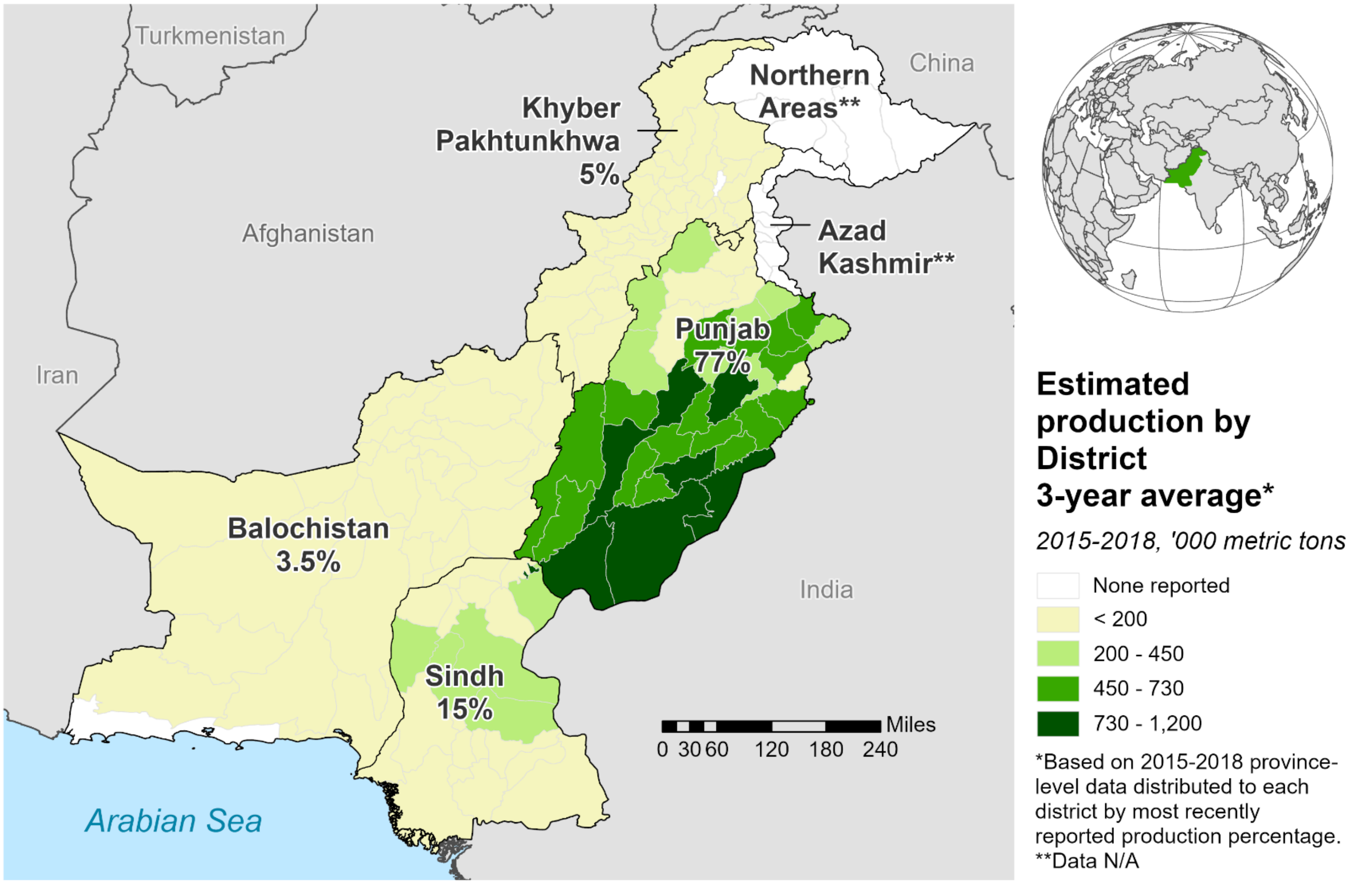
Engineering a Wheat Yield Feature from Historical Data



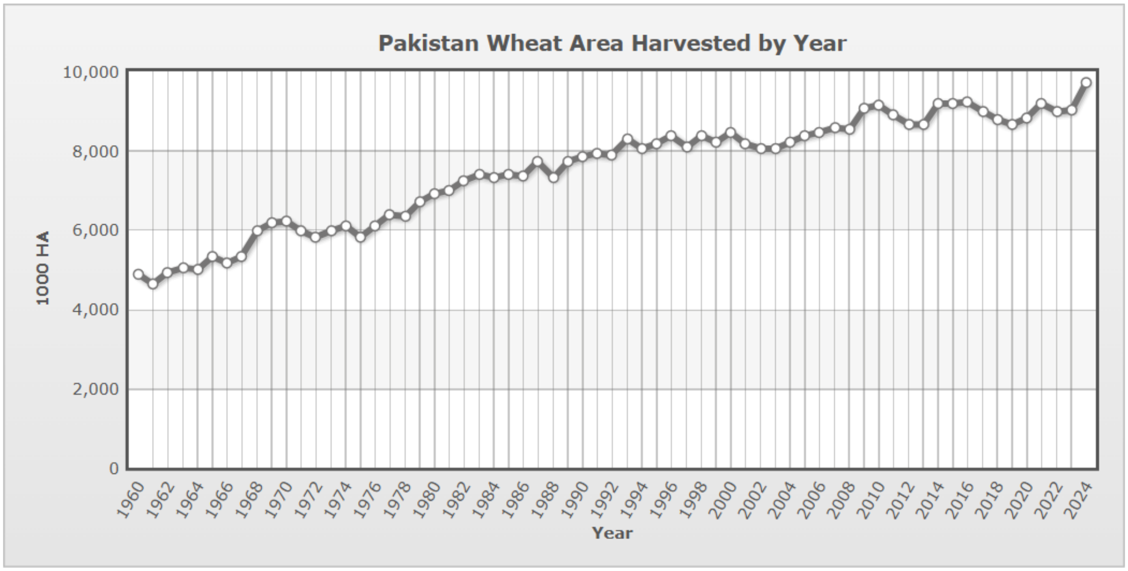
*Based on 2015-2018 province-level data distributed to each district by most recently reported production percentage.
**Data N/A

