Predicting Selenium content from Nutritional Labels in Seafood

Halle Ritter

Why Selenium (Se) in Fish?

- Harm reduction
- Some studies show protective factor against Hg toxicity
  - (Regardless of whether it is or not, factor of interest to consumers)
- Not listed on nutritional labels

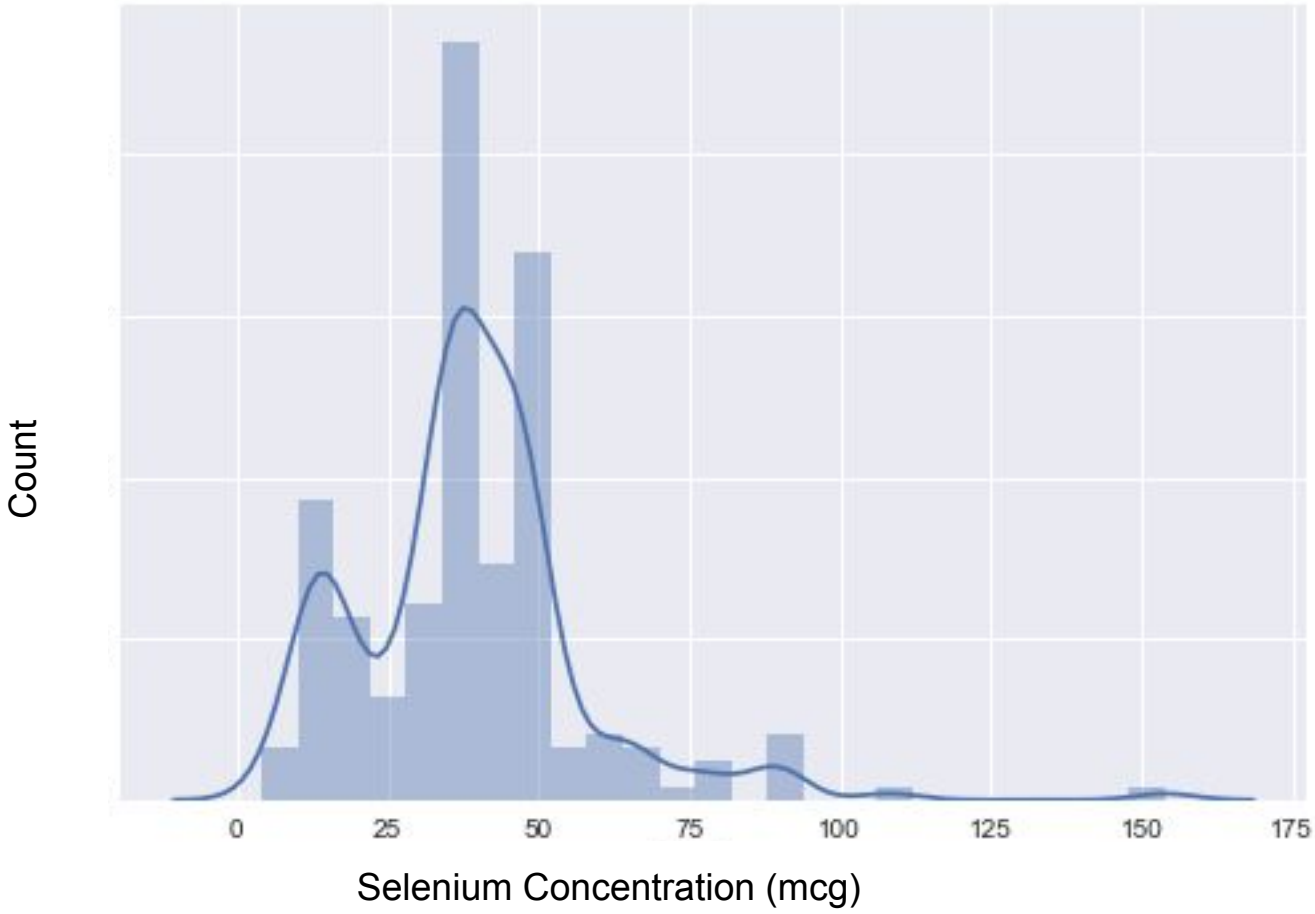
Objective

Create model **that a consumer could use in-store** to predict Se content (DV) from nutrients on label of fish (IVs/features).

