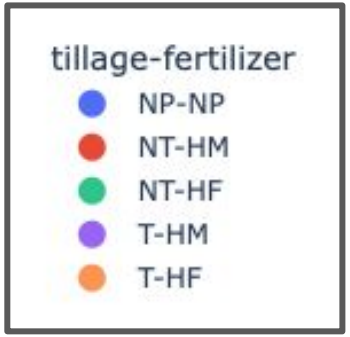
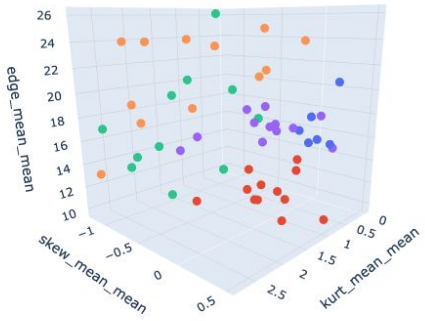
Quantifying heterogeneity and distinguishing soils

8/3/23

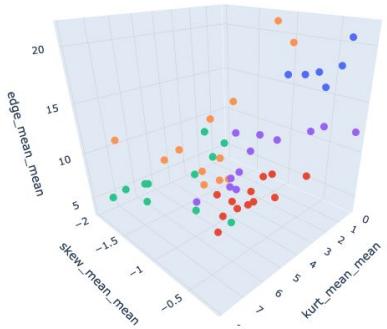
Previously

- I extracted 3 averaged statistical metrics on sparsely sampled horizontal x-ray slices (every 50 horizontal slices).
  - Skewness, kurtosis (both on a 50 px sliding window, sparsely sampled), and sobel edge convolution
  - For each scan, I averaged those values into one of 4 depth bins.
  - So 12 total features for each scan: skew, kurt and sobel averages at 4 depth bins.

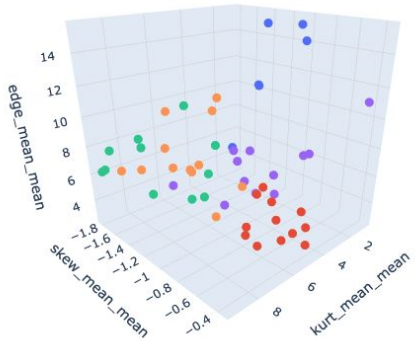




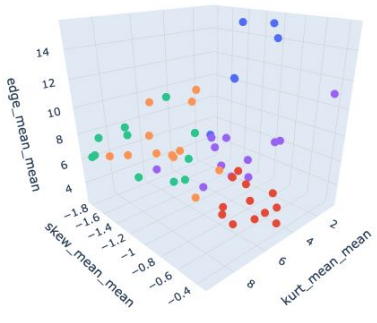
0 - 3.3 cm



3.3 - 6.6 cm



6.6 - 9.9 cm



9.9 - 12.9 cm

Features are predictive of soil type

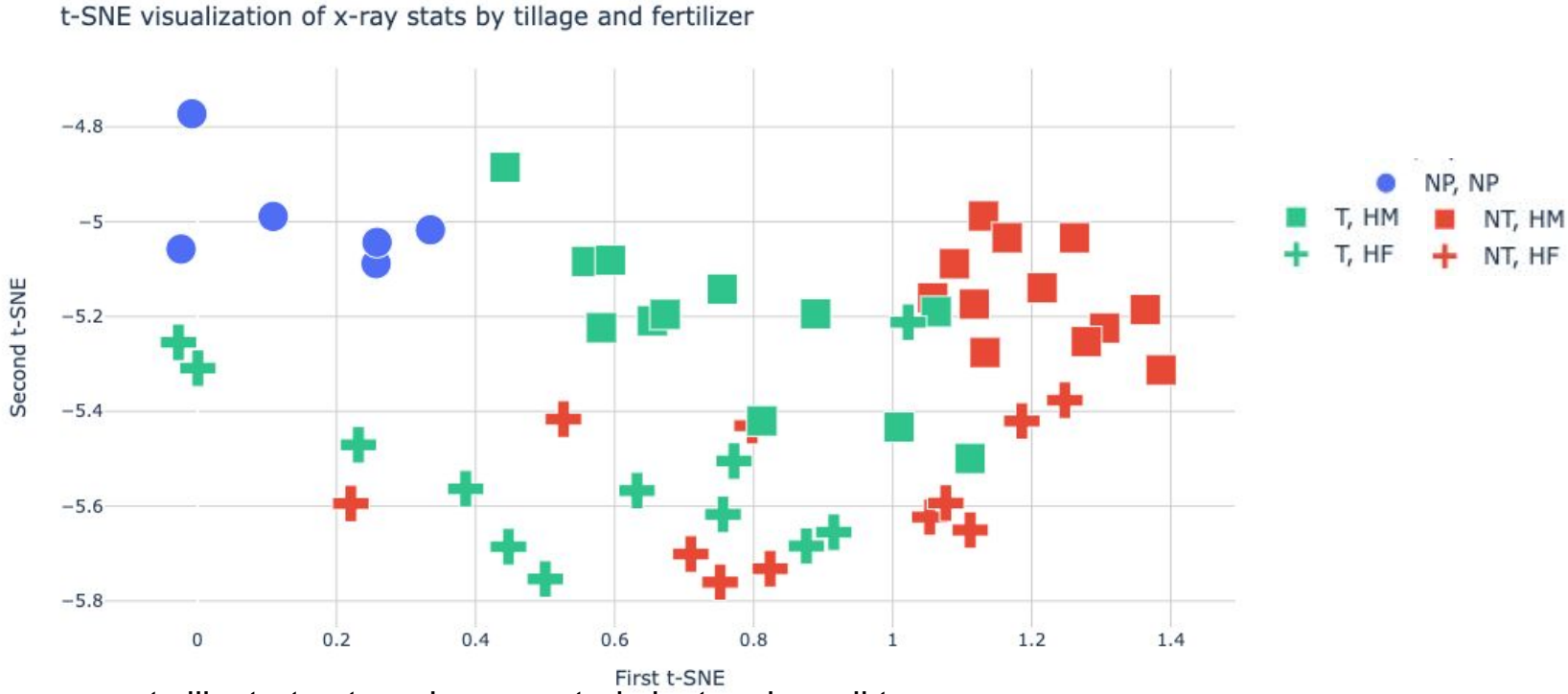
- I trained a support vector machine (SVM, a classification technique) to **predict soil type** from the 12 statistical features calculated for each scan:
  - SVM is traditionally a binary classification technique, but we have more than 2 soil types. The python library sklearn has default support for multi-class SVM. I used the default one-v-rest approach and a linear kernel to predict membership.
  - To determine the accuracy of the SVM model, I used **cross-validation with 5 folds** (for each fold, the model was trained on 80% of the scans to predict the soil type of other 20% of the scans), and compared the accuracy of the predictions.
    - I used stratified group k-folds to preserve the distribution of each soil type in the folds.
  - Accuracy = total correct predictions / total predictions
  - Results:

