Nonparametric Analysis of **US Dairy Production and Consumption**

Teo Bucci
Filippo Cipriani
Gabriele Corbo
Andrea Puricelli

POLITECNICO MILANO 1863

Nonparametric Statistics (8 CFU)
MSc. Mathematical Engineering
Politecnico di Milano

February 17th, 2023

Dataset presentation

The data consists of **yearly** measurements of production and consumptions of **US dairy** products.

Although there are many variables available, we are **focusing** on:

Production
- **Milk price [$]**
- Milk production [lbs]
- Hay price [$] etc.

Consumption [lbs/person]
- Milk (milk, yogurt, cottage, …)
- Cheese (mozzarella, cheddar, swiss, …)

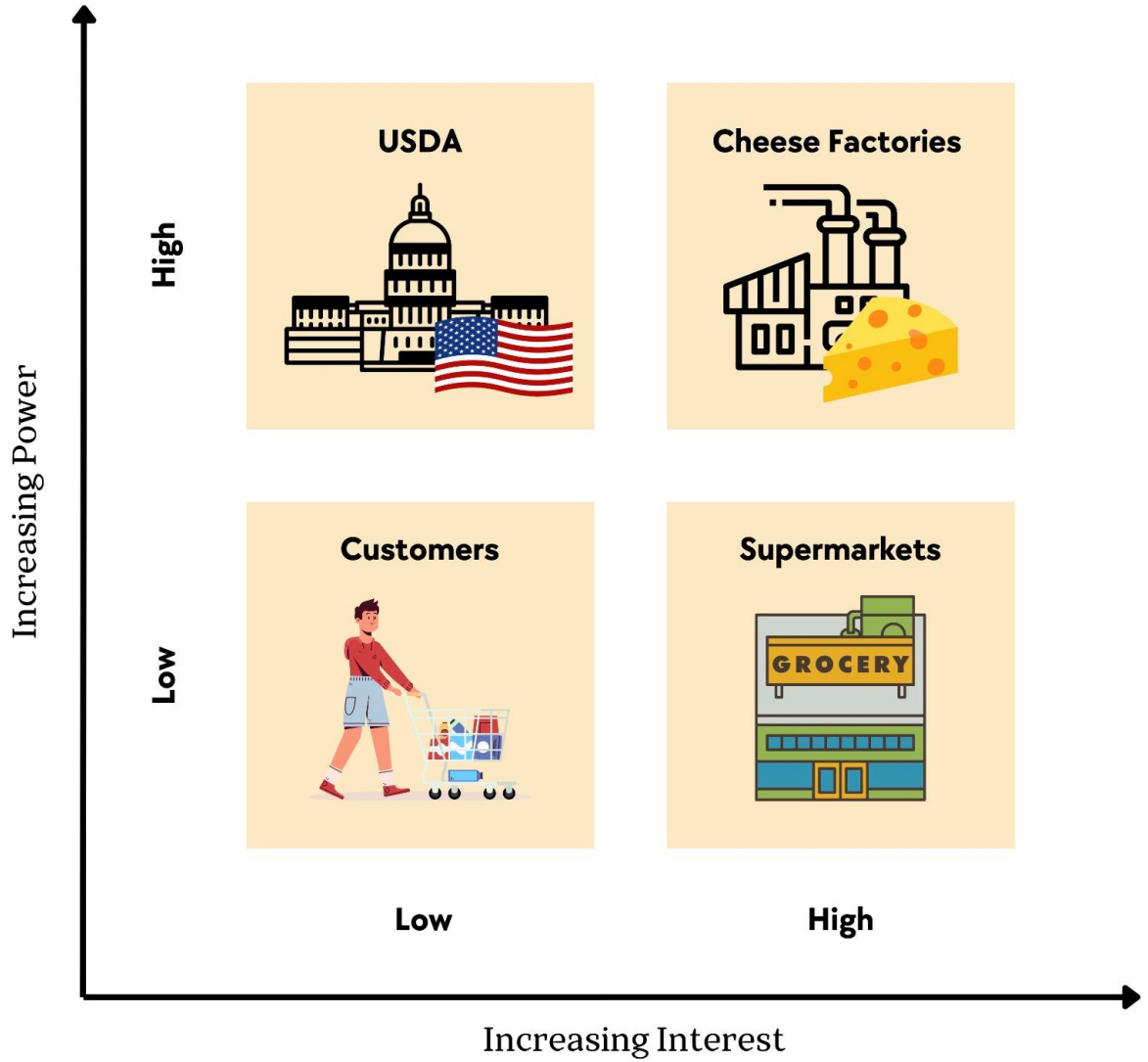
1980 – 2014: Data provided by tidytuesday

