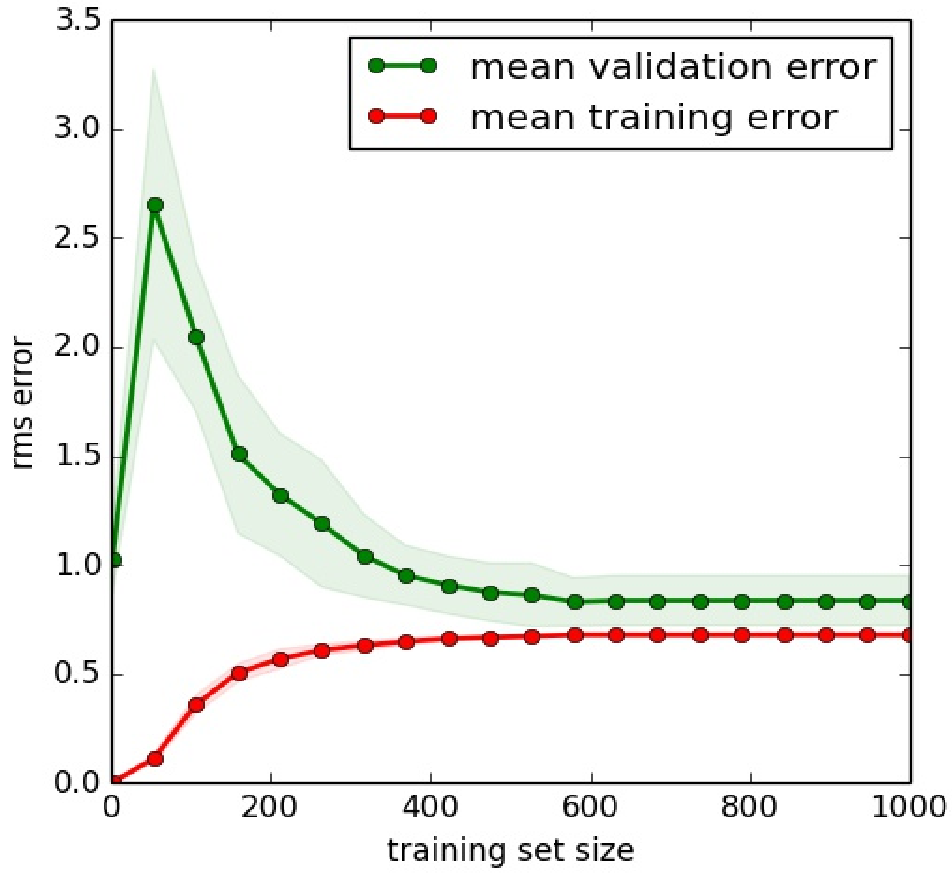


- pair most important features
- train 2D model with same hyperparameters as full model

- only for plotting
- still get most storms

Conclusions

- r^2 score $\max(G)|_{24\ \mathrm{hrs}} = 0.85$
- targets not $\in \mathbb{R}$?
  - classifiers advantageous
- ensemble models with modern hardware
- find most important features
  - less black box than neural nets



Future

- converge quickly: more features
- solar data: radio bursts
- solar wind coupling functions: could compare performance of many

References

Bala and Reiff, 2012, SpWe, ‘Improvements in short-term forecasting of geomagnetic activity’

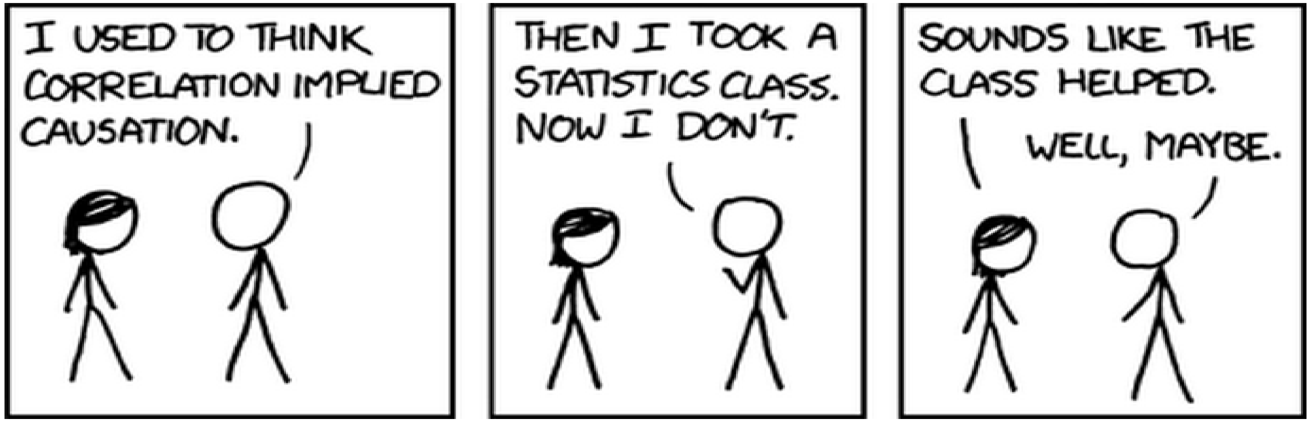
Hastie et al. 2009, Springer,   Elements Statistical Learning II

