Author: Slava Zagriichuk

NEAR INFRARED SPECTRA ANALYSIS OF

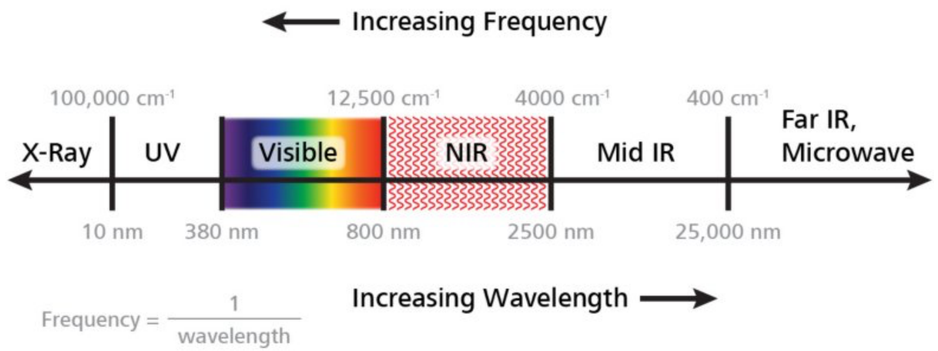
MILQ COMPOSITION

WHAT DOES MILQ CONTENT?

- Explain briefly the technological process and the nature of the given data.
- Show and describe the distribution of essential substances.
- Show and describe the spectrograms.
- Build models to predict milk composition using spectrograms.

WHY AND HOW DO WE DO SO?

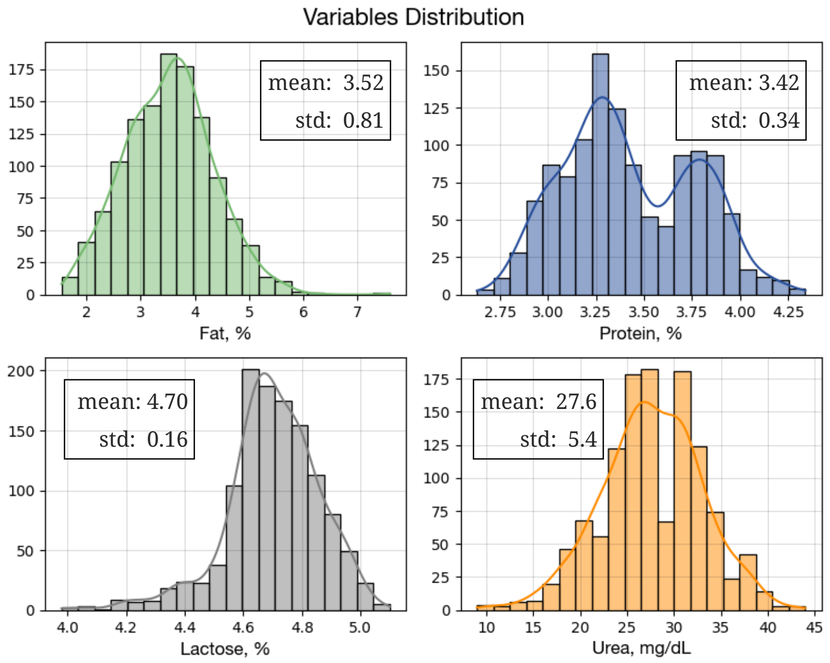
- HOW?
  - Why do we see different colors? Materials absorb certain parts of the light spectrum and reflect or transmit others. The colors we see are the reflected or transmitted parts. The spectrum itself is produced by a light source, such as a bulb or the sun.
  - By observing color and brightness, we can compare the turbidity and color of water, which can indicate the concentration of substances in it. While this method is not so precise, as human perception in principle, it allows us making some comparative analysis.
  - Fortunately, there is a broader spectrum beyond visible light, and we have fancy sources of electromagnetic radiation and sensors that allow us to measure things much more accurately. However the principles behind these measurements are similar to what I described above, and this field of study is called spectroscopy.
- WHY?
  - We perform measurements to gain control over processes. With the help of computers, spectroscopy has become easier and efficient method to research the matter.
  - In this study, the Near Infrared (NIR) part of the spectrum was used to avoid spoiling the milk. You can find NIR on the spectrum in the image below.



