A Match Adjusted R-Squared Method for Defining Products within Census Administrative Trade Data

Helen McCulley
Senior Mathematical Statistician
Division of Price Statistical Methods/Branch of International Prices

2023 FCSM Research & Policy Conference
October 26, 2023

International Price Program

- The International Price Program (IPP) of the U.S. Bureau of Labor Statistics produces the Import/Export Price Indexes (MXPI).

- The MXPI are **Principal Federal Economic Indicators,** measuring the changes in prices of nonmilitary goods and services traded between the United States and the rest of the world.

Survey Response Decline

- Response rates to most federal surveys have been declining for many years. The decline accelerated during the COVID-19 pandemic as household and business patterns were greatly interrupted.
- The IPP will soon replace 30-40% (by trade weight) of its directly-collected index pricing data with administrative trade data.

Census Trade Data

- The data comes from trade transactions filed with US Customs and Border Protection and is cleaned and edited by the U.S. Census Bureau for statistical purposes.
- These import and export transaction records, referred to as Census Trade Data (**CTD**), are used by IPP at an aggregate level for both index weighting and survey sampling.
- The primary field for classifying products within CTD is the 10-digit Harmonized System (**HS**) code.
- Transactions include dollar value and quantity.

Matched Model Approach

- Combinations of CTD fields will be used to group transactions into unique **product varieties.**
- The CTD field combination chosen for an area should:
  1. Result in products that are consistently traded over time.
     - More fields can lead to decreased consistency of trade
  2. Result in product varieties with prices that are close to the representative mean prices of the varieties.
     - More fields can decrease the “intra-product” price variability

Index Calculation

- Within homogeneous trade areas, these specific combinations of CTD field values are used to create product varieties whose monthly prices are tracked to calculate Tornqvist unit value indexes at the HS 10-digit level.
- The current system aggregates these unit value indexes to the appropriate import or export price indexes using the IPP's current modified Laspeyres index calculation method.

Census Trade Data Fields

- Recent file layouts showed ~118 fields on import transaction records and ~86 fields on export records
- The HS Code and the Country Code (origin or destination) must always be included in field combinations
- Narrowed down to 12 possible import fields and 8 possible export fields to be used in product variety formation

Import Fields for Product Definition

Description	Key Code
Harmonized Commodity Classification code (HTSUSA number)	H
Country of origin code, Schedule C codes	C
Country of origin subcode (special trade agreements)	B
Manufacturers identification number truncated to 8-digits	M
Foreign port code, Schedule K codes	F
Customs port and district of entry, Schedule D Codes	D
Destination State	S
Importer identification number	I
Ultimate consignee, US party to whom the overseas shipper sold the imported merchandise	U
Related party trade	R
Foreign trade zone	Z
Customs entry type code	T

Export Fields for Product Definition

Description	Key Code
Harmonized Commodity Classification code (HTSUSA number)	H
Country of destination code, Schedule C codes	C
Employer identification number	E
District and port of exportation, Schedule D Codes	D
U.S. State of Origin	S
Zip code	Z
Domestic or foreign status of merchandise	F
Related party trade	R

Possible Import Field Combinations

- Must contain H (HS code) and C (country of origin)

- 10 additional fields including the null set gives $\sum_{0 \leq n \leq 10} \binom{10}{n} = 2^{10} = 1024$ combinations

$$K = \{k_1, k_2, k_3 \ldots, k_{1024}\} = \{\{HC\}, \{HCM\}, \{HCS\}, \ldots, \{HCMSBDPIGRTU\}\}$$

Field Combination Selection

- The combination of fields used to form product varieties are determined at the BEA 5-digit index level.
- Early research
  - Short, medium, and long field combinations - compared index results
  - Linear regression
  - Predictor screening (bootstrap forest partitioning)
  - Variable importance (Boruta method)
- Match-adjusted R-squared (MARS) - addresses both product continuance and adequate product definition

Match-adjusted R-squared

- Antonio G. Chessa's work with scanner data use in the Dutch CPI
- The **MARS** score is the product of μ, the degree of product match with the base period, and R, a measure of product homogeneity.
- As μ increases (choosing less detailed keys) R tends to decrease
- As R increases (choosing more detailed keys) μ tends to decrease

Just a Little Notation

- C_t is the set of all CTD records i in time t
- Individual transaction record's quantity and value are q_i and v_i
- The set of all possible key combinations for imports is $K=\{k_j | \ 1 \leq j \leq 1024\}$.
- The set of product varieties created by grouping all transactions with the same field values for all included key fields as defined by k_j is $\{i_{k_j}\}$.
- Each of these product varieties will have the summed value $v_{i_{k_j}}$ and the summed quantity $q_{i_{k_j}}$ across all transactions as defined by the k_j grouping.
- $B_{0,t}$ is the set of products that are imported both in the base period and in period t.

Degree of Product Match

For month t and key combination k_j,

$$\mu_{k_j,t} = \frac{\sum_{i_{k_j} \in B_{0,t}} q_{i_{k_j},t}}{\sum_{i \in C_t} q_{i,t}} \Rightarrow$$

$$\frac{\text{Sum of product quantities at time t if products are also in the base period}}{\text{Sum of quantities at time t}}$$

for a particular item defining key combination

Degree of Product Match

- Bounded by 0 and 1
- Less detailed item keys (fewer included CTD fields) result in higher degree of product match
- More detailed item keys result in lower degree of product match
- Base period dependent

Product Homogeneity

$$R_{k_j,t} = \frac{\sum_{i_{k_j} \in B_t} q_{i_{k_j},t} \left(\frac{v_{i_{k_j},t}}{q_{i_{k_j},t}} - \frac{v_t}{q_t} \right)^2}{\sum_{i \in C_t} q_{i,t} \left(\frac{v_{i,t}}{q_{i,t}} - \frac{v_t}{q_t} \right)^2}$$

$R_{k_j,t}$ is the proportion of explained variance in product prices, relative to the variance of transaction unit prices.