Pedregosa et al., 2011, JMLR, 'Scikit-learn: Machine Learning in Python‘

Wing *et al.*, 2005, JGR, ‘Kp forecast models’

© NERC All rights reserved

Correlation

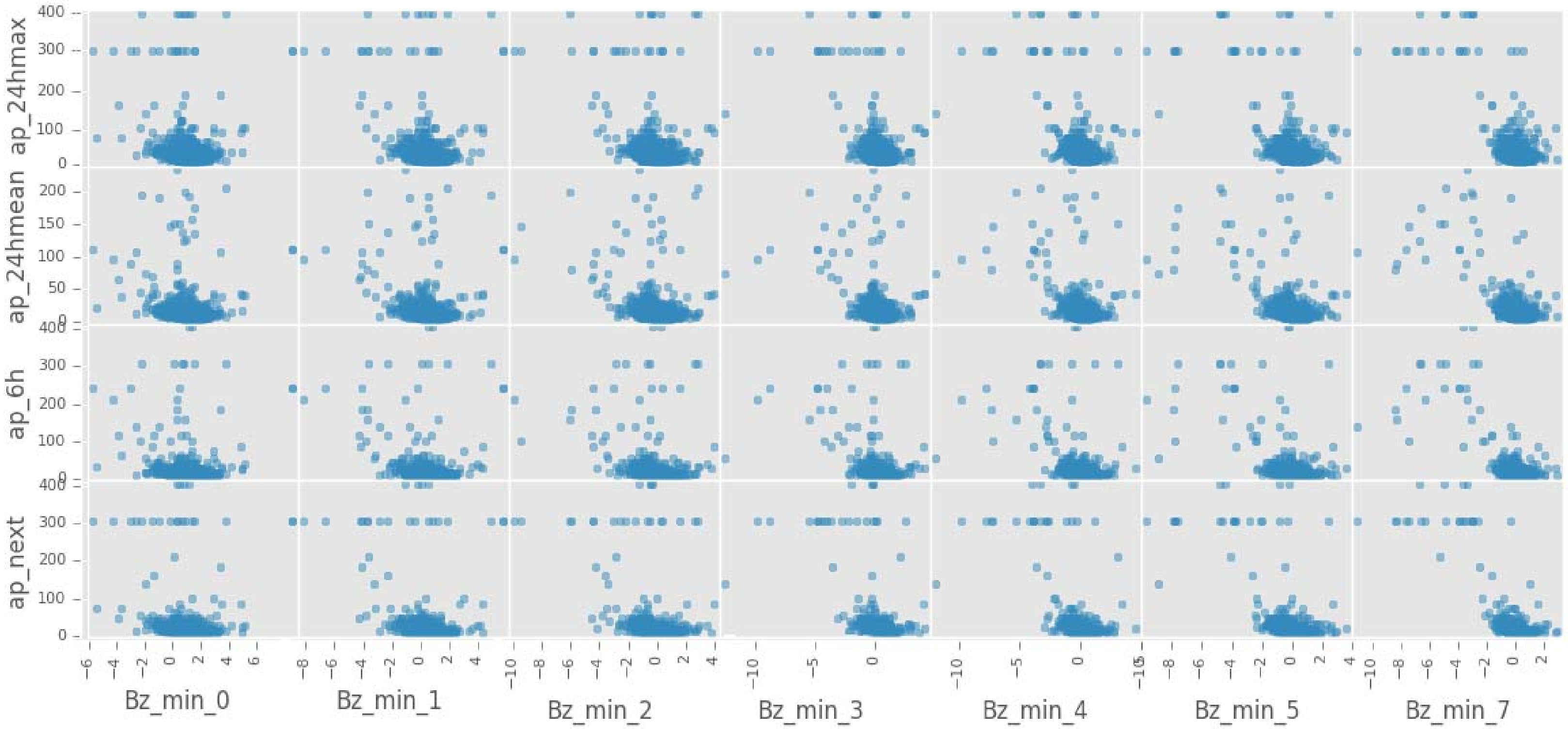


Title text: Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.