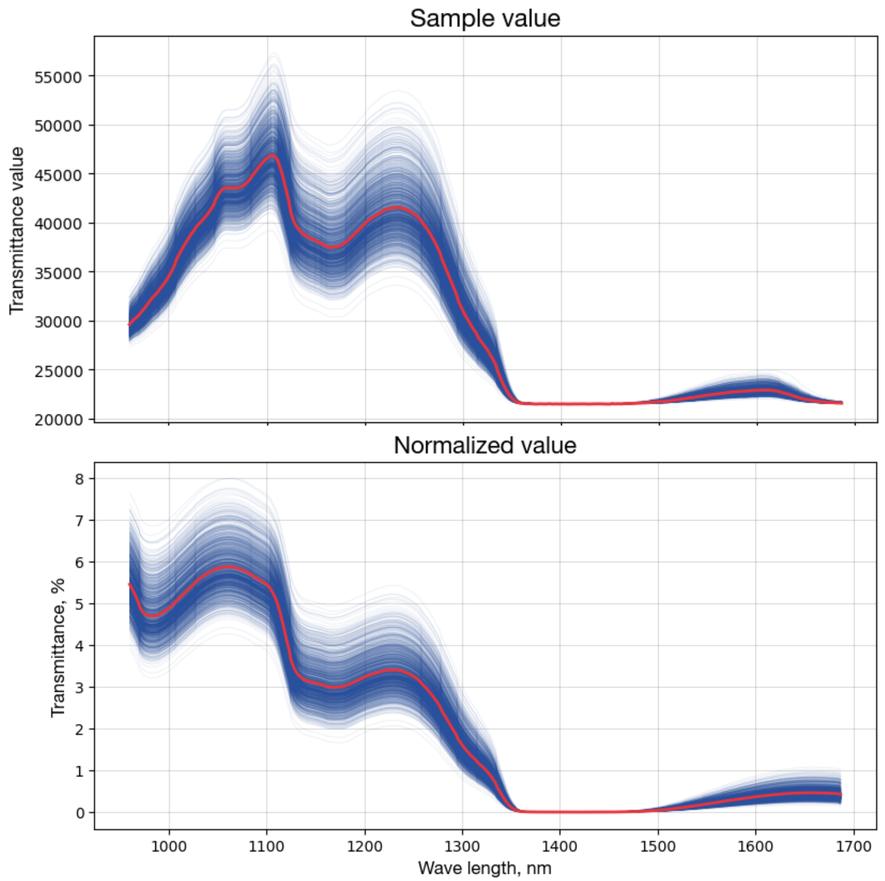
WHAT AND HOW WAS MEASURED?

- WHAT?
  - The content of essential substances in the milk was measured using precise tests. Metadata includes information such as date and time, yield in liters, and the time elapsed since the previous yield.
  - Transmittance: digital values directly from the sensors and relative transmittance in percent calculated with certain formula.
  - Number of substances and metadata parameters is 14. Number of measurements parameters is 256 x 4, i.e. 1024.
  - Number of samples is 1224.
- HOW?
  - As a source of radiation there was used 16x16 diode matrix with the wavelength range 960-1690 nm. Sensor is digital with range 0-65535. Each resulted value is the average of 100 measurements.
  - To measure relative transmittance there was made additional measurements in the darkness and in the lightness. So there are 256 just digital values, 256 values in darkness, 256 values in lightness and 256 relative transmittance that was calculated using formula described in the source materials.
- IMPORTANT !
  - Detailed information about the research method and the equipment used can be found in the pdf file in the same folder or at the link: https://www.sciencedirect.com/science/article/pii/S235234092300834X

HOW DOES THE DATASET LOOKS LIKE?

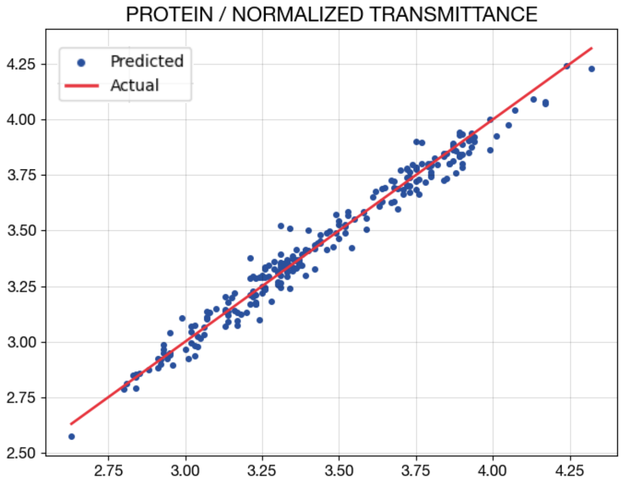


- Distributions of essential substances in milk.