Classification Type:	Tillage-Fertilizer	Tillage	Fertilizer
Fold 1 accuracy	0.73	0.91	1.00
Fold 2 accuracy	0.82	0.91	0.91
Fold 3 accuracy	0.73	0.91	0.73
Fold 4 accuracy	0.90	0.70	0.90
Fold 5 accuracy	0.90	0.90	0.90
Mean accuracy	**0.81**	**0.87**	**0.89**

Reducing dimensionality to extract qualitative insights

- Principal Component Analysis (PCA) is great for data that has mostly linear relationships
- Our data is rather complex; even for a single depth bin, the points lie on a curved 3D manifold rather than a cluster that resembles a line or plane.
- I instead used **t-distributed stochastic neighbor embedding (t-SNE),** a **non-linear** dimensionality reduction technique, to map data from 12 features to 2 components.
- **The goal was to understand how the soil-types seperate and characterize those differences**.

Reducing dimensionality to extract qualitative insights



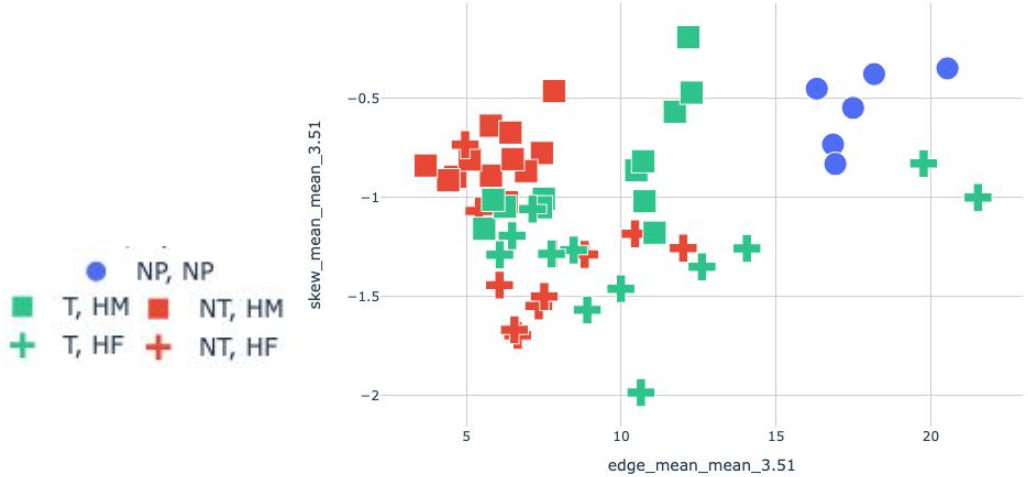
t-SNE components illustrate strongly separated clusters by soil type.

ANOVA:

- Significant differences in t-SNE_1 by both fertilizer (P~1E-8), and tillage (P~1.8E-5)
- Significant differences in t-SNE_2 by fertilizer (P~1E-12)

Reducing dimensionality to extract qualitative insights

- t-SNE is excellent for visualizing clusters.
- A downside is no direct or deterministic mapping between original features and t-SNE components
- But we can still observe strong correlations for different components:
- t-SNE_1 - strongly correlated with sobel edge features; coefficients below:
  - edge_mean_mean_10.14 -0.631848
  - edge_mean_mean_0.0 -0.757746
  - edge_mean_mean_6.825 -0.860923
  - **edge_mean_mean_3.51 -0.931149**
- t-SNE_2 - strongly correlated with skew and kurtosis (which seem negatively correlated with each other) coefficients below:
  - **skew_mean_mean_3.51 0.877922**
  - skew_mean_mean_0.0 0.834085
  - skew_mean_mean_6.825 0.747661
  - skew_mean_mean_10.14 0.582238
  - kurt_mean_mean_0.0 -0.626756
  - kurt_mean_mean_3.51 -0.785171
  - kurt_mean_mean_10.14 -0.797555
  - kurt_mean_mean_6.825 -0.860753



Conclusion

- Skew, Kurtosis, and Sobel edge seem to be good metrics to differentiate soils
- Changes in skew and kurtosis seem indicative of fertilizer vs manure.
- Changes in edges seems indicative of tillage differences
- Depth seems important, especially to differentiate tillage of high fertilizer soils
- Future directions:
    - How do these metrics connect with physical metrics like drainage, nutrient transport, etc? More experiments required.
    - How do these metrics connect to pore structure metrics?
    - How do bulk density measurements compare against x-ray calculated density values?