lack of correlation $\not\Rightarrow$ lack of causation

© NERC All rights reserved

A small slice of our feature-space



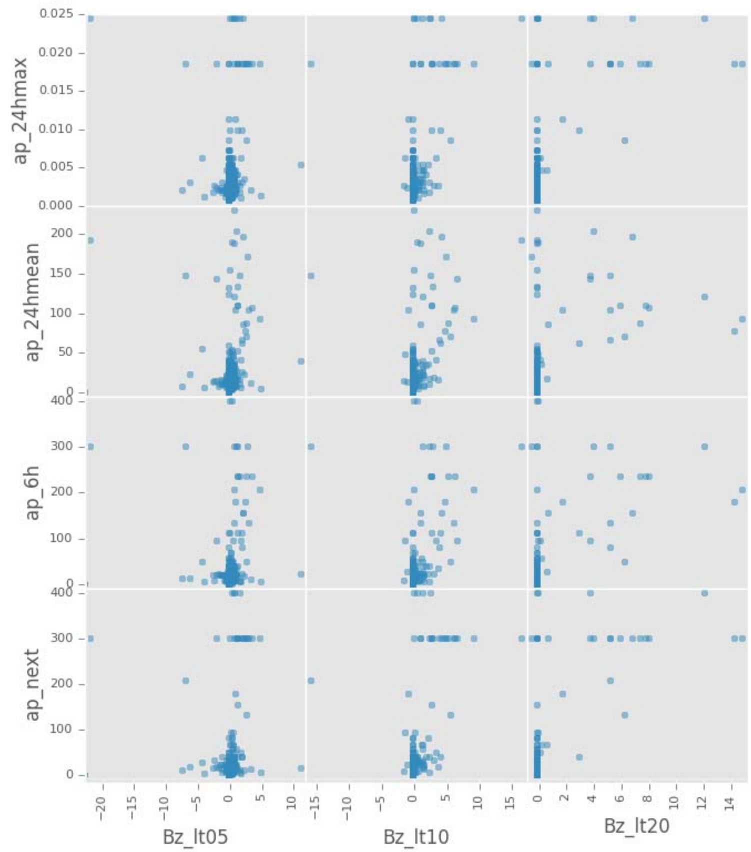
target ap

min(Bz)| few hours ago

© NERC All rights reserved

A small slice of our feature-space

target ap



time Bz below threshold

© NERC All rights reserved

Abstract, do not show

Geomagnetic indices are ubiquitous parameterizations of storm-time magnetic conditions. Their prediction is one goal of space weather forecasting and they are required inputs for a variety of models. Despite much recent progress; human space weather forecasters, unlike terrestrial weather forecasters, cannot yet rely on physical models to make whole-system predictions. We trial various data-driven models: seeking robust; accurate; and fast predictions of the Ap index, and derived values, as an operational forecaster aid.

Machine learning (ML) is a branch of statistics focused on making accurate predictions when presented with novel data. ML techniques underpin much of our online lives: from web-search and recommendation systems to fraud detection. Modern computer hardware and ML libraries allow models to be: regularly re-trained with the latest data, and optimized over ever larger parameter spaces.

Index prediction presents a number of challenges for statistical models. The most important events to predict are, potentially infrastructure damaging, large storms. However, they are rare events and distributions of geomagnetic activity are positively skewed with very heavy tails. We present strategies for dealing with the rarity of large storms; both predicting them accurately, and being able to quantify a model's large-storm predictive power. We also demonstrate schemes for data cleaning and assimilation including the integration of disparate data types within a single model.

A variety of algorithms are trained including models in both local and global parameter space, we inter-compare them and benchmark them against an existing operational auto-regressive model. We use various metrics, many of which show the ML methods predict storms better than the existing approach.

© NERC All rights reserved

Metrics

- many studies use R^2 between true and predicted as success criterion
- may not be appropriate given fat-tail of ap (and Kp)

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_y = \sum_i (y_i - \bar{y})^2$$

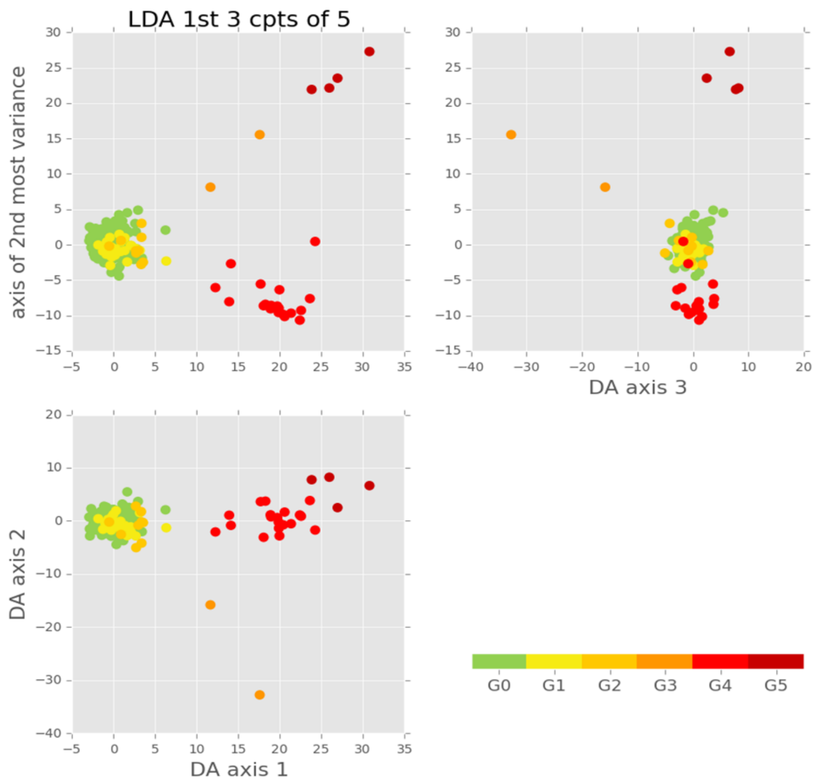
$$S_{pred} = \sum (y_i - y_{pred\ i})^2$$