Product Homogeneity

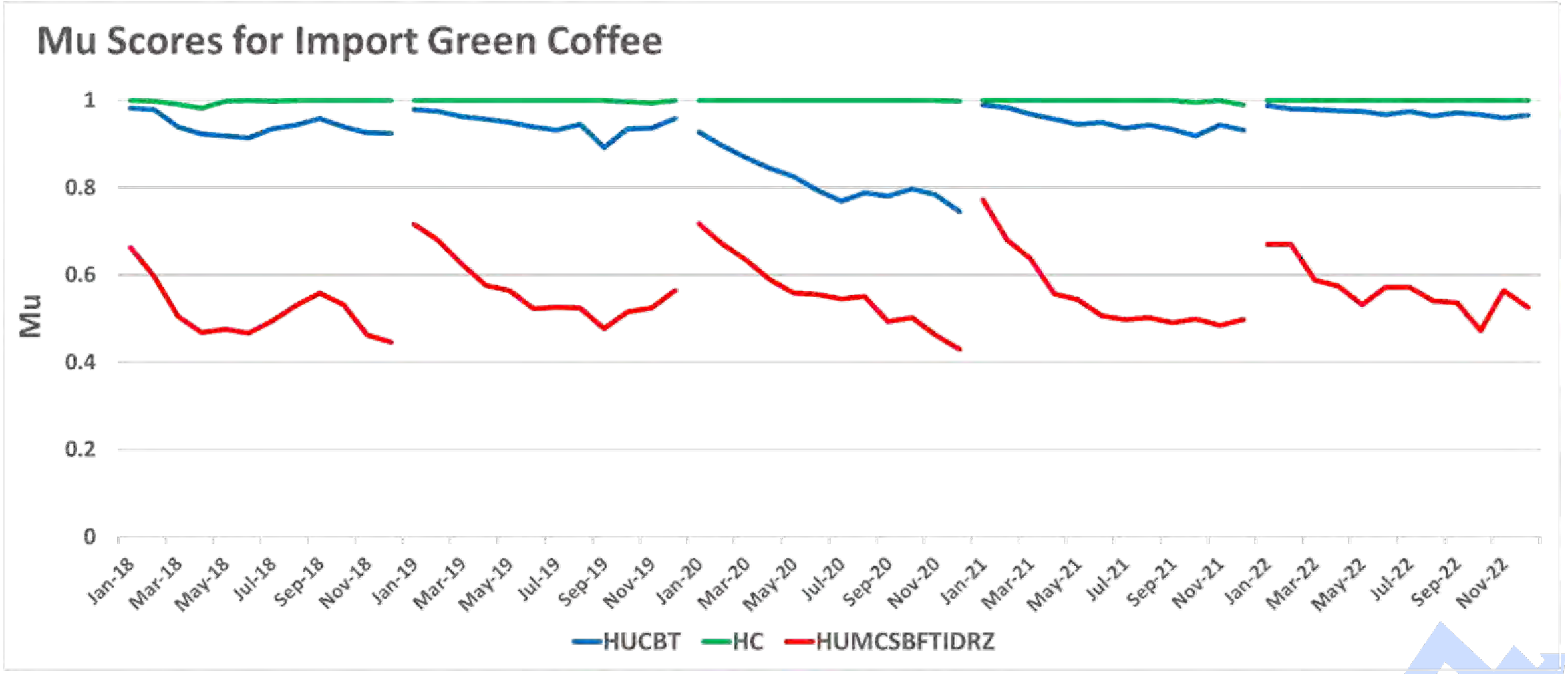
- Base period is irrelevant
- Also bounded by 0 and 1
- More detailed keys have higher R values
- Less detailed keys have lower R values

Match-Adjusted R^2 - MARS

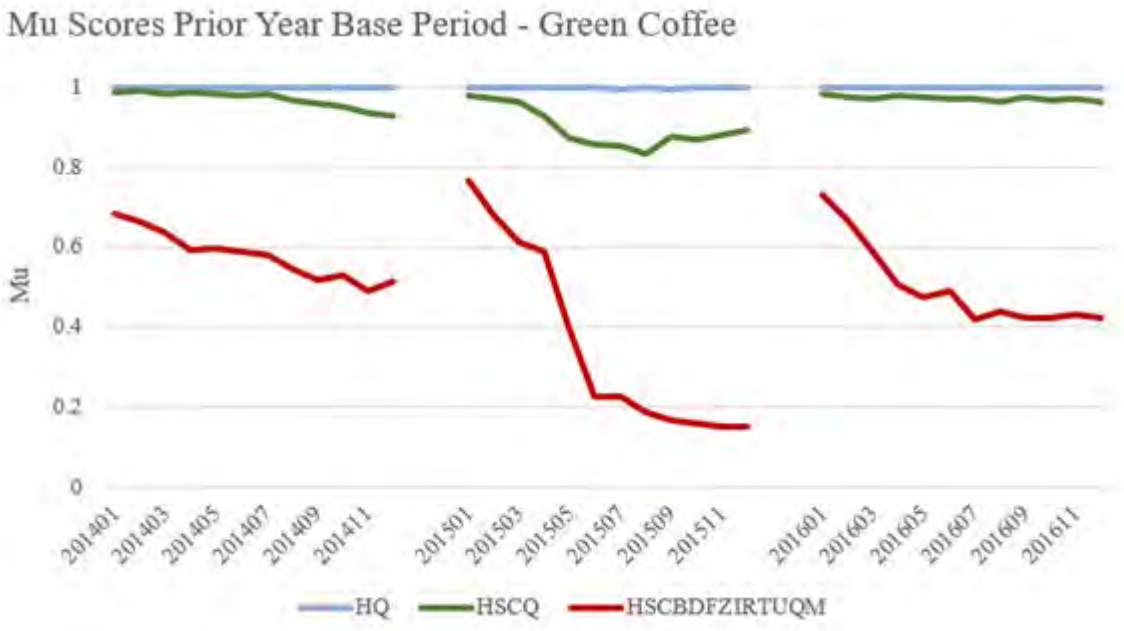
$$M_{k_j,t} = \mu_{k_j,t} R_{k_j,t}$$

- Bounded by 0 and 1
- Base period dependent
- Allows the selection of a key combination, corresponding to the largest value of *M*, that works to both minimize product churn and maximize homogeneity.