+

We retrieved data to fill the observations up to the past year

2015 – 2021

Stakeholders analysis



1

Key players

High power - High interest

Cheese producers and manufacturers

2

Informed or consulted players

High interest - Low power

Retailers and distributors

High power - Low interest

Government - USDA

3

Monitored players

Low power - Low interest

End users and customers

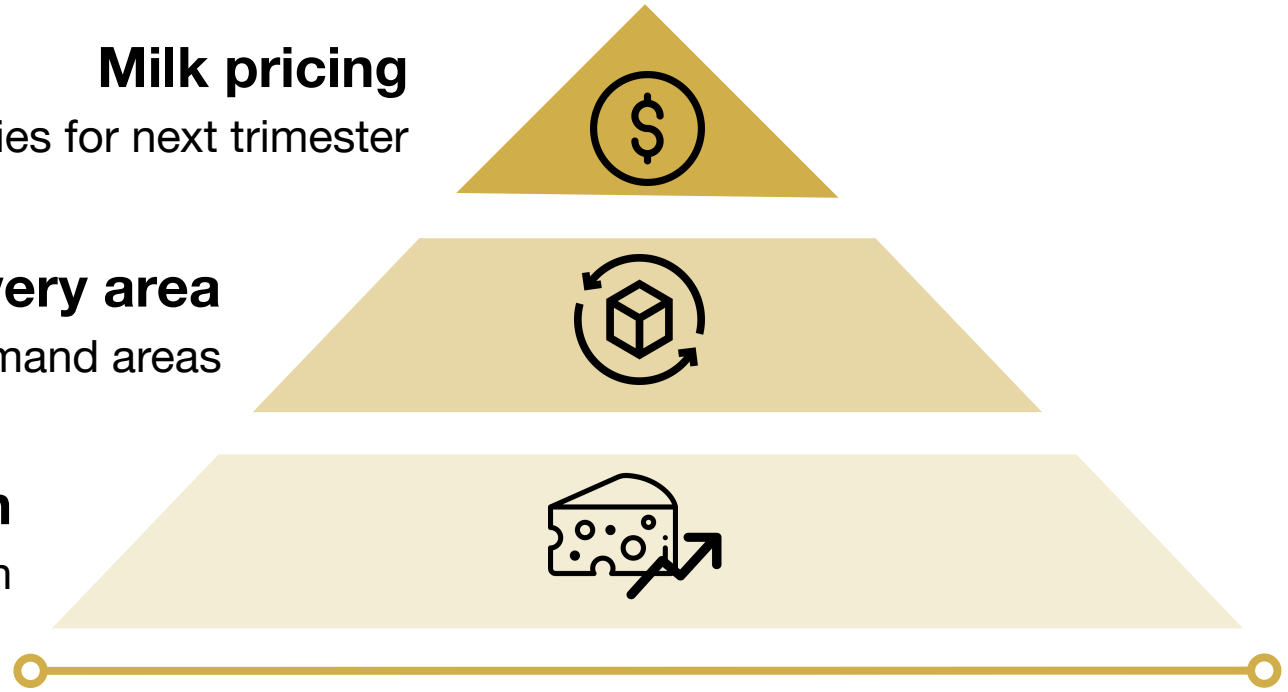
Research question

December 2022, a new competitor from **Salt Lake City, Utah** enters the market.
What should be their strategy?

Milk pricing
Schedule retailer deliveries for next trimester

Milk production quantity and delivery area
Based on sales, focus on high demand areas

Cheese type selection
Produce cheeses trending in consumption



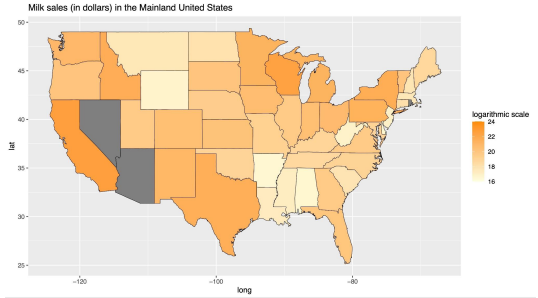
Data Pipeline

Adjust prices for **inflation**:[1]

$$\text{2022 USD value} = \frac{\text{CPI in 2022}}{\text{CPI in 1980}} \cdot \text{1980 USD value}$$

Embed **spatial** data

Milk and dairy sales (in [$]) **and population** for most counties of the US



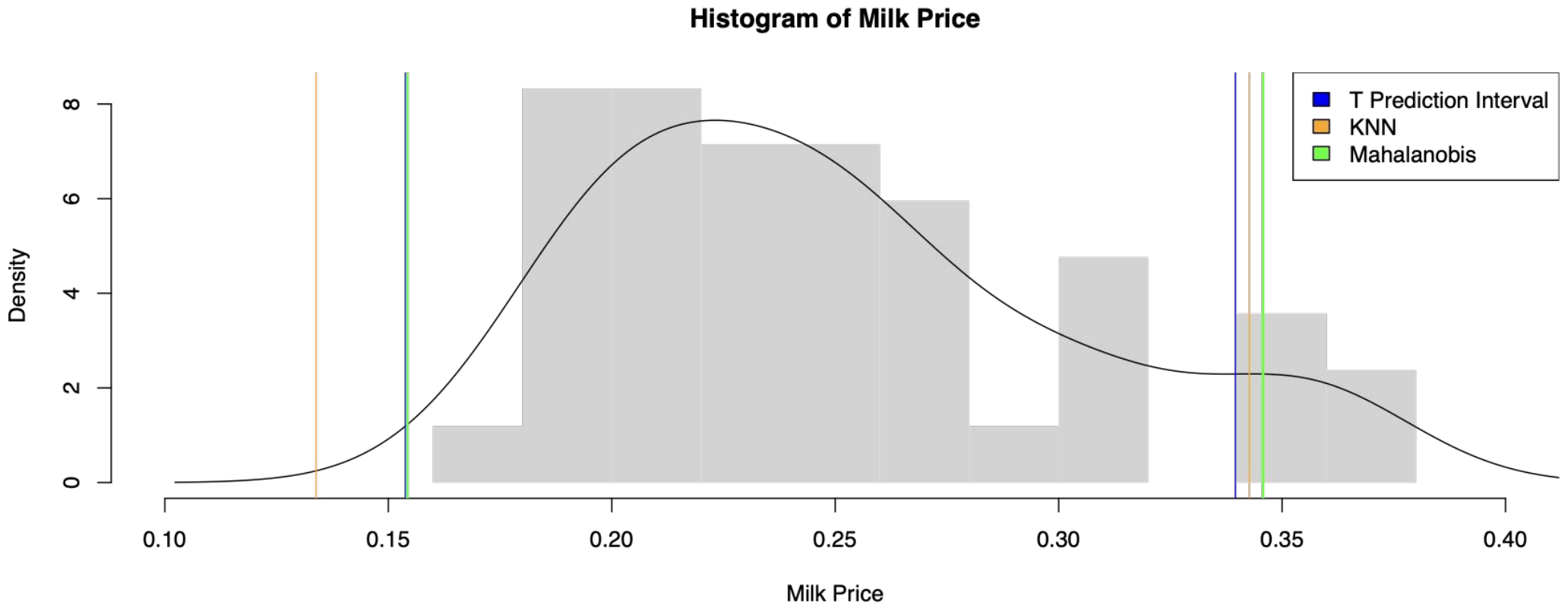
[1]CPI: Consumer Price Index

Conformal

Conformal Prediction

The 3 conformal intervals are respectively:
Blue: Using **T Prediction** interval.

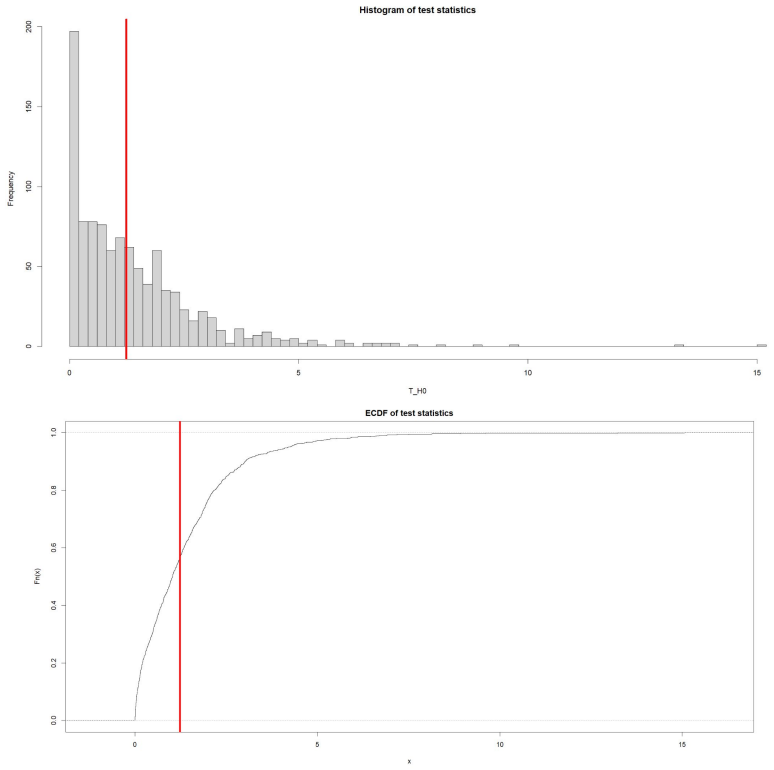
Yellow: Created using the **KNN** distance.
Green: Created using the **Mahalanobis** distance.