$$R^2 = 1 - \frac{S_{pred}}{S_y}$$

© NERC All rights reserved

Forget regression: classify

- other decompositions that separate G levels
- different algorithms
- stratified train, parameter fit splitting
- easier cross validation



best so far

RandomForestClassifier

G next score 0.93 pm 0.06

G max 24 h 0.85 pm 0.14

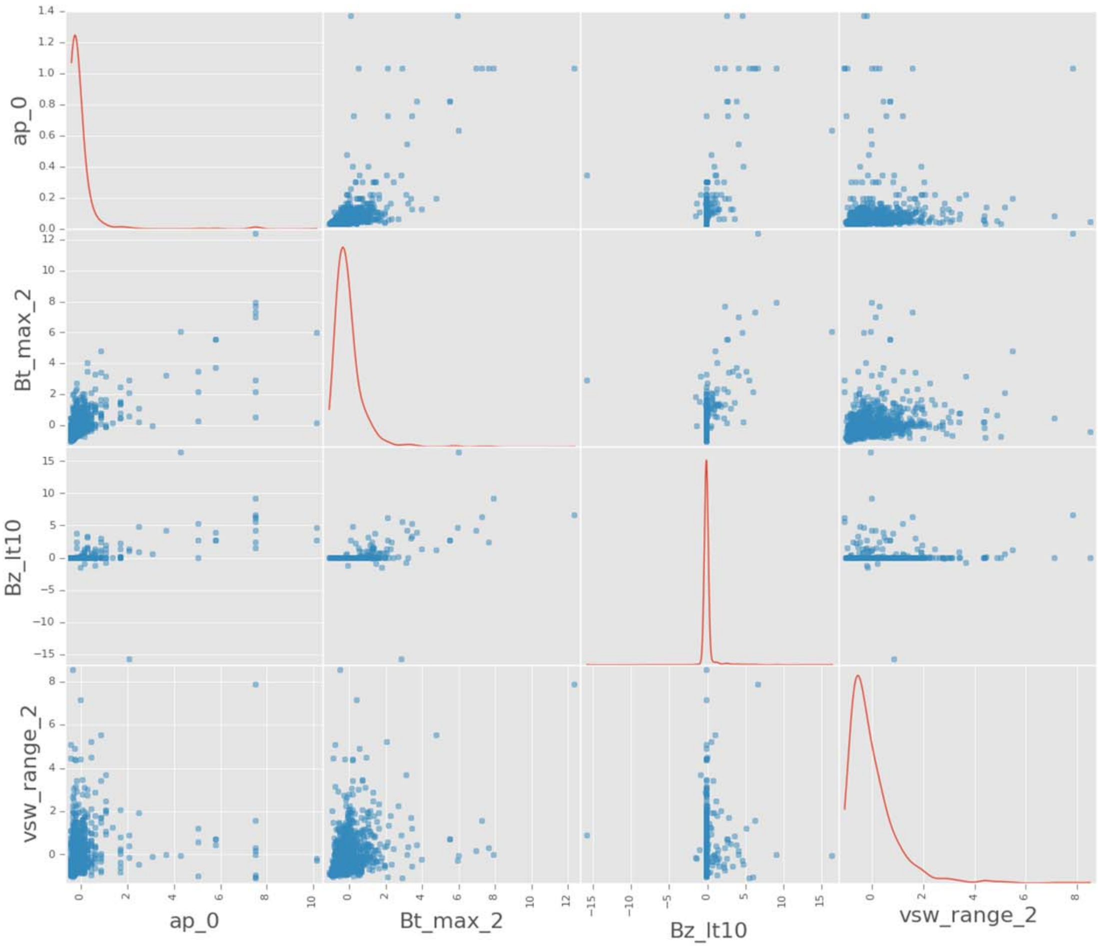
	G0_pred	G1_pred	G2_pred	G3_pred	G4_pred	G5_pred
G0_true	197	2	0	0	0	0
G1_true	0	0	0	0	0	0
G2_true	1	1	0	0	0	0
G3_true	0	1	0	0	1	0
G4_true	0	0	0	0	2	0
G5_true	0	0	0	0	0	1