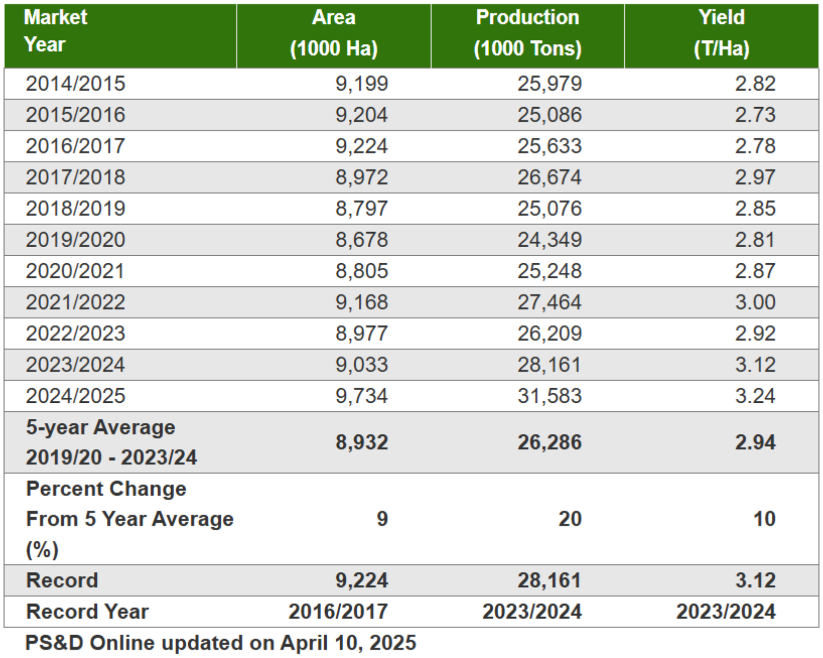
USDA Foreign Agricultural Service
U.S. DEPARTMENT OF AGRICULTURE

Source: Agricultural Marketing Information Service; Pakistan Bureau of Statistics; Crop Reporting Service

Objective

To enhance our dataset for predictive modeling and feature importance analysis, we propose engineering a new feature: Wheat Yield (measured in tons per hectare), based on historical production and area harvested data available from the USDA.





Market Year	Area (1000 Ha)	Production (1000 Tons)	Yield (T/Ha)
2014/2015	9,199	25,979	2.82
2015/2016	9,204	25,086	2.73
2016/2017	9,224	25,633	2.78
2017/2018	8,972	26,674	2.97
2018/2019	8,797	25,076	2.85
2019/2020	8,678	24,349	2.81
2020/2021	8,805	25,248	2.87
2021/2022	9,168	27,464	3.00
2022/2023	8,977	26,209	2.92
2023/2024	9,033	28,161	3.12
2024/2025	9,734	31,583	3.24
5-year Average 2019/20 - 2023/24	**8,932**	**26,286**	**2.94**
Percent Change From 5 Year Average (%)	**9**	**20**	**10**
Record	**9,224**	**28,161**	**3.12**
Record Year	**2016/2017**	**2023/2024**	**2023/2024**

PS&D Online updated on April 10, 2025

Available Data

- We are provided with two critical time-series datasets from the USDA:
- Wheat Production (1000 metric tons) – e.g., 2023: 28,161,000 MT
- Area Harvested (1000 hectares) – e.g., 2023: 9,033,000 HA
- These values span from 1960 to 2024 and are available on an annual basis for Pakistan.

Yield Calculation Formula

We define Yield as:

Yield (tons/hectare) = Production (1000 MT) / Area Harvested (1000 HA)

Since both are expressed in thousands, the units cancel out, and the formula is simplified:

Yield = Production / Area Harvested

Example Calculation (2023):

- Production = 28,161 (1000 MT)
- Area Harvested = 9,033 (1000 HA)

Yield = **28,161 / 9,033 ≈ 3.12 tons/hectare**

This value is consistent with other regional wheat yield estimates and provides a realistic baseline for training.

Proposed Approach

1. Combine the two datasets (Production + Area Harvested) into a single dataframe indexed by year.
2. Calculate yield for each year using the formula above.
3. Merge yield with local feature data, such as soil nutrients, rainfall, temperature, etc., assuming aligned year-wise data is available.
4. Use the engineered Yield feature as the target variable for a regression model to predict wheat productivity based on environmental and soil factors.

Benefit of This Feature

Adding a real, externally validated yield column will:

- Improve model accuracy.
- Allow cross-country benchmarking.
- Enable explainability through feature importance analysis.