GAM

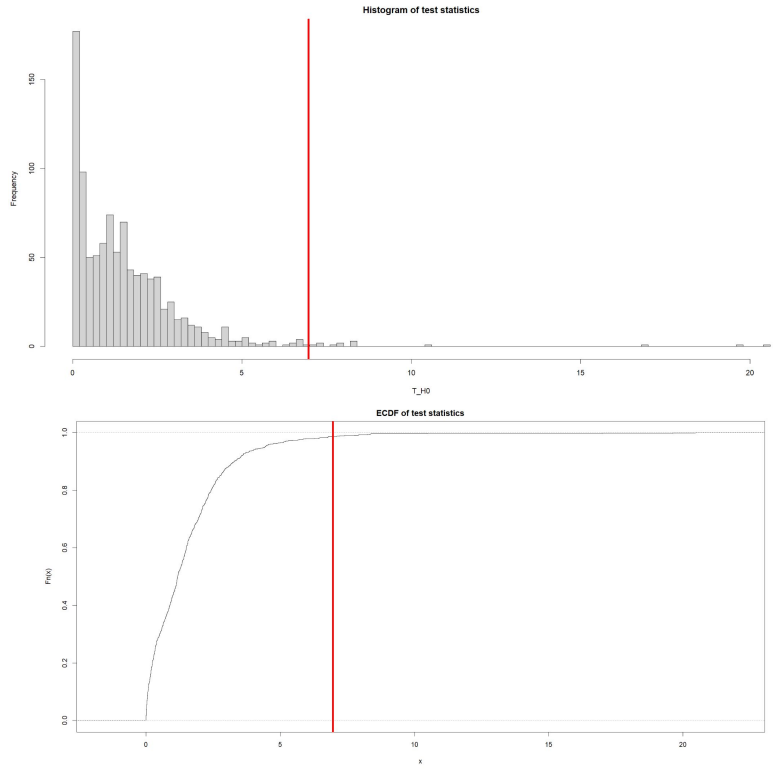
We performed multiple **Permutation Tests** to assess the significance of each covariate. For example:

$$H_0 : \beta_{\texttt{Slaughter cow price}} = 0 \quad vs \quad \beta_{\texttt{Slaughter cow price}} \neq 0$$



P-value: **0.431**

$$H_0 : \beta_{\texttt{Dairy ration}} = 0 \quad vs \quad \beta_{\texttt{Dairy ration}} \neq 0$$