© NERC All rights reserved

Dataset

- ACE and ground magnetometer data from 1998 to 2015

- Define a set of
  - $max(v_{SW})|_{last\ 5\ hours}$
  - $max(ap)|_{27\ days\ ago}$
  - $range(|B_{IMF}|)|_{4\ to\ 5\ hrs\ ago}$
  - $time(B_{z,\ IMF} < 10\mathrm{nT})|_{last\ 24\ hrs}$
  - $\ldots$
- ~100 features in all

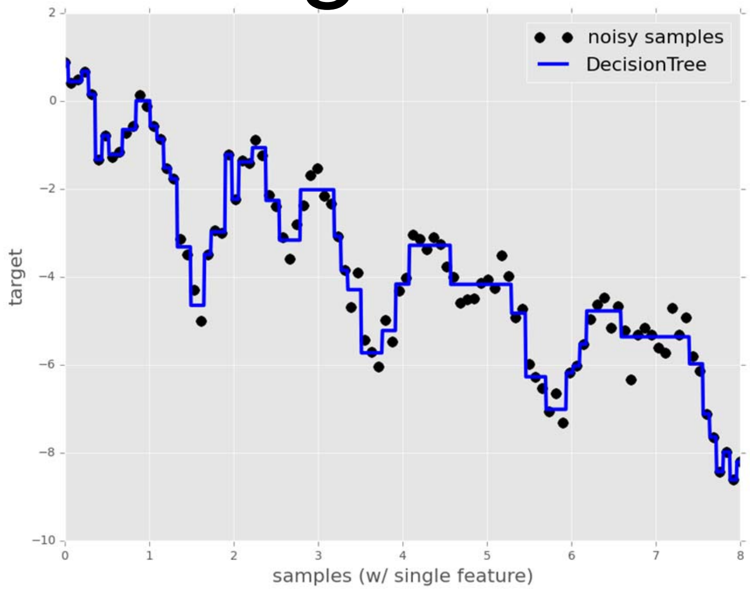


© NERC All rights reserved

An (in the end not very) interesting result

- Decision tree regressor can predict
  1. `ap_24hmax`
  2. `ap_24hmean`
  3. `ap_6h`
  4. `ap_next`

  in 1 pass



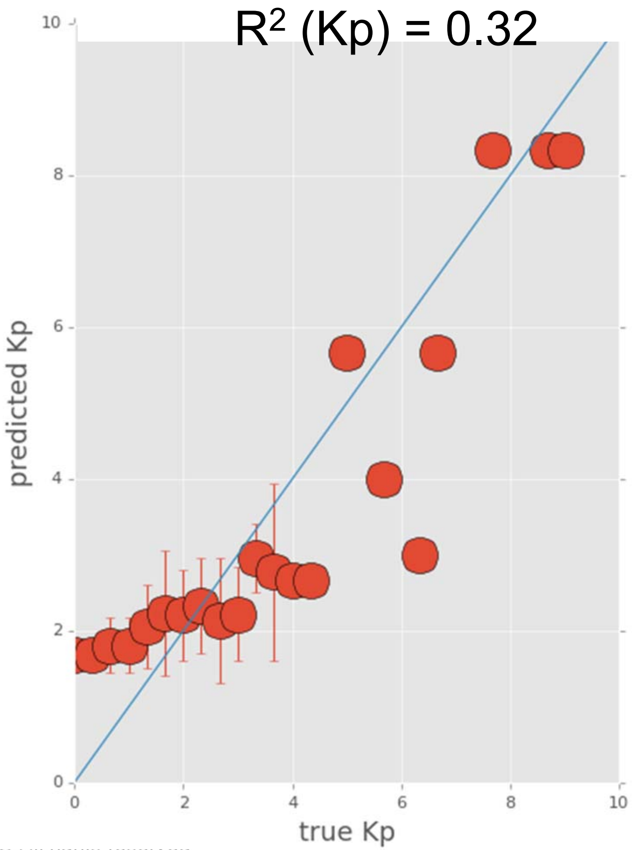
- promotes sparse features
  - keep only 7/100
    i. ap_1,
    ii. ap_0,
    iii. vsw_mean_4,
    iv. Bt_mean_7,
    v. Bt_range_7,
    vi. ap_11