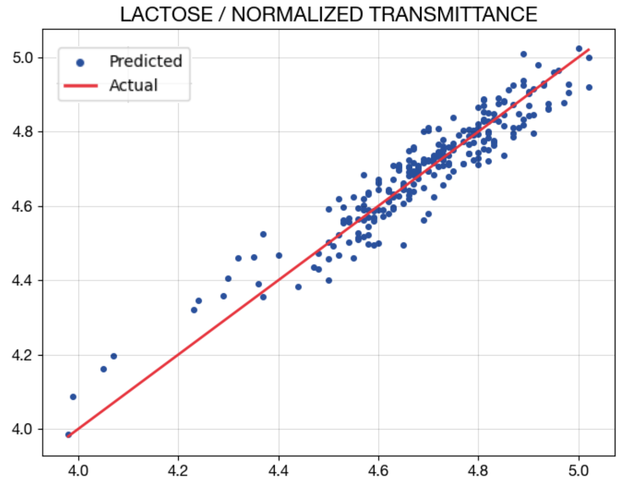


- Spectrograms with both digital and normalized transmittance values.
- Each blue line represents 1 sample, red line is the overall average.

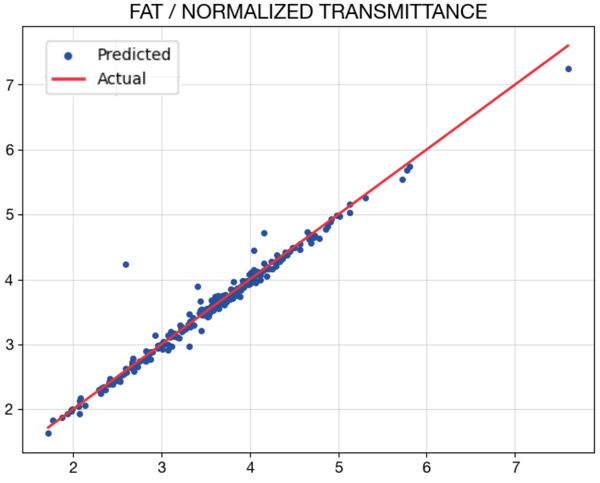
LINEAR REGRESSION: ACTUAL VS PREDICTED



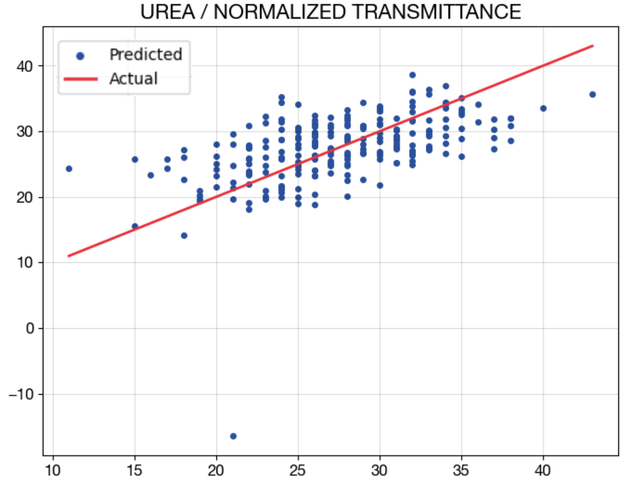
- R2 Score: 0.974
- MSE: 0.004
- MAE: 0.043
- Unit: %



- R2 Score: 0.894
- MSE: 0.003
- MAE: 0.045
- Unit: %



- R2 Score: 0.971
- MSE: 0.019
- MAE: 0.064
- Unit: %



- R2 Score: 0.037
- MSE: 24.950
- MAE: 3.720
- Unit: mg/dL

DO THE METHOD AND LINEAR REGRESSION MODEL WORK?

- Actually they do for protein, fat and lactose. For protein and fat linear regression model results are pretty consistent, for lactose a bit worse but still good.
- Linear regression model completely fails to predict the urea content, although it doesn’t mean that spectroscopy doesn’t work, maybe just wrong choice of the model.
- What could be improved?
    - Despite data has high quality, probably it could make sense to cut the extreme values to ensure smoother performance of linear regression model.
    - From train/test results comparison I noticed that models are slightly overfitted (it hasn’t been shown in this presentation), therefore cross-validation can be used to avoid overfitting.
    - Try other models for lactose to try to improve results, for example neural network model.
    - Definitely try other models for urea. Probably read about it’s properties to find out is it even possible to detect it with this wavelength range.