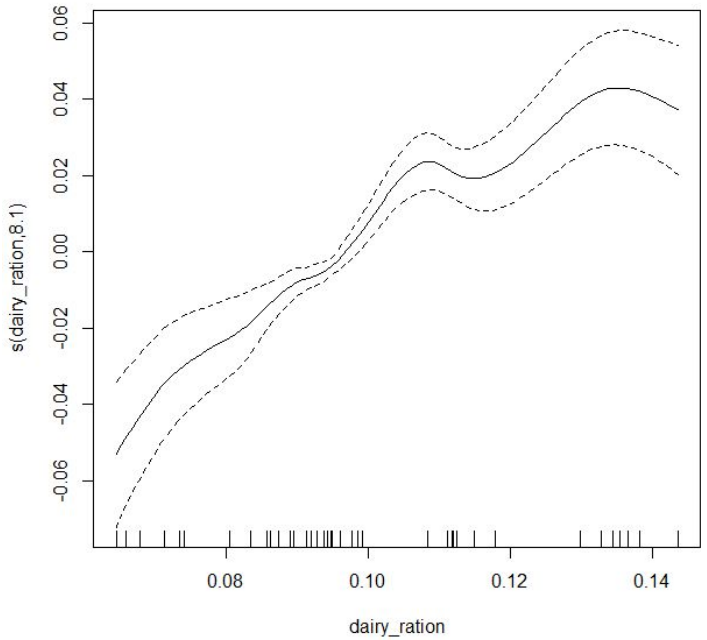


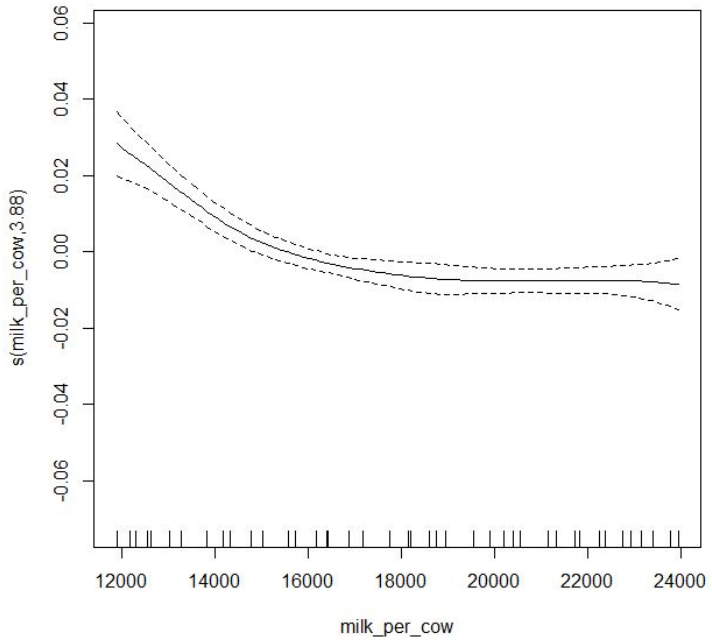
P-value: **0.014**

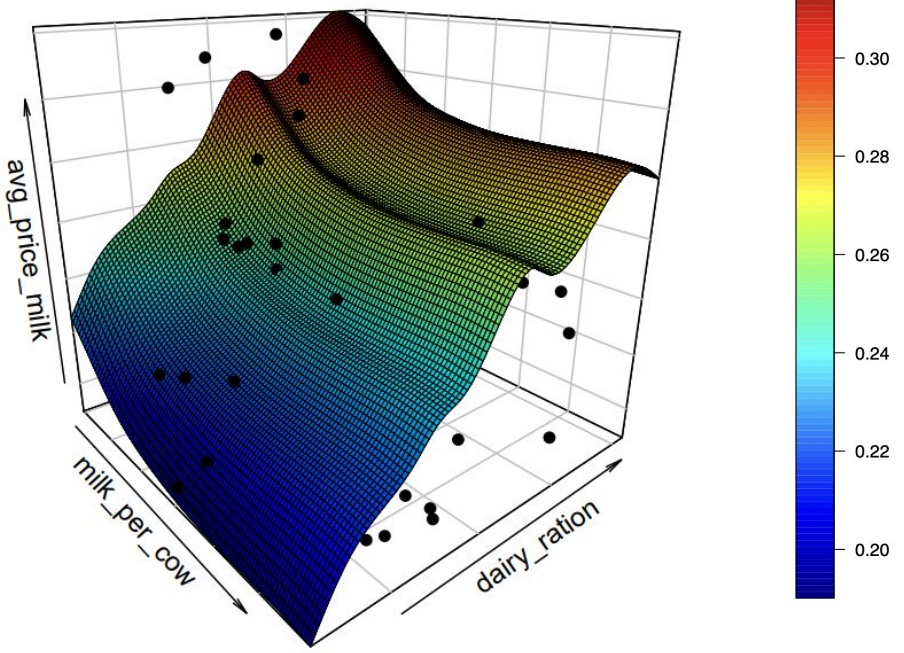
GAM

$$\begin{aligned}\texttt{avg\_price\_milk}_i = \beta_0 &+ \beta_1 \cdot \texttt{milk\_cow\_cost\_per\_animal}_i \\ &+ \beta_2 \cdot \texttt{milk\_volume\_to\_buy\_cow\_in\_lbs}_i \\ &+ \beta_3 \cdot \texttt{milk\_feed\_price\_ratio}_i \\ &+ f_1(\texttt{dairy\_ration}_i) + f_2(\texttt{milk\_per\_cow}_i) + \epsilon_i\end{aligned}$$

$$\beta_1, \beta_3 > 0, \quad \beta_2 < 0$$







GAM

Milk pricing

Delivery deals with private retailers are commonly stipulated on a trimestral basis.

Cost of a milk cow
Price of a dairy ration
Dairy cow feed for amount of milk produced

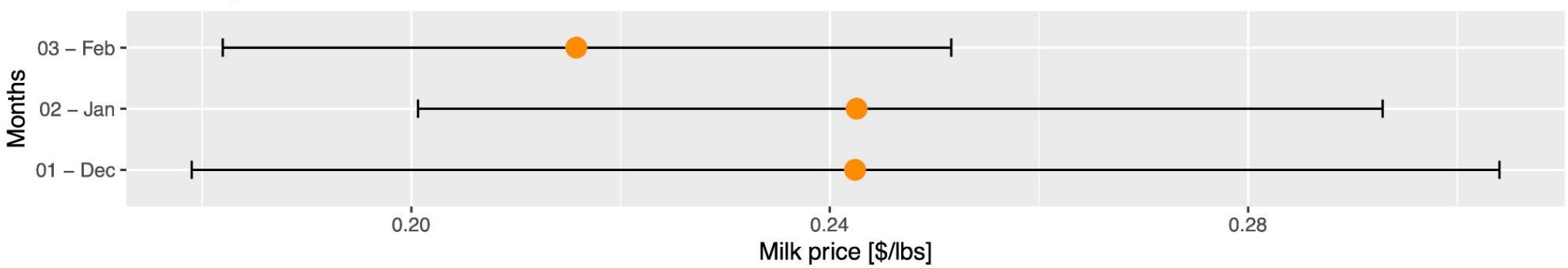
+

Milk produced per cow
Dairy cow value in milk production potential

USDA Projections released by the **USDA**
December, January, February

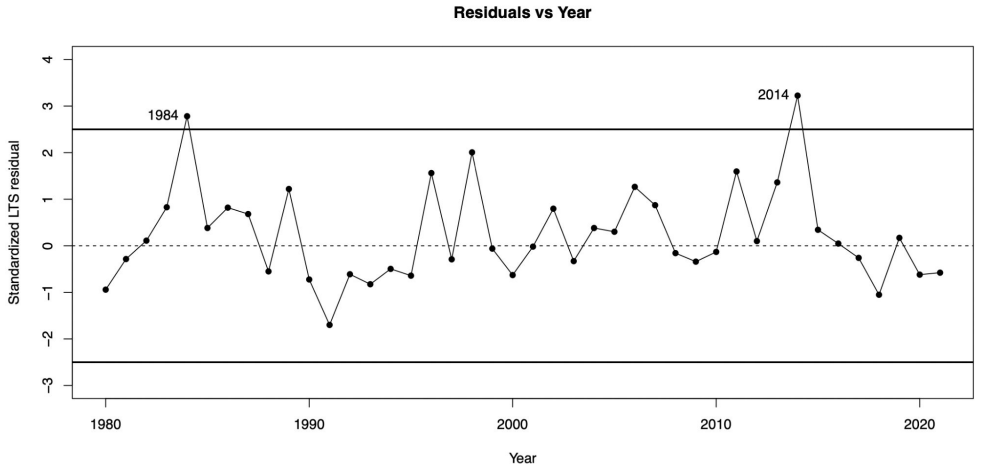
Median values from previous years
2010 – 2021

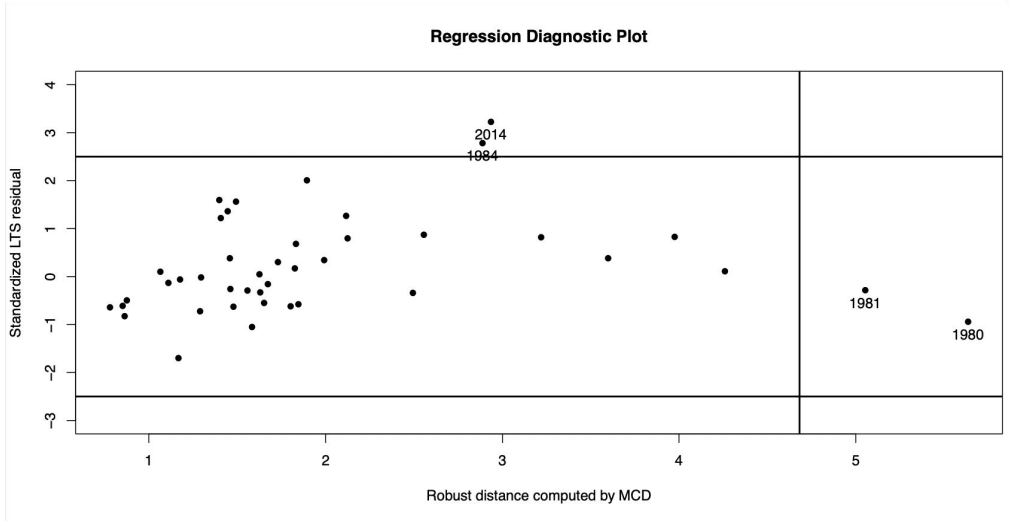
Predicted milk price for trimester + Reverse Percentile (Bootstrap) intervals α=0.1