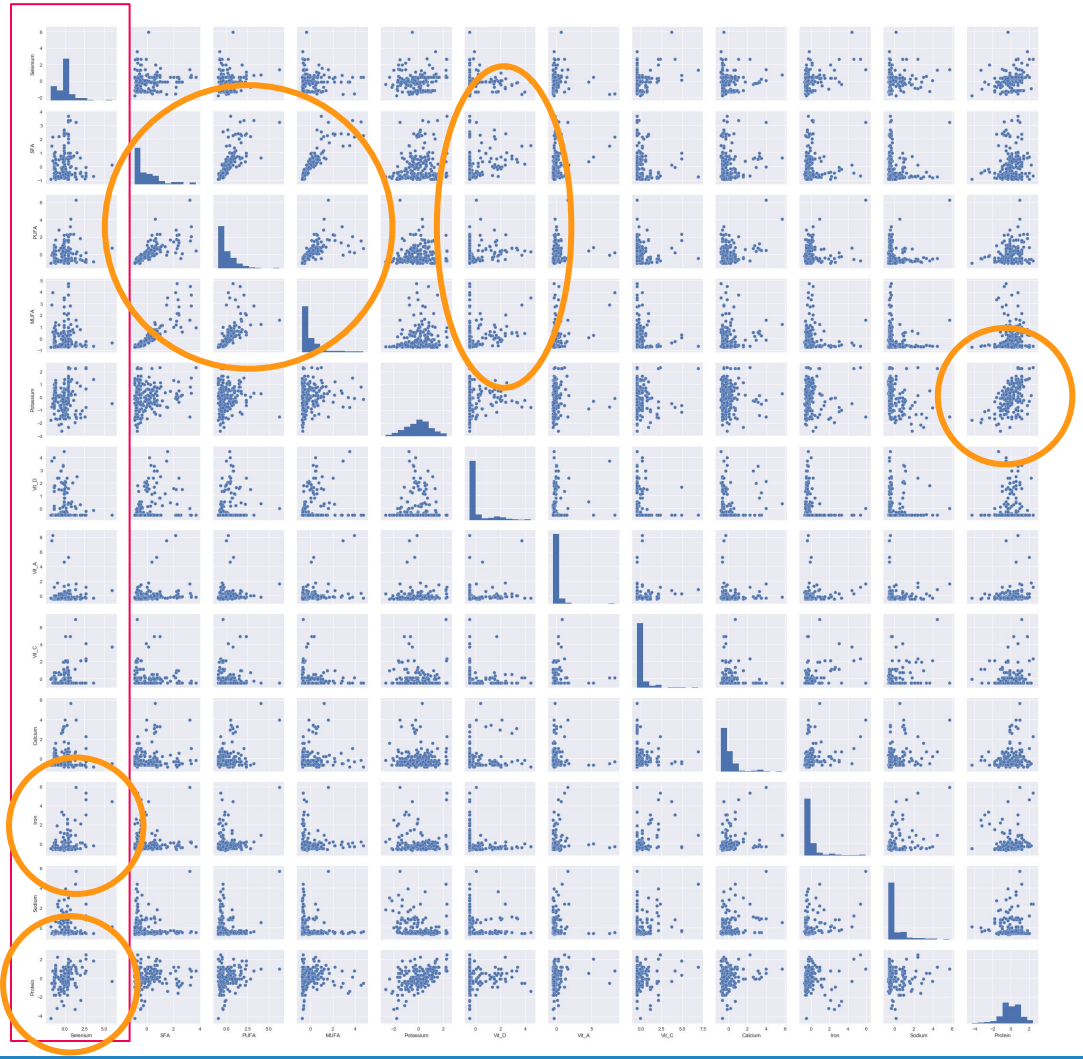
Data Characteristics

- Wrote script to download from nutritionvalue.org , from USDA
- 206 entries, 168 in no-outlier set
- 11 factors: Protein; Vit's A, C, D; Fe, Na, K, Ca, SFA, PUFA, MUFA

	Selenium	SFA	PUFA	MUFA	Potassium	Vit_D	Vit_A	Vit_C	Calcium	Iron	Sodium	Protein
Fish												
Fish, raw, ling	36.5	0.120	0.220	0.090	379.0	0.0	2.0	0.0	3.0	4.0	6.0	38.0
Fish, raw, cusk	36.5	0.130	0.280	0.090	392.0	0.0	1.0	0.0	1.0	5.0	1.0	38.0
Fish, raw, carp	12.6	1.083	1.431	2.328	333.0	988.0	1.0	3.0	4.0	7.0	2.0	36.0
Fish, raw, spot	36.5	1.450	1.090	1.330	496.0	0.0	2.0	0.0	1.0	2.0	1.0	38.0
Fish, raw, scup	36.5	0.640	1.030	0.560	287.0	47.0	2.0	0.0	4.0	3.0	2.0	38.0

