# Deriving a Data-Driven Acreage and Yield Imputation Method for Suppressed MASC Yield Data

# 1. Background

This report documents the evolution of a reproducible, year‑agnostic framework for estimating missing acreage and yield data in the Manitoba Agricultural Services Corporation (MASC) Variety Browser. The analysis focuses on reconstructing suppressed records for crop–variety–municipality combinations that fall below MASC’s disclosure thresholds (fewer than three farms or less than 500 acres). What began as an acreage‑only model has since been refined into a paired acreage‑and‑yield imputation system that maintains fidelity to official provincial totals while generalizing across crop years.

# 2. Source Data

Two main inputs form the basis of this work:

• masc\_df — the full merged dataset containing both valid and suppressed records.

• masc\_summary.csv — the provincial annual summary containing:

year records varieties total\_acres yield\_tonnes\_per\_acre

The summary file provides the reference total acreage and average yield per acre for each year. For example, 2024 reported 9,945,150.2 harvested acres with an average yield of 2.039 tonnes per acre. These serve as ground truth constraints for reconstruction.

# 3. Early Attempts and Dead Ends

Initial approaches treated the missing entries as random gaps and explored various allocation methods. Correlation analysis between reported acres and missing variety counts per municipality showed only moderate relationships, and RM‑level balancing approaches produced implausible per‑variety acreages (often exceeding 2,000 acres). Attempts to weight by number of null varieties or by crop percentage of total area over‑emphasized high‑acreage crops such as Red Spring Wheat and Canola, resulting in exaggerated totals. These experiments highlighted the need for a farm‑based unit of measure and proportional weighting rooted in real observed farm‑level behaviour.

# 4. Establishing a Farm‑Level Basis

To ground the model in physical scale, the valid portion of the dataset was grouped by crop to calculate:

avg\_acres\_per\_farm = acres / farms  
avg\_yield\_per\_farm = yield\_tonnes / farms

These averages describe how much land and production a typical farm contributes for each crop. They became the foundation for scaling missing entries, replacing arbitrary percentage‑based weights with empirically derived farm‑scale behaviour. Crops with fewer than 10,000 reported acres were excluded from direct weighting to ensure stability.

# 5. Transition to Stable and Proxy Mapping

While reconstructing acreage, it became clear that not all crop names appear each year. To make the process reproducible across multiple years (2017–2024), two classification layers were introduced:

• Stable crops — crops with at least 10,000 reported acres in valid\_df. Their averages are reliable reference anchors.  
• proxy\_map — a dictionary mapping rare or missing crop names to their most agronomically similar stable crop (for example, mapping 'POLISH CANOLA' or 'RAPESEED' to 'ARGENTINE CANOLA').

When a crop in missing\_df is not found in valid\_df, the proxy\_map substitutes it with the appropriate stable reference. If the crop remains unmatched, it inherits the median average from all stable crops. This ensures that no crop, regardless of naming differences or rarity, is left without an imputation basis.

# 6. Acreage Imputation Method

Each missing crop–variety record is assumed to represent one farm. Using average acres per farm for each crop, the unreported provincial acreage (acres\_diff) is distributed proportionally according to:

weight = avg\_acres\_per\_farm \* n\_missing  
weight\_norm = weight / weight.sum()  
crop\_imputed\_total = weight\_norm \* acres\_diff  
imputed\_acres = crop\_imputed\_total / n\_missing

This approach scales the missing acreage according to both the average farm size and the number of missing varieties within each crop. The final acreage allocation matched the exact provincial shortfall:

Acreage check: 2122848.799999999 expected: 2122848.799999999

# 7. Extending the Framework to Yield

To complete the production picture, yield imputation was added using the same proportional logic. The masc\_summary.csv provides an average yield per acre for the year, from which the total provincial yield (yield\_tonnes\_per\_acre × total\_acres) was calculated. The difference between this total and the summed valid yield from masc\_df is yield\_diff:

yield\_diff = (summary\_yield\_sum - masc\_df['yield\_tonnes'].sum())

Average yield per farm (avg\_yield\_per\_farm) was used in place of avg\_acres\_per\_farm to weight the missing crops:

weight = avg\_yield\_per\_farm \* n\_missing  
weight\_norm = weight / weight.sum()  
crop\_imputed\_total\_yield = weight\_norm \* yield\_diff  
imputed\_yield\_tonnes = crop\_imputed\_total\_yield / n\_missing

This produced an exact provincial reconciliation:

Yield check: 6163074.6625 expected: 6163074.6625000015

As with acreage, each missing record’s yield estimate is tied directly to empirical farm productivity patterns rather than arbitrary distribution.