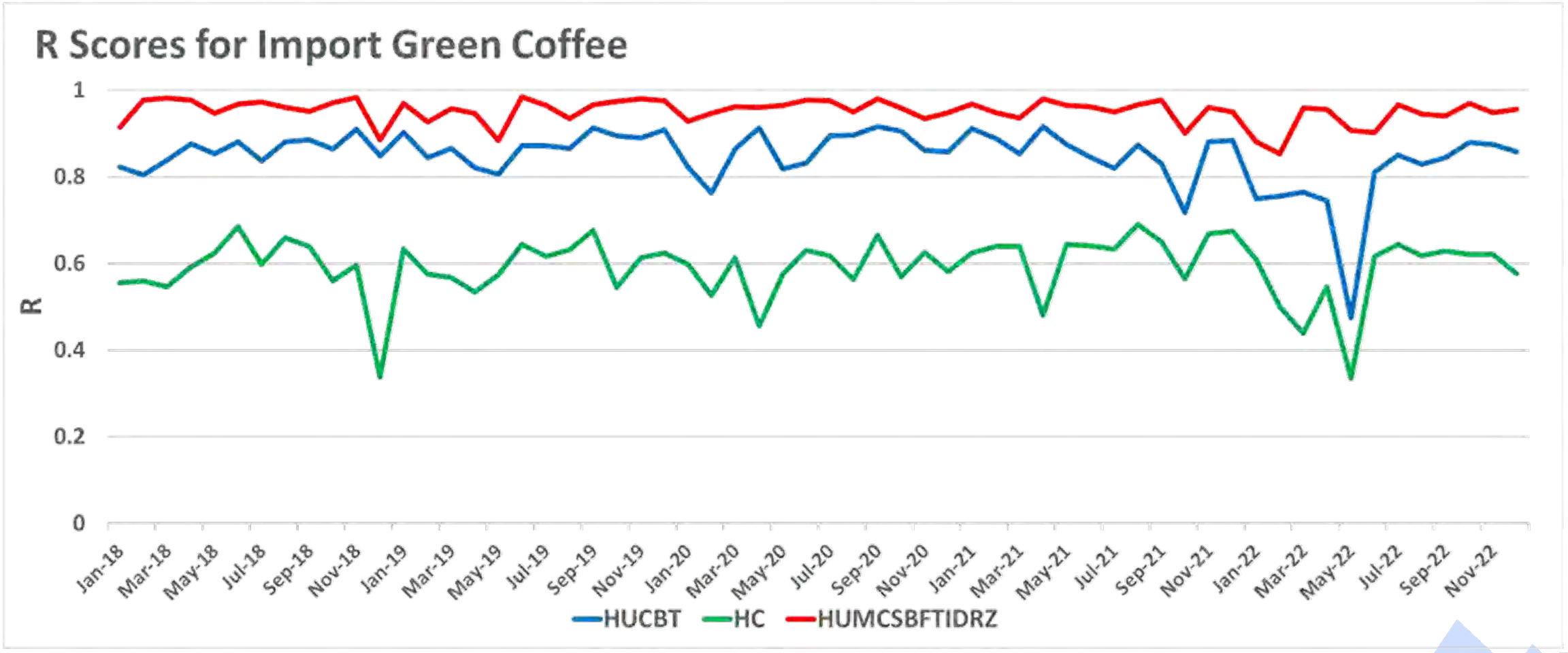
Examples – Import Green Coffee



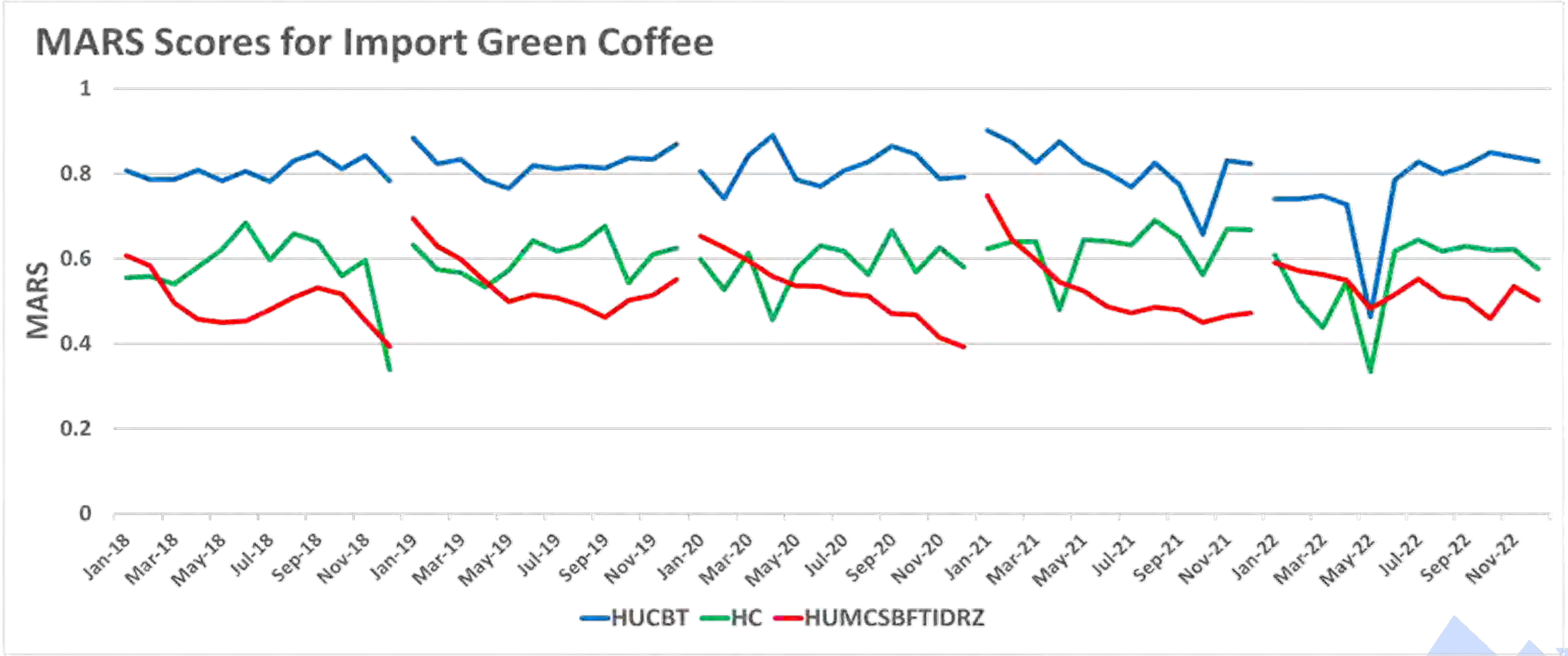
Base Period Dependency of Mu



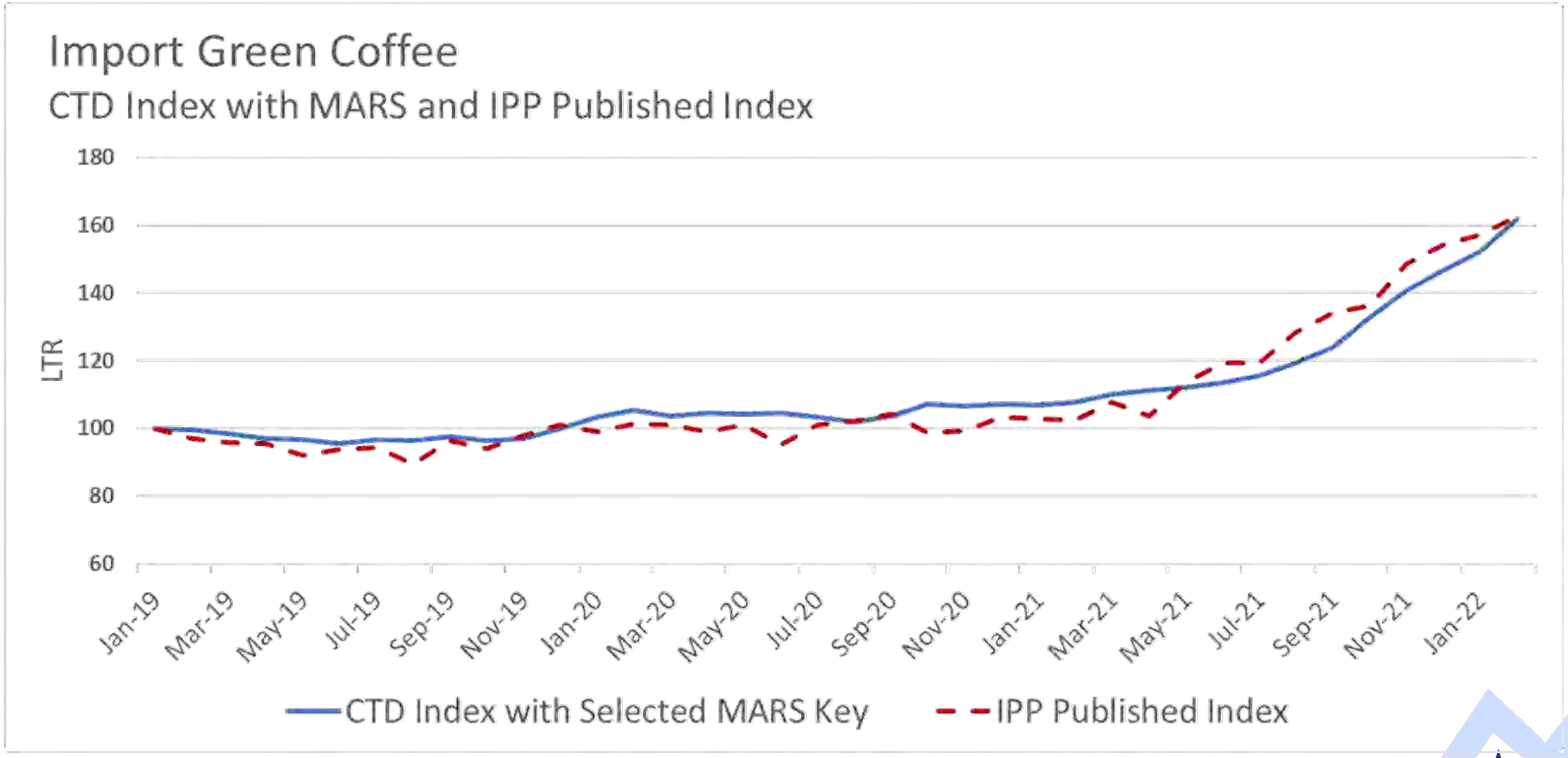
Examples – Import Green Coffee



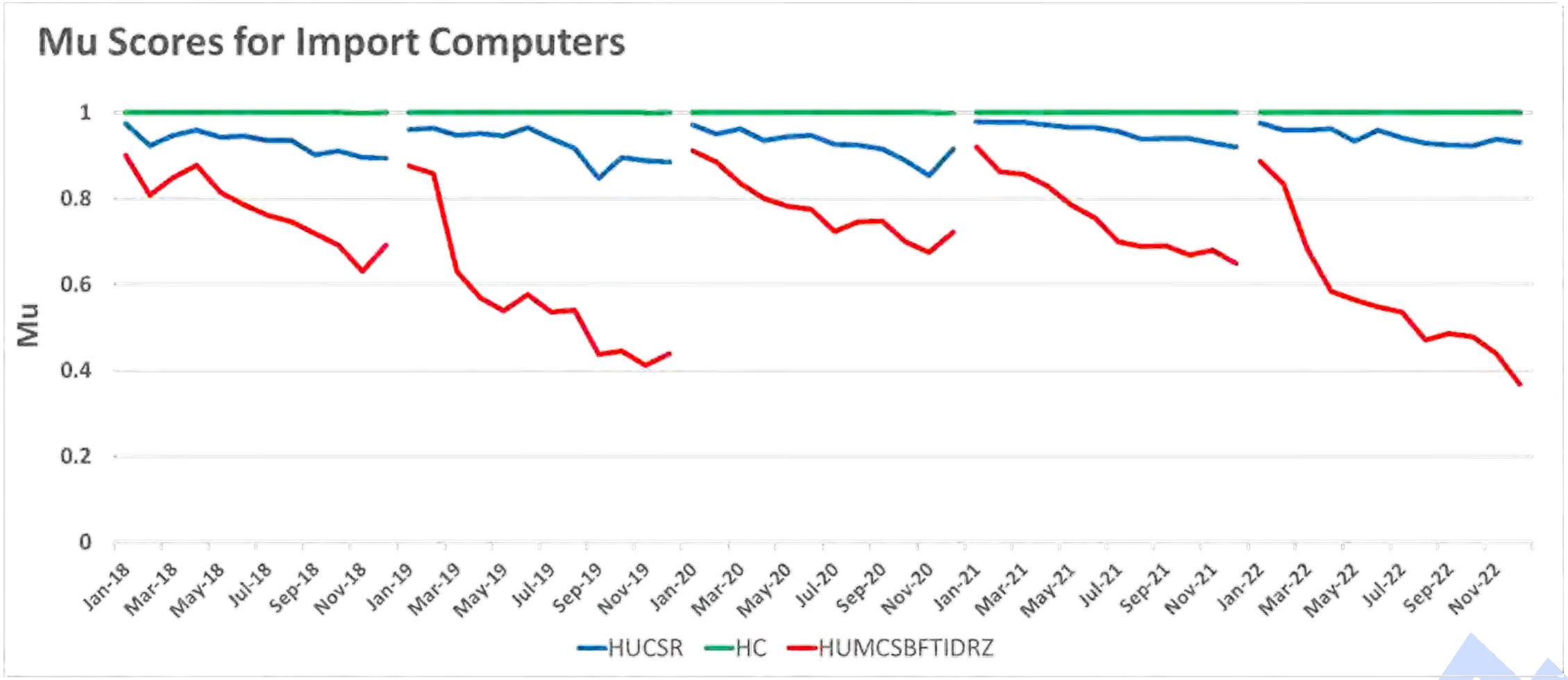