Robustness

Robust Regression





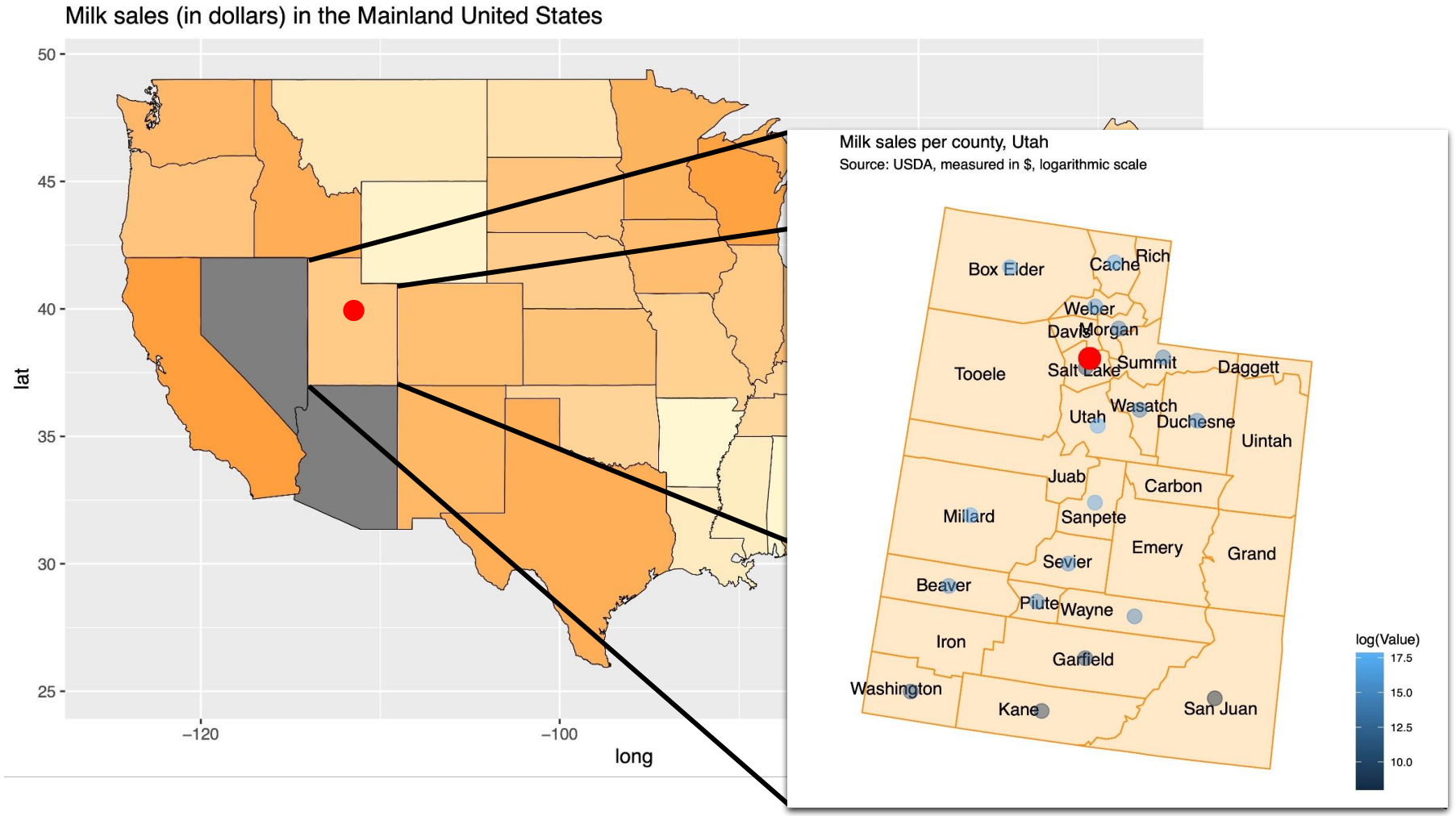
We can note two vertical outliers, corresponding to the years **1984** and **2014**.

1984 — **Dairy Price Support Program (DPSP)**, aimed at stabilizing milk prices.

2014 — The **export demand** for cheese from main US import partners hits an all-time high.

Spatial Analysis

Estimating the quantity of milk requires understanding the product demand in the neighbouring counties in **Utah**, where it will be delivered.

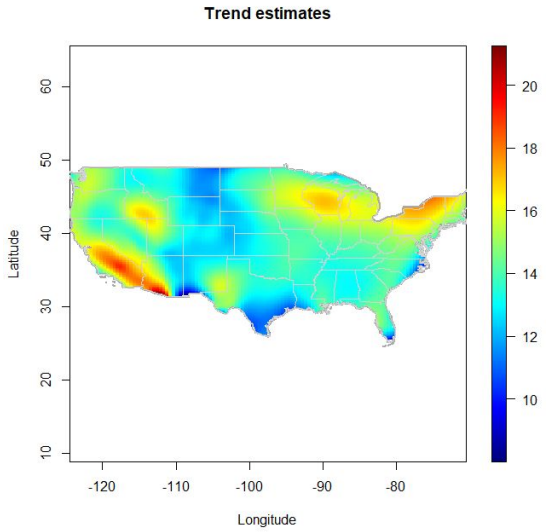


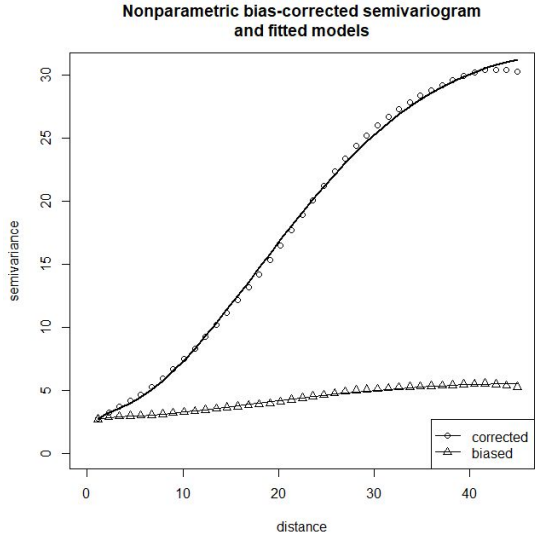
The spatial data collected

- milk+dairy sales [$]
- population

covers only **part of the counties** in the US, while many have **missing values**.