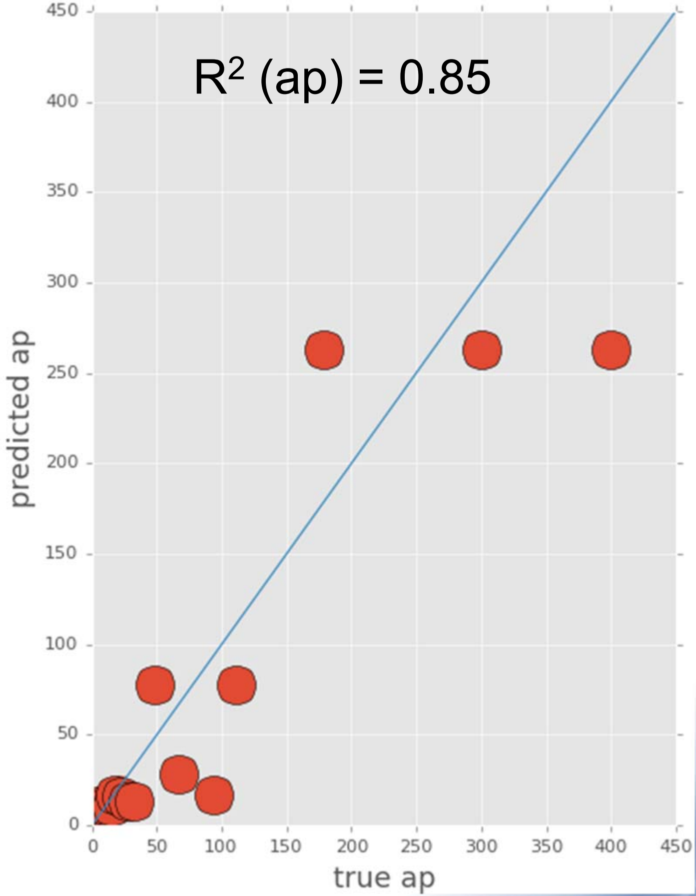
- trained in ~seconds

total R^2 = 0.78 (all targets 1-4)
cf 0.79 Wing APL2 predicting Kp 4 hours ahead

© NERC All rights reserved

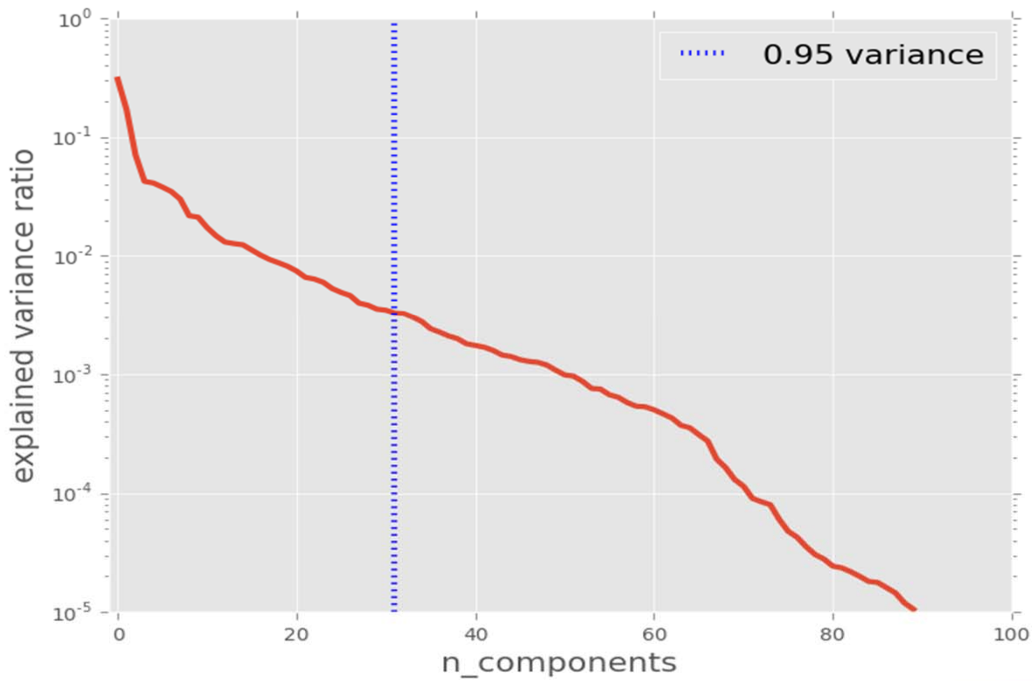
Metrics


R² (Kp) = 0.32
predicted Kp
true Kp
0
2
4
6
8
10


R² (ap) = 0.85
predicted ap
true ap
0
50
100
150
200
250
300
350
400
450

Curse of dimensionality

- Now have a ~ 100 dimensional data set
  - On a good day I might be able to visualize things in 3



- Some approaches require

  $x_i \not\propto x_j$

- But we have
  - shocks:

    $n_{proton} \propto |B_{IMF}|$

  - MHD flux freezing:

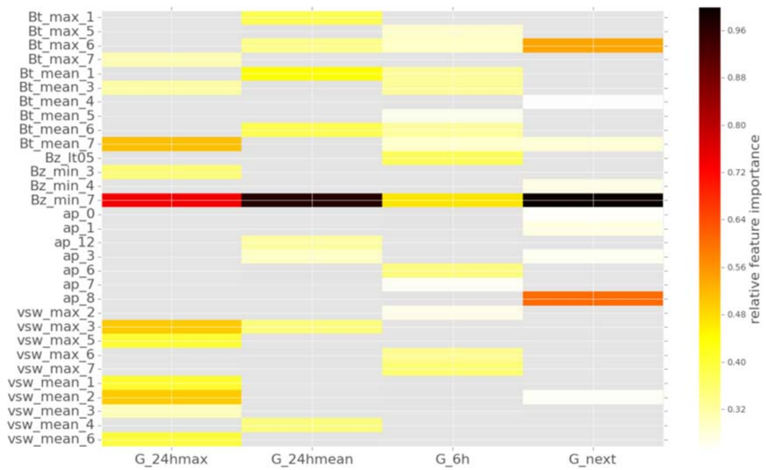
    $\Delta v_{SW} \propto \Delta |B_{IMF}|$

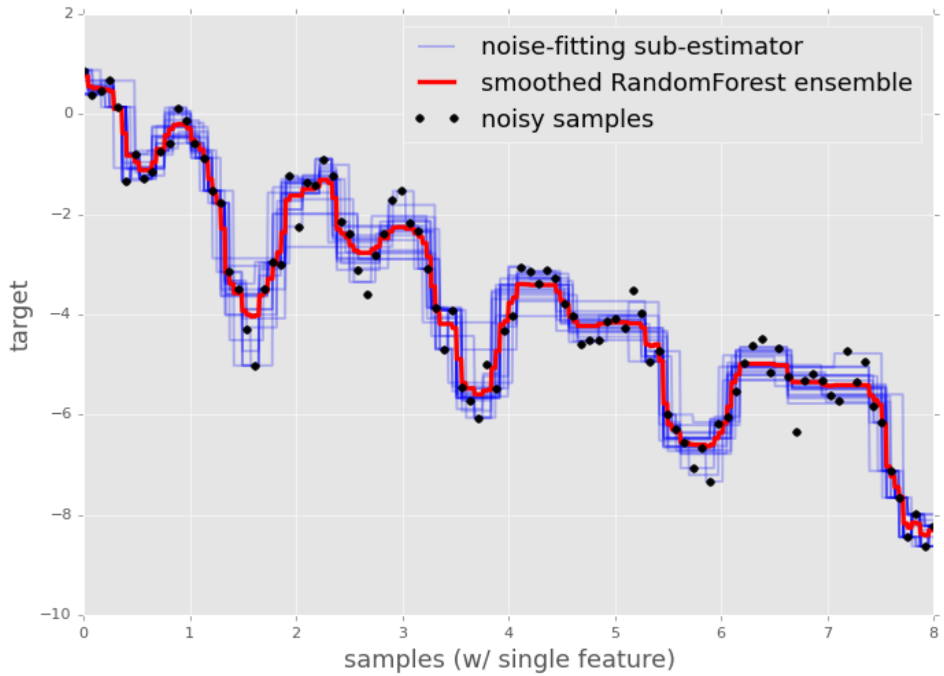
- timeseries:

  autocorrelation

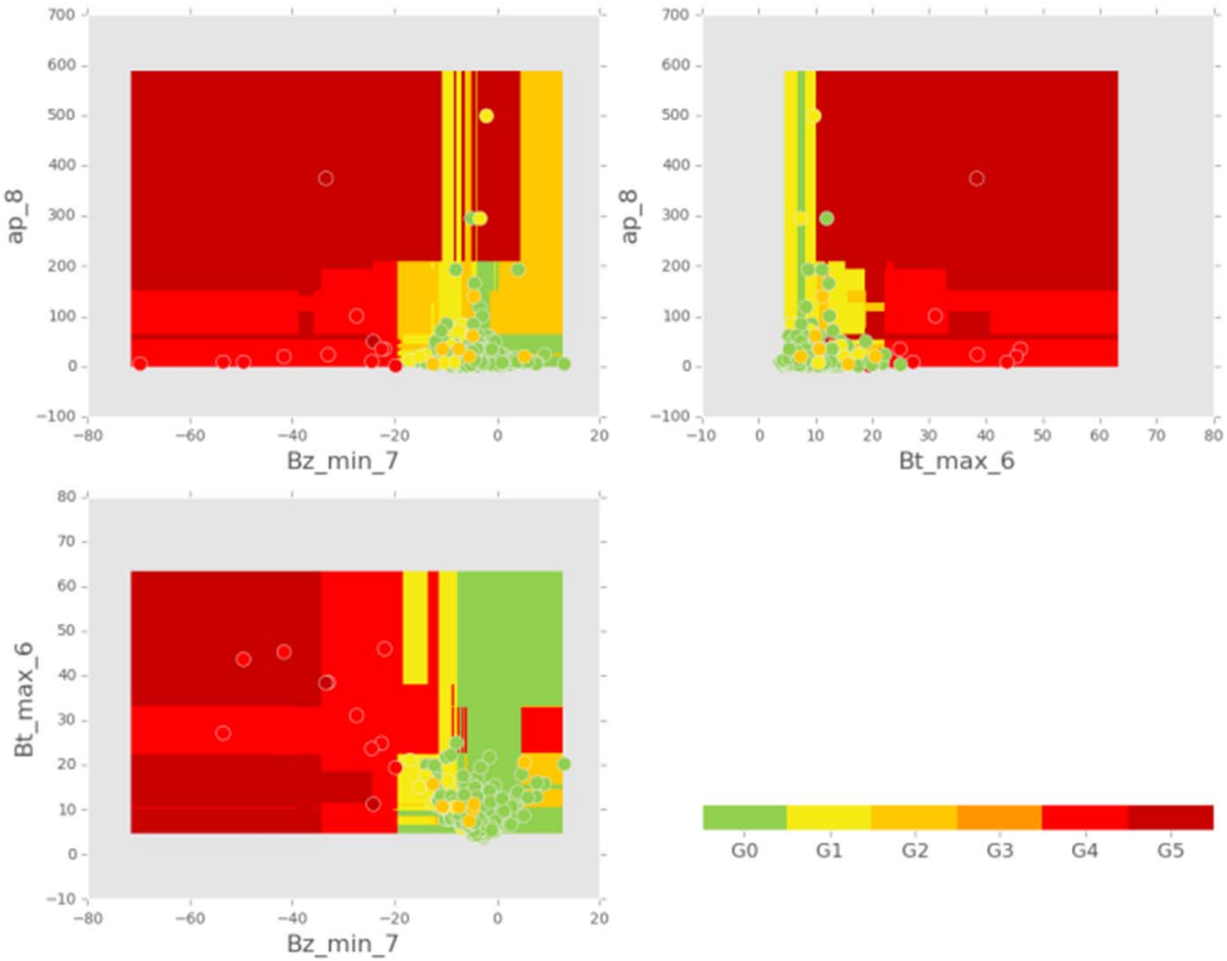
© NERC All rights reserved

Random forest classifier feature importances > 0.25 max (all targets)





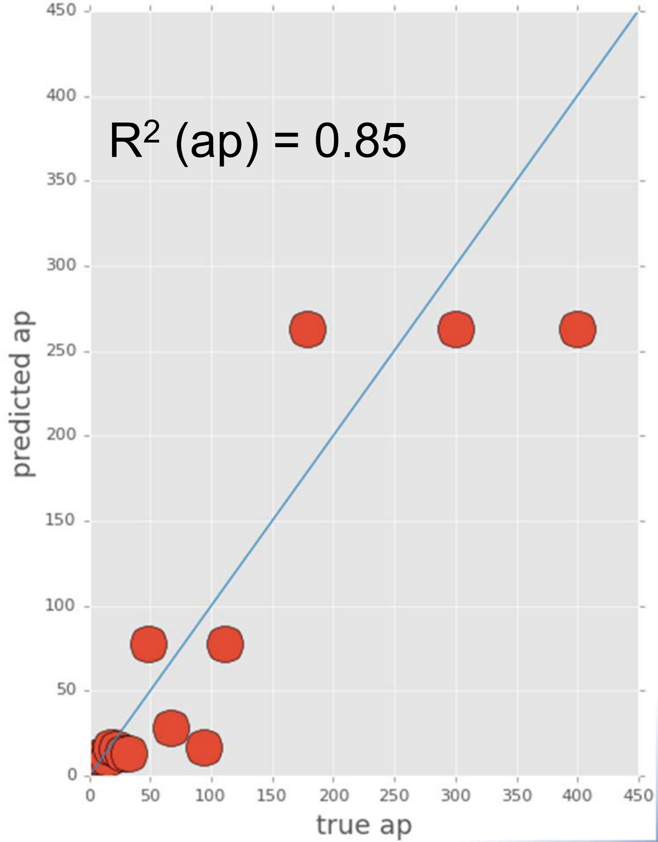
© NERC All rights reserved



© NERC

Metrics

- DecisionTreeRegressor single for model `ap_6h`
- R^2 (ap) = 0.85
- just good at predicting quiet
- R^2 may not be appropriate given fat-tail of ap (and Kp)
- similar argument for MSE



© NERC All rights reserved