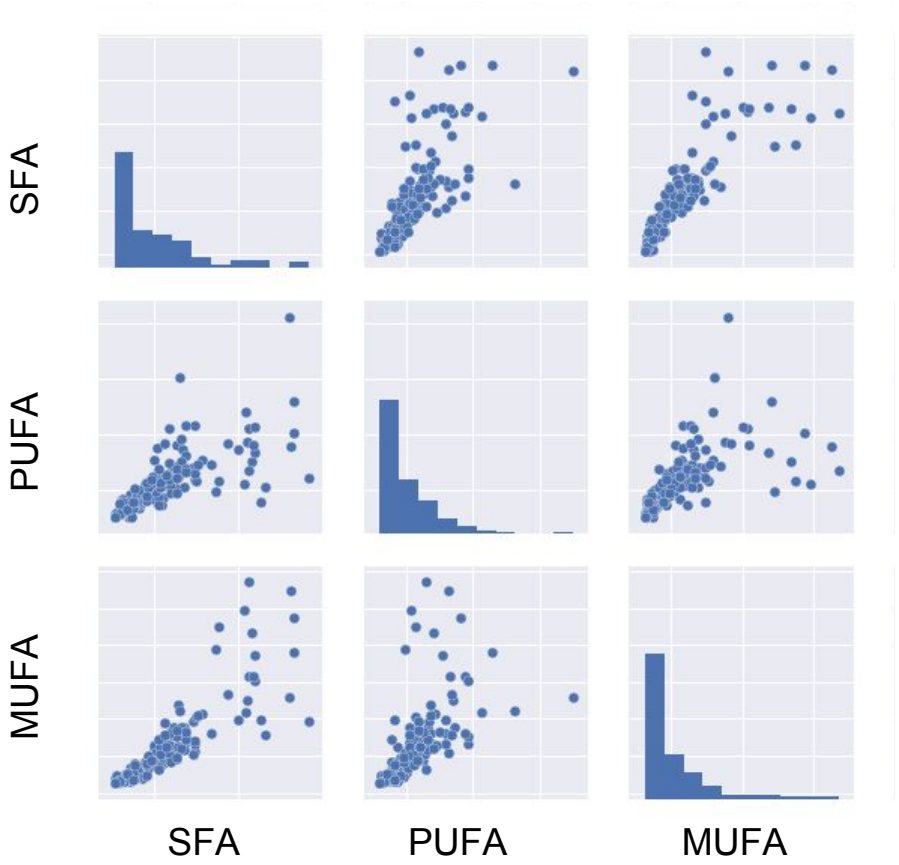

Count
0
25
50
75
100
125
150
175
Selenium Concentration (mcg)