# 8. Integration and Final Dataset

After acreage and yield imputations, valid\_df and missing\_df were merged into a unified dataset with an indicator column:

valid\_df['imputed'] = 0  
missing\_df['imputed'] = 1

Each missing entry was assigned farms = 1, acres = imputed\_acres, and yield\_tonnes = imputed\_yield\_tonnes. These fields replicate the structure of valid\_df, producing a complete table of 8,761 entries for 2024 with no missing acreage or yield.

# 9. Key Insights and Generalization

The most significant realization was that yield could be imputed using the same structure as acreage once it was expressed on a per‑farm basis. This symmetry allowed both sides of agricultural production—land use and output—to be balanced simultaneously against provincial totals. The final framework is reproducible across years and insensitive to variations in naming conventions or reporting thresholds. By anchoring all calculations in farm‑level averages, it provides consistent year‑to‑year scaling even as crop mixes evolve.

# 10. Conclusion

This dual‑imputation model replaces ad‑hoc manual adjustments with a reproducible, data‑driven procedure. It reconciles reported and unreported acreage and yield for all crops, leveraging valid\_df’s internal structure and the proxy system for unmatched crops. Each missing record is conceptually one farm, scaled by empirically derived averages that preserve provincial totals. The resulting dataset aligns with MASC’s published summaries and is ready for integration into downstream biomass, yield‑mapping, and economic analyses.

# 11. Appendix

# stable crop anchors (never remapped)

stable\_crops = [

"ARGENTINE CANOLA", "RED SPRING WHEAT", "SOYBEANS",

"FIELD PEAS", "BARLEY", "OATS", "GRAIN CORN",

"PASTURE (TAME/PERENNIAL)", "COARSE HAY"

]

