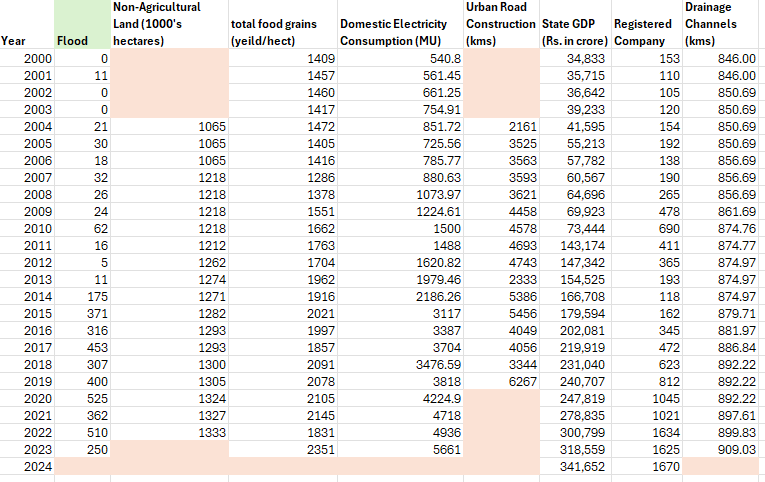
**Creating Flood Count Data**

1. Daily rainfall data - [link](https://nwdp.nwic.in/dataset/rainfall-assam-manual-daily/resource)
2. Consider 90th percentile (extreme rainfall) as indicator of flood (Actuarial Climate Index uses