Pairplot matrix: Every variable plotted against every other variable

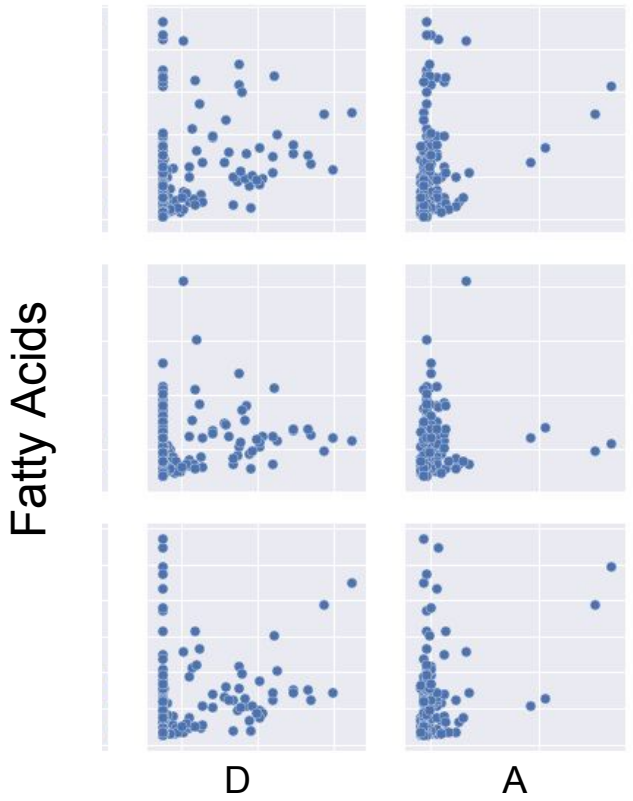


Circled areas of interest detailed on slides following

Intercorrelation of Fatty Acids

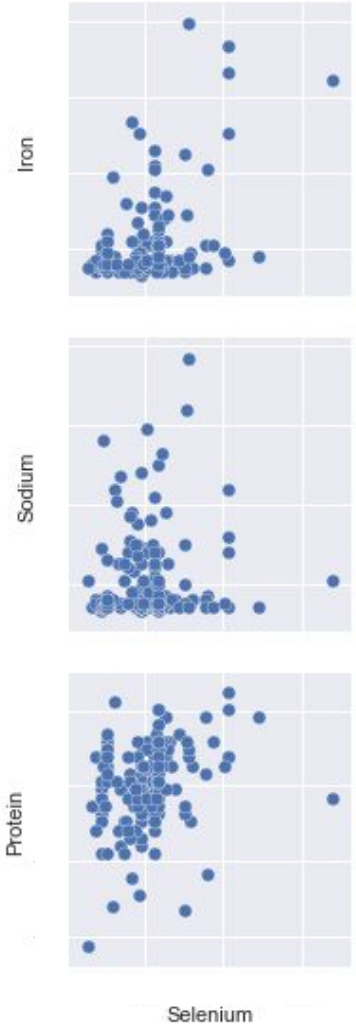


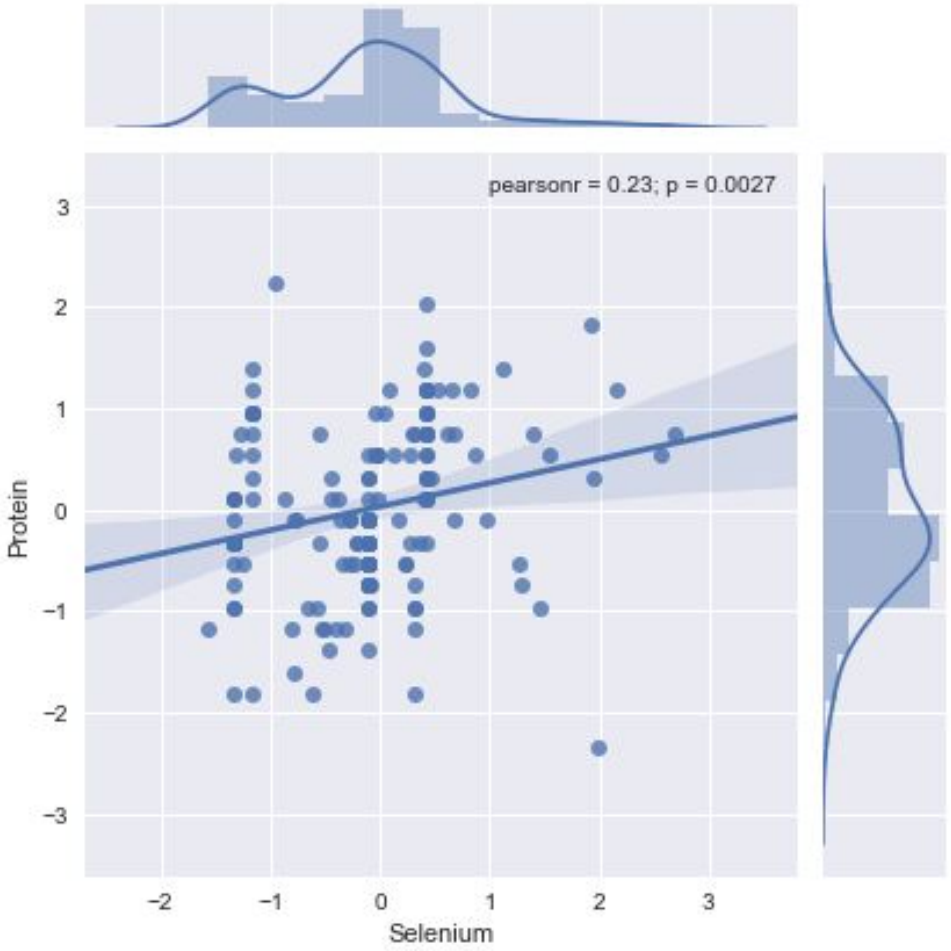
Fat-soluble vitamins



Selenium Correlations?