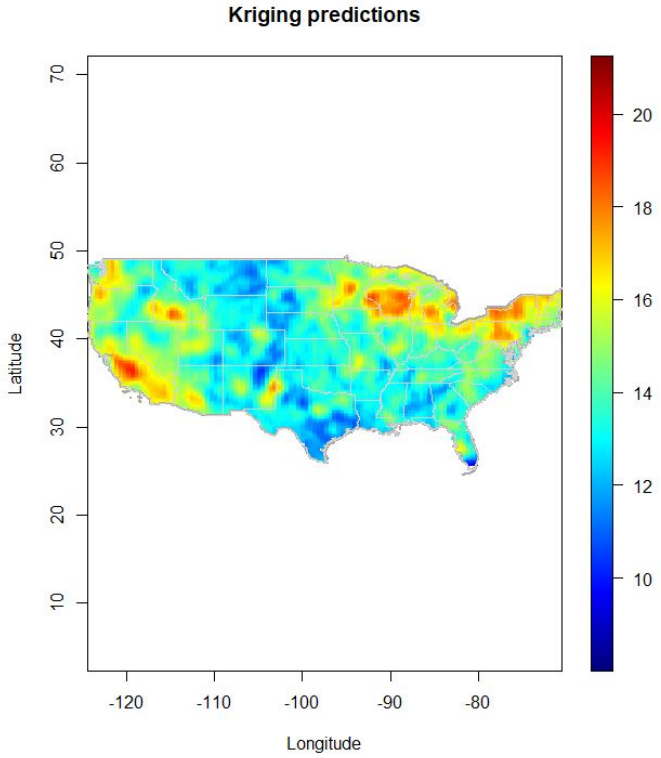
Spatial Analysis

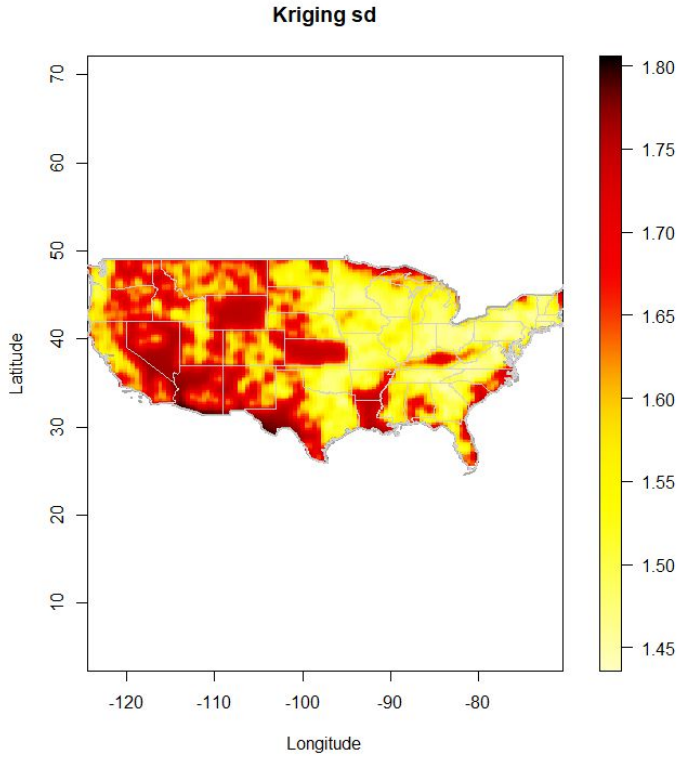




$Y(\mathbf{x}) = \mu(\mathbf{x}) + \epsilon(\mathbf{x})$ is a second-order stationary process.

The **local linear estimator** for $\mu(\cdot)$ at location x is obtained by solving a least squares minimization problem. The semivariogram is estimated nonparametrically as well through a **local linear estimate obtained from residuals.**