Examples – Import Green Coffee



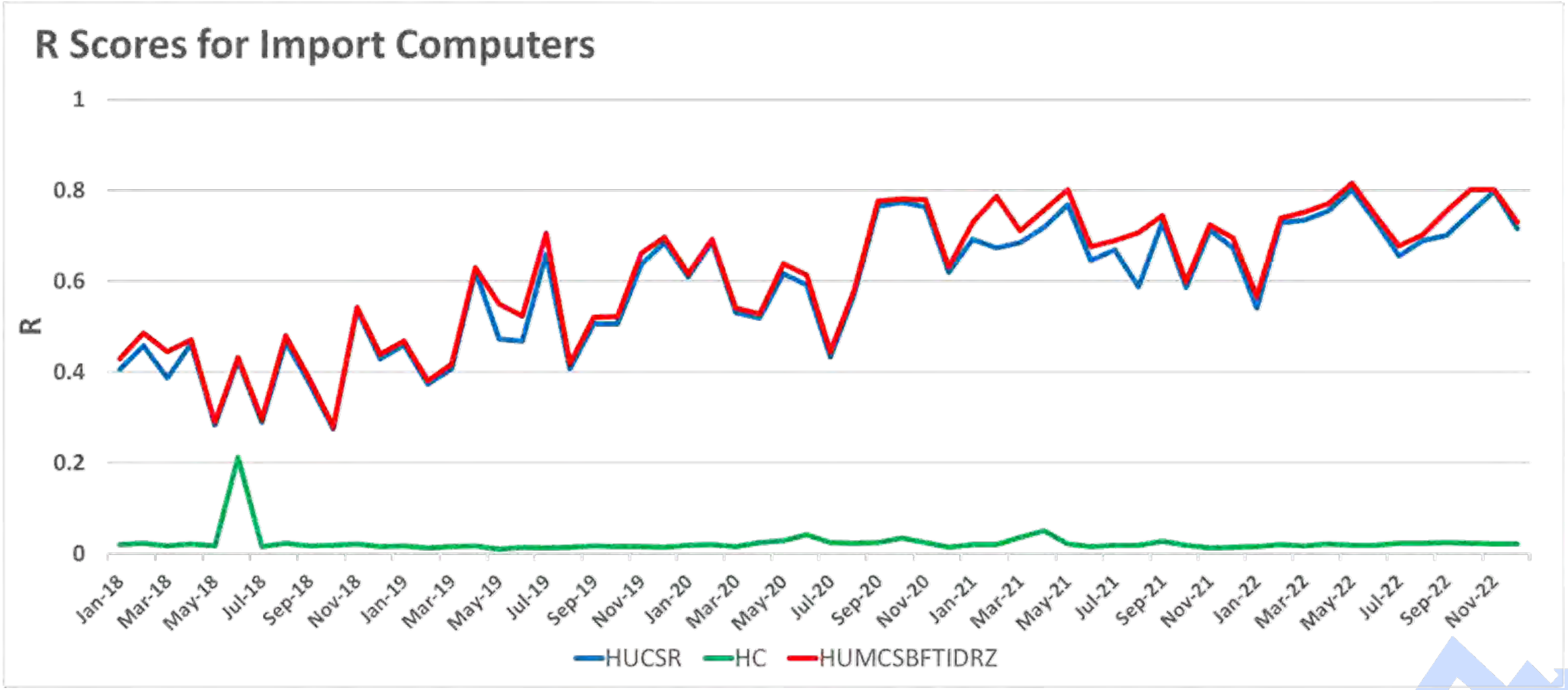
Examples – Import Green Coffee



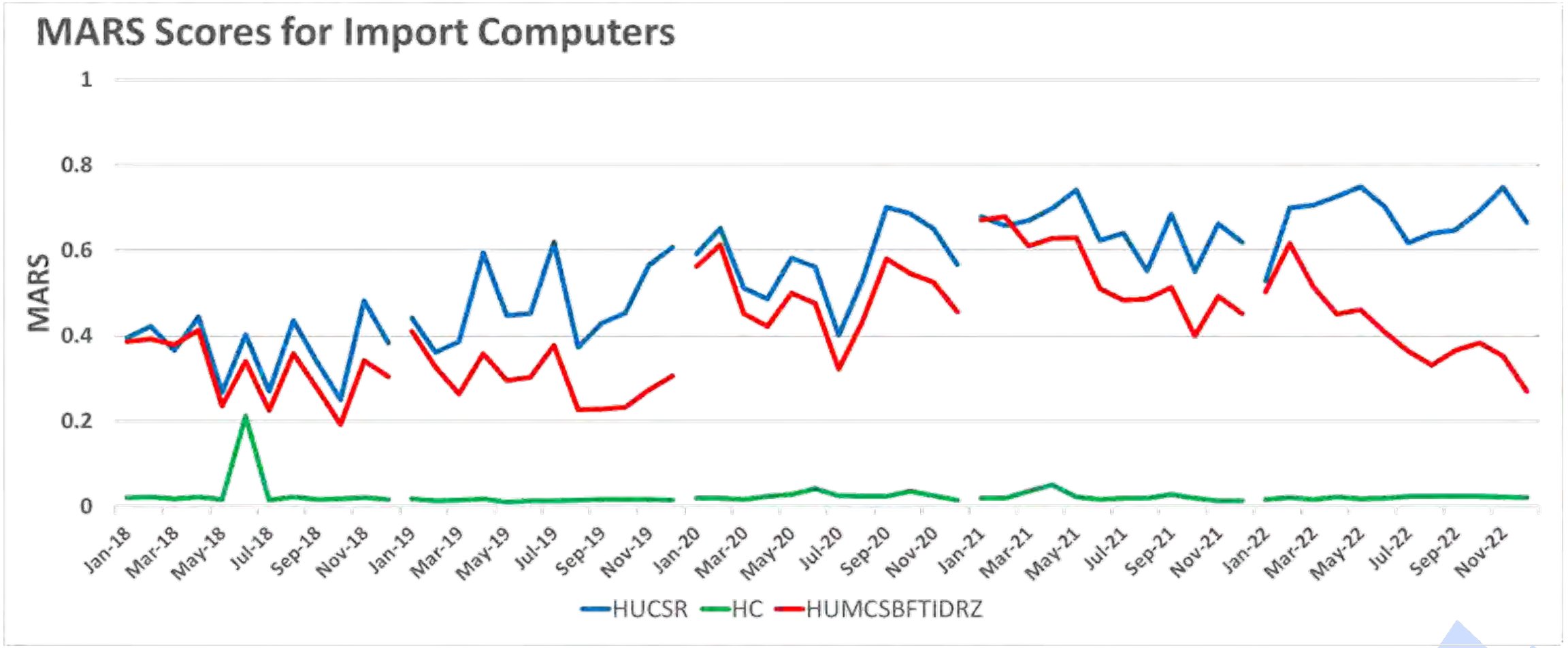
Examples – Import Computers



Examples – Import Computers

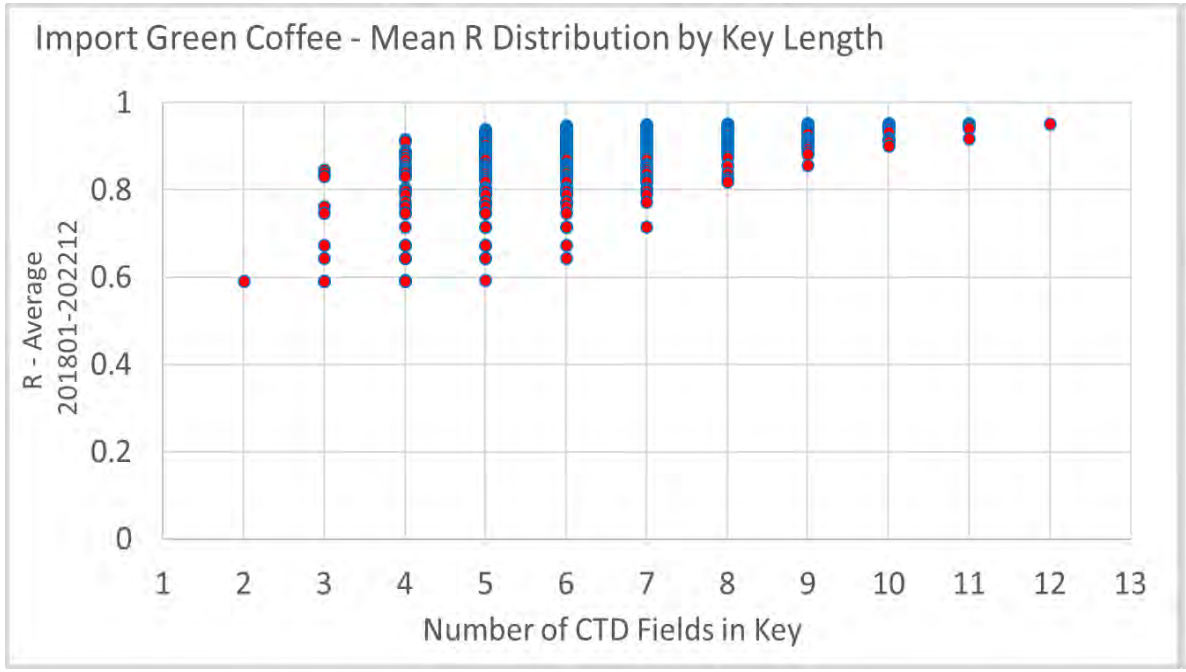