# universal proxy map covering all MASC-listed crops

proxy\_map = {

# oilseeds

"POLISH CANOLA": "ARGENTINE CANOLA",

"RAPESEED": "ARGENTINE CANOLA",

"OIL SUNFLOWERS": "ARGENTINE CANOLA",

"NON-OIL SUNFLOWERS": "ARGENTINE CANOLA",

"HEMP GRAIN": "SOYBEANS",

"ORGANIC HEMP GRAIN": "SOYBEANS",

"MUSTARD": "ARGENTINE CANOLA",

"FLAX": "ARGENTINE CANOLA",

"ORGANIC FLAX": "ARGENTINE CANOLA",

"PHACELIA": "ARGENTINE CANOLA",

# cereals and small grains

"RED SPRING WHEAT": "RED SPRING WHEAT",

"NORTH. HARD RED WHT": "RED SPRING WHEAT",

"PRAIRIE SPRING WHEAT": "RED SPRING WHEAT",

"OTHER SPRING WHEAT": "RED SPRING WHEAT",

"DURUM WHEAT": "RED SPRING WHEAT",

"HARD WHITE WHEAT": "RED SPRING WHEAT",

"EXTRA STRONG WHEAT": "RED SPRING WHEAT",

"WINTER WHEAT": "RED SPRING WHEAT",

"WINTER TRITICALE": "RED SPRING WHEAT",

"TRITICALE": "RED SPRING WHEAT",

"EMMER WHEAT": "RED SPRING WHEAT",

"SPELT": "RED SPRING WHEAT",

"ORGANIC DURUM WHEAT": "RED SPRING WHEAT",

"ORGANIC E.S. WHEAT": "RED SPRING WHEAT",

"ORGANIC H.W. WHEAT": "RED SPRING WHEAT",

"ORGANIC N.H.R. WHT": "RED SPRING WHEAT",

"ORGANIC P.S. WHEAT": "RED SPRING WHEAT",

"ORGANIC R.S. WHEAT": "RED SPRING WHEAT",

"ORGANIC SPR WHT OTHR": "RED SPRING WHEAT",

"ORGANIC WINTER WHEAT": "RED SPRING WHEAT",

"OPEN POL. FALL RYE": "RED SPRING WHEAT",

"HYBRID FALL RYE": "RED SPRING WHEAT",

"ORGANIC O P FALL RYE": "RED SPRING WHEAT",

"OPEN POL SILAGE CORN": "GRAIN CORN",

"OPEN POLLINATED CORN": "GRAIN CORN",

"ORGANIC BARLEY": "BARLEY",

"ORGANIC OATS": "OATS",

"BARLEY": "BARLEY",

"BUCKWHEAT": "OATS",

"OATS": "OATS",

"MIXED GRAIN": "BARLEY",

"PROSO MILLET": "OATS",

"MILLET (FOXTAIL SEED)": "OATS",

"MILO/GRAIN SORGHUM": "GRAIN CORN",

# pulses and legumes

"SOYBEANS": "SOYBEANS",

"PINTO BEANS": "SOYBEANS",

"BLACK BEANS": "SOYBEANS",

"WHITE PEA BEANS": "SOYBEANS",

"SMALL RED BEANS": "SOYBEANS",

"CRANBERRY BEANS": "SOYBEANS",

"KIDNEY BEANS": "SOYBEANS",

"OTH DRY EDIBLE BEANS": "SOYBEANS",

"ADZUKI BEANS": "SOYBEANS",

"FABABEANS": "FIELD PEAS",

"FIELD PEAS": "FIELD PEAS",

"LENTILS": "FIELD PEAS",

"ORGANIC FIELD PEAS": "FIELD PEAS",

# root and vegetable crops

"PROC POTATOES-IRRIG": "FIELD PEAS",

"PROC POTATOES-DRYLND": "FIELD PEAS",

"TABLE POTATOES": "FIELD PEAS",

"SUGAR BEETS": "FIELD PEAS",

"CARROTS": "FIELD PEAS",

"CAULIFLOWER": "FIELD PEAS",

"ASPARAGUS": "FIELD PEAS",

"CUCUMBERS": "FIELD PEAS",

"SWEET CORN": "FIELD PEAS",

"RED BEET": "FIELD PEAS",

"RUTABAGAS": "FIELD PEAS",

"COOKING ONIONS": "FIELD PEAS",

"SWEET POTATO": "FIELD PEAS",

"CABBAGE": "FIELD PEAS",

"PARSNIPS": "FIELD PEAS",

"QUINOA": "FIELD PEAS",

# forage, hay, pasture

"ALFALFA": "COARSE HAY",

"ALFALFA/GRASS MIX.": "COARSE HAY",

"ALFALFA EST OPTION": "COARSE HAY",

"ALFALFA GRASS EST OP": "COARSE HAY",

"ALFALFA SEED EST OPT": "COARSE HAY",

"COMMON ALFALFA SEED": "COARSE HAY",

"PED. ALFALFA SEED": "COARSE HAY",

"COARSE HAY": "COARSE HAY",

"GREENFEED": "COARSE HAY",

"FORAGE ESTABLISHMENT": "COARSE HAY",

"CLOVER & BROMEGRASS HAY": "COARSE HAY",

"CLOVER RD CNRYG & TIM HAY": "COARSE HAY",

"INTERCROP MIXTURE": "COARSE HAY",

"SWEET CLOVER": "COARSE HAY",

"SWEET CLOVER (SEED)": "COARSE HAY",

"SWEET CLOVER EST OPT": "COARSE HAY",

"SAINFOIN (FORAGE)": "COARSE HAY",

"CICER MILKVETCH (FORAGE)": "COARSE HAY",

"HAIRY VETCH": "COARSE HAY",

# grass and seed crops

"GRASSES": "PASTURE (TAME/PERENNIAL)",

"GRASSES (SEED)": "PASTURE (TAME/PERENNIAL)",

"GRASSES (SOD)": "PASTURE (TAME/PERENNIAL)",

"PASTURE (TAME/ANNUAL)": "PASTURE (TAME/PERENNIAL)",

"PASTURE (TAME/PERENNIAL)": "PASTURE (TAME/PERENNIAL)",

"PER. RYEGRASS SEED": "PASTURE (TAME/PERENNIAL)",

"ANNUAL RYEGRASS SEED": "PASTURE (TAME/PERENNIAL)",

"TALL FESCUE SEED": "PASTURE (TAME/PERENNIAL)",

"TIMOTHY SEED COMMON": "PASTURE (TAME/PERENNIAL)",

"PED. TIMOTHY SEED": "PASTURE (TAME/PERENNIAL)",

"ALSIKE CLOVER (FORAGE)": "PASTURE (TAME/PERENNIAL)",

"ALSIKE CLOVER (SEED)": "PASTURE (TAME/PERENNIAL)",

"BIRDSFOOT TREFOIL (SEED)": "PASTURE (TAME/PERENNIAL)",

"BIRDSFOOT TREFOIL FORAGE": "PASTURE (TAME/PERENNIAL)",

"FALL GRASS EST OPTN": "PASTURE (TAME/PERENNIAL)",

"FALL TIMOTHY EST OPT": "PASTURE (TAME/PERENNIAL)",

"SPRING GRASS EST OPT": "PASTURE (TAME/PERENNIAL)",

"SPRING TIMOTHY EST O": "PASTURE (TAME/PERENNIAL)",

"LUPINS (FORAGE)": "PASTURE (TAME/PERENNIAL)",

"LUPINS (SEED)": "PASTURE (TAME/PERENNIAL)",

"CARAWAY": "PASTURE (TAME/PERENNIAL)",

"CANARYSEED": "PASTURE (TAME/PERENNIAL)",

# silage, corn, and miscellaneous

"SILAGE CORN": "GRAIN CORN",

"NON-CORN SILAGE": "GRAIN CORN",

"TOO WET TO SEED": "COARSE HAY",

"UNUSED LAND": "COARSE HAY",

"BUSH/YARD/SLOUGH": "COARSE HAY",

}