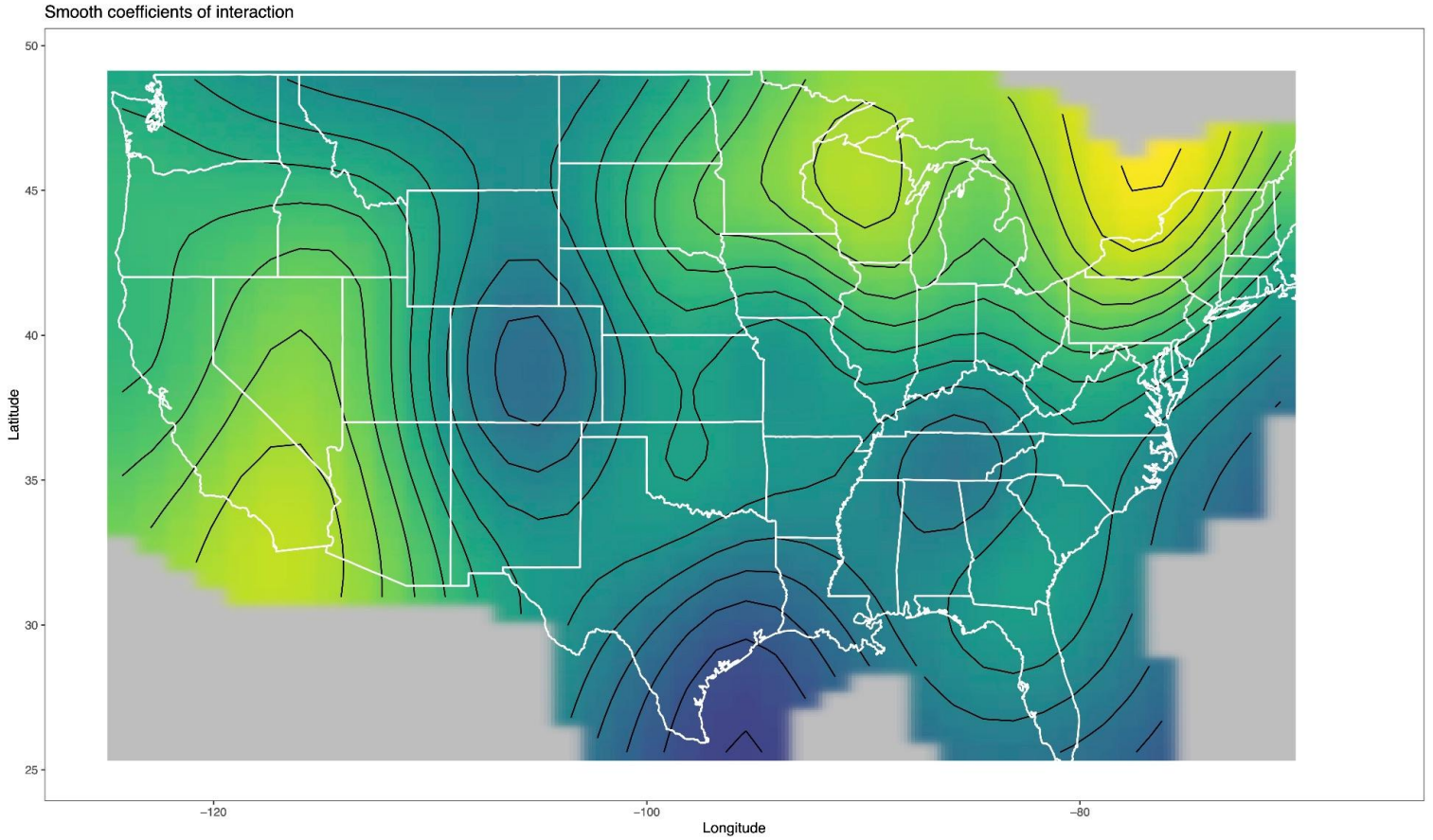


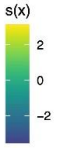


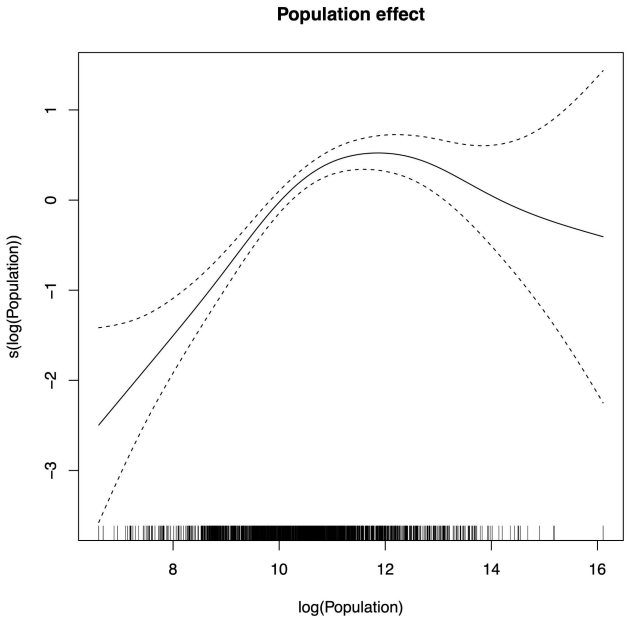
Spatial models with GAM

Fit **additive model** as alternative way of modeling the spatial dependency of dairy and milk sales.
Effect of the **coefficients**:

$$\texttt{log(dairy\_sales)} \sim f(\texttt{latitude} \cdot \texttt{longitude}) + f(\texttt{log(Population)})$$

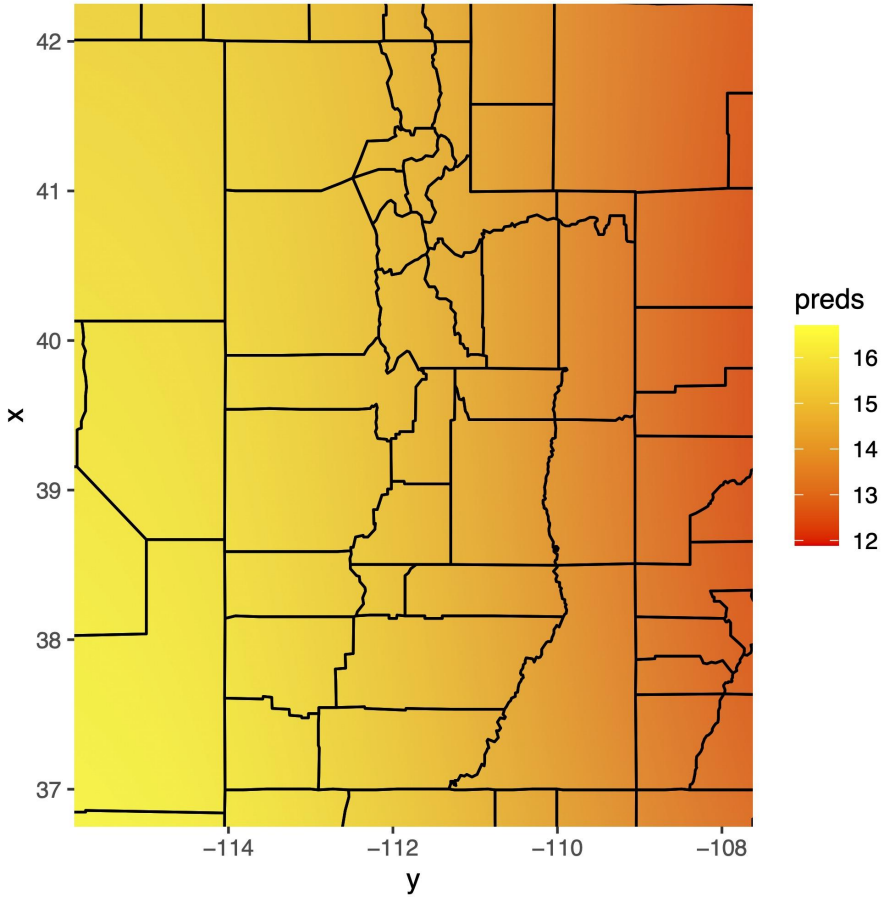


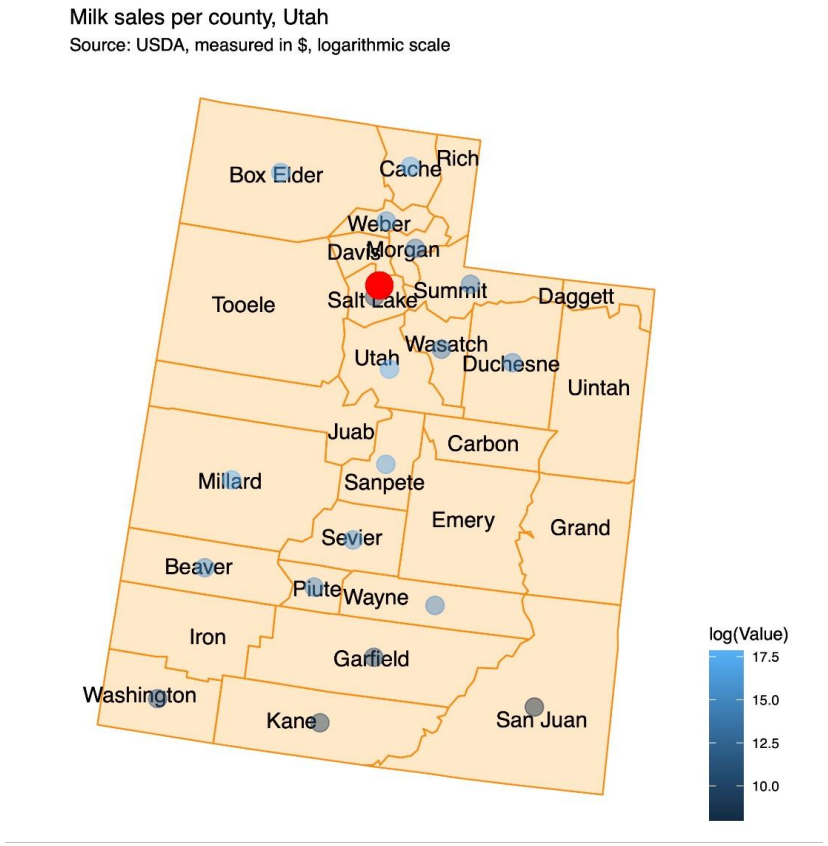




Spatial models with GAM

Use model and population data to get prediction for dairy sales in missing Utah counties.
Both Kriging and GAM predictions suggest focusing deliveries on **Western counties of Utah, specifically North-Western counties.**