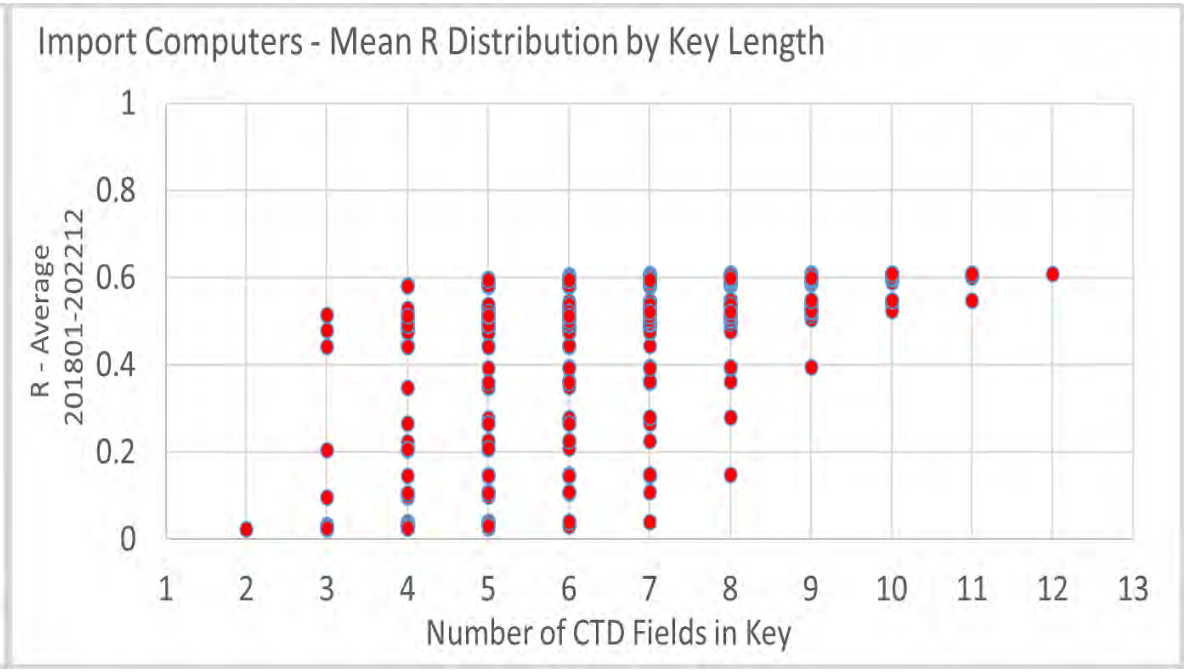
Examples – Import Computers



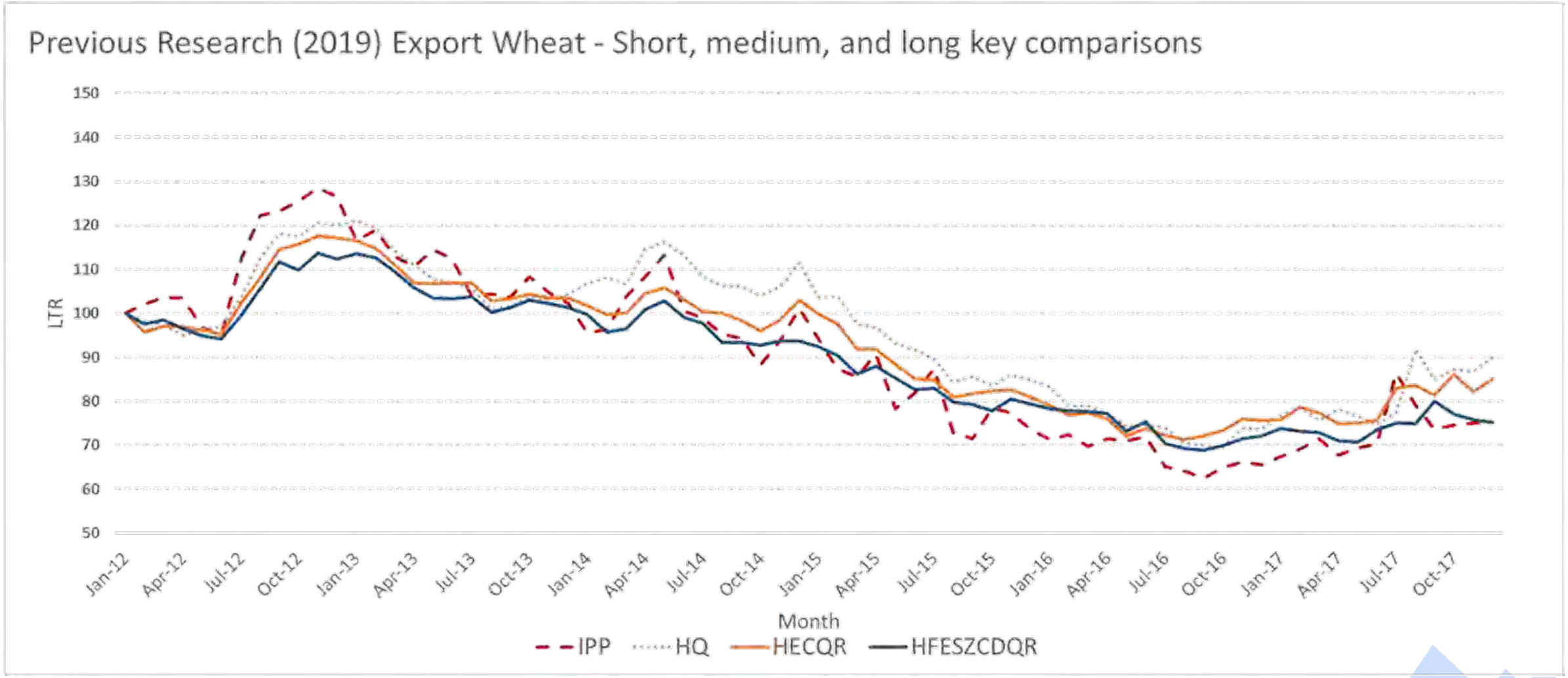
Distribution of R

Homogeneous and Nonhomogeneous Areas

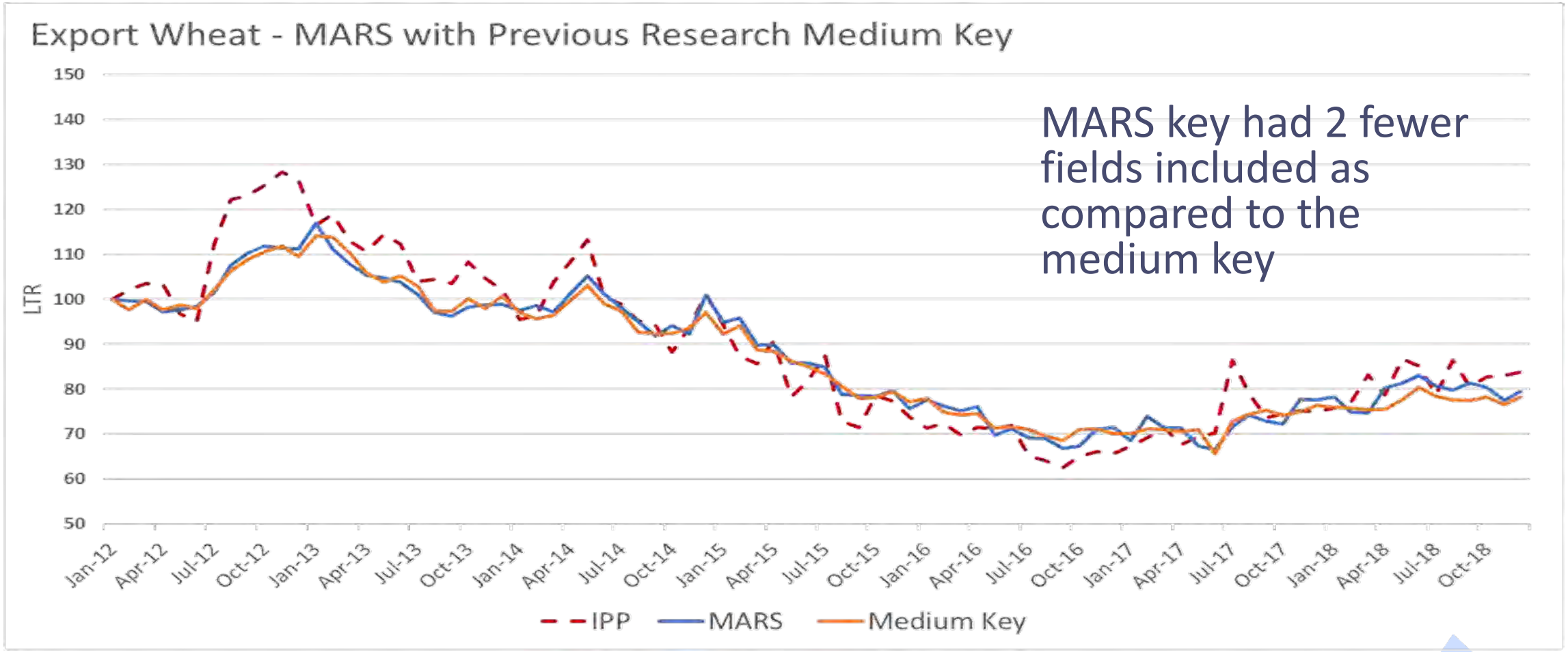




Export Wheat – Short, Medium, Long Keys



Export Wheat – MARS Key and Medium Key



Additional Resources

Department of Labor, Bureau of Labor Statistics. “Comment Request.” Federal Register 88, no. 174 (Sept 11, 2023): 62402 https://www.govinfo.gov/content/pkg/FR-2023-09-11/pdf/2023-19486.pdf.

Chessa, A.G. (2019). MARS: A Method for Defining Products and Linking Barcodes of Item Relaunches. Paper presented at the 16th Meeting of the Ottawa Group on Price Indices, 8-10 May 2019, Rio de Janeiro, Brazil. Available at https://eventos.fgv.br/sites/eventos.fgv.br/files/arquivos/u161/product_definition_with_mars_chessa_og19.pdf

Contact Information

Helen McCulley
Senior Mathematical Statistician
Division of Price Statistical Methods/Branch of International Prices
www.bls.gov/mxp
202-691-6907
mcculley.helen@bls.gov