Not immediately clear from visual examination of scatterplot




pearsonr = 0.23; p = 0.0027
Protein
Selenium
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

Analysis Considerations

- Optimize for interpretable model
  - Not necessarily most complex fit
- Small data set drives cross-validation method
- Suboptimal qualities of data set reflect underlying reality
  - Size, multicollinearity, feature set

Random Forest Machine Learning Models

	Raw R^2	K-fold cross-val R^2	Ranked Feature Importances
Full Data Set	0.58	0.21	('Iron', 0.18), ('Protein', 0.14), ('Potassium', 0.14), ('Vit_C', 0.11), ('Calcium', 0.09), ('PUFA', 0.08), ('MUFA', 0.07), ('SFA', 0.06), ('Sodium', 0.06), ('Vit_A', 0.04), ('Vit_D', 0.03)
Outliers Removed	0.56	0.15	('Potassium', 0.18), ('Protein', 0.14), ('Calcium', 0.14), ('PUFA', 0.10), ('SFA', 0.09), ('Sodium', 0.08), ('MUFA', 0.07), ('Iron', 0.07), ('Vit_C', 0.06), ('Vit_D', 0.04), ('Vit_A', 0.03)

Regressions

	Model	R^2	Cross-val R^2	Feature Importance (top)
Simple Linear Regression	All features degree =1	0.35	0.11	('Iron', 0.48), ('Protein', 0.41), ('Vit_A', 0.19)
Polynomial Pipeline	All features degree =3	1	-7500	n/a: Overfit
Polynomial- Pick and Choose	All features degree = 1; protein & iron = 3	0.37	0.059	('Protein', 0.48), ('Iron2', 0.34), ('Vit_A', 0.20),

Feature Importances b/w Models

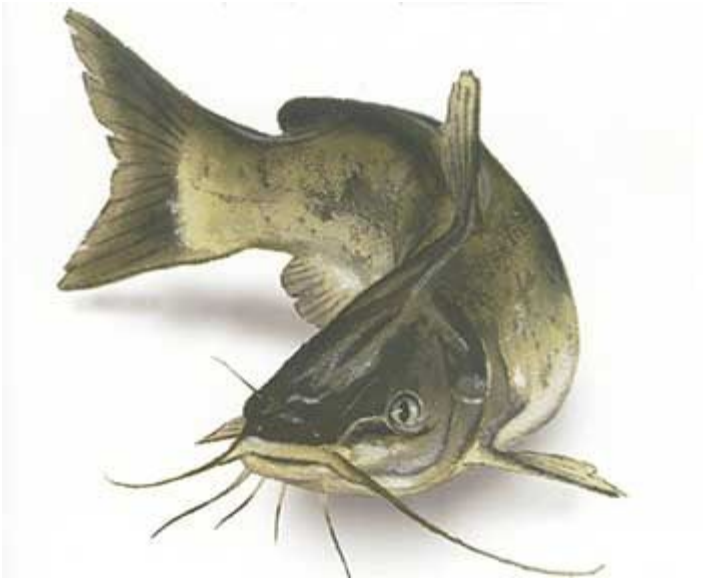
- Usually unimportant features:
  - Vitamins (esp. C and D)
- Occasionally important:
  - Iron, protein/potassium
  - ...but feature importances rearrange between models = low model confidence

Conclusions

- Not much useable predictive ability for Se with this data set
  - Not much correlation- makes sense biochemically
  - This is a feature of the data set, not models: models accurately demonstrate lack of underlying correlation
- Interesting other correlations (e.g. fatty acids)
- Negative results are results!

Future Directions

- Other strategies of outlier removal
- Better Se data source
- Omega fatty acid ratios
- Join Hg data set



Top Fish by Se Content

	Selenium	(ug/100g)
Fish		
Mollusks, moist heat, cooked, Pacific, oyster	154.0	
Fish, dry heat, cooked, fresh, yellowfin, tuna	108.2	
Fish, raw, yellowfin, fresh, tuna	90.6	
Mollusks, moist heat, cooked, mixed species, cuttlefish	89.6	
Mollusks, moist heat, cooked, common, octopus	89.6	
Mollusks, moist heat, cooked, blue, mussel	89.6	
Fish, dry heat, cooked, orange, roughy	88.3