Bayesian Clustering

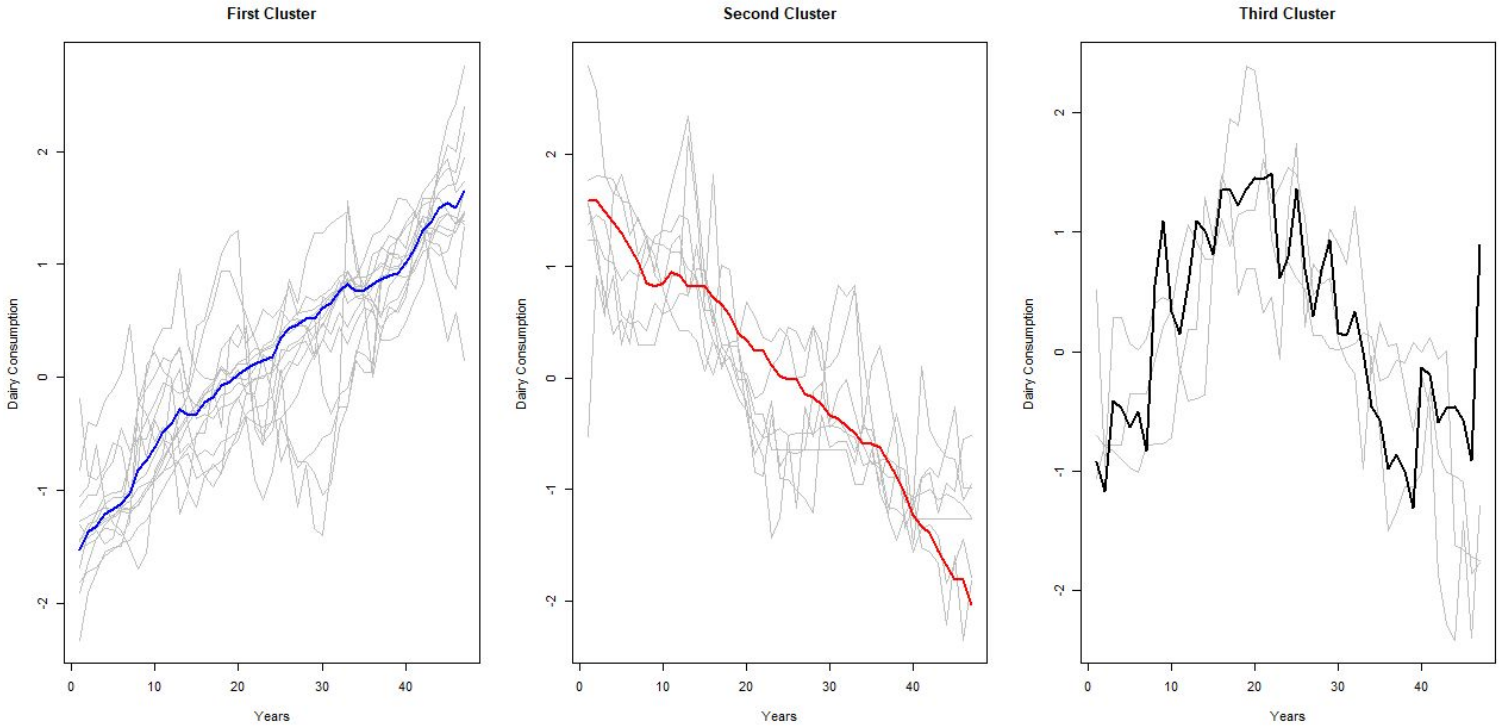
$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\alpha}_i + \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots, n$$

$$\theta_{it} = \rho\theta_{i,t-1} + \nu_{it} \quad \text{with } \nu_{it} \sim \mathcal{N}(0, \sigma_\theta^2)$$

$$\gamma_i = (\boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$$

$$\gamma_i \mid G \overset{\text{iid}}{\sim} G, \quad i = 1, 2, \ldots, n$$

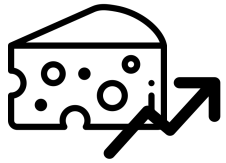
$$G \sim \mathcal{DP}(a, b, G_0)$$



Starting from a **Dirichlet Process** we developed a nonparametric Bayesian clustering.

Through an **outlier detection**, we verified that there were no outliers in the three clusters.

Cluster analysis



Cheese type selection

Produce types of cheese trending in consumption

Conservative strategy

Invest in the production of **upward trending** products identified by nonparametric bayesian clustering:

Cheddar, Mozzarella, Muenster, Cream and Neufchatel, Fluid yogurt, Butter, American cheese, Evaporated and canned bulk and skim milk, Dry buttermilk.

Aggressive strategy

Invest **also** in types of cheese identified in the **third cluster**: cheese who have experienced a downfall in recent years but are beginning a new positive trend: Dry whole milk, Dry whey (milk protein)

Conclusions

Conclusions

Milk pricing

👉 GAM + Bootstrap intervals

Delivery areas

👉 Kriging and GAM for spatial data

Cheese type selection

👉 Bayesian Nonparametric clustering

Further developments

- Consider possible ways of embedding the **import/export** data in our current model
- Find a way to relate dairy sales from the spatial analysis to the optimal quantity of milk to produce (and cows to buy…)

Thank you for the attention!

References

- rfordatascience/tidytuesday. (2019, January 28). GitHub. Retrieved December 12, 2022, from https://github.com/rfordatascience/tidytuesday
- USDA ERS - Dairy Data. (n.d.). USDA ERS - Dairy Data. Retrieved December 12, 2022, from https://www.ers.usda.gov/data-products/dairy-data/
- Bergmann, D., O'Connor, D., & Thümmel, A. (2015). **Seasonal and cyclical behaviour of farm gate milk prices**. *British Food Journal*.
- Nicholson, C. F., & Stephenson, M. W. (2015). **Milk price cycles in the US dairy supply chain and their management implications**. *Agribusiness*, *31*(4), 507-520.
- Nieto-Barajas, L. E., & Contreras-Cristán, A. (2014). **A Bayesian nonparametric approach for time series clustering**. *Bayesian Analysis*, *9*(1), 147-170.

Goals

We identified our main stakeholders in **cheese factories** with one of the following **needs:**

Improvements to the production chain

Identify which dairy products have seen their **consumption increase** and which **decline** over time, and the reasons behind

The aim is to **understand potential improvements** to production chain and **marketing strategy** of declining goods

Optimal price strategy of a new competitor

A **new potential competitor** wants to decide whether to **join the market** or not, the problem is to identify the **potential of its resources**

It’s therefore important to have a clear idea of the **pricing** of goods and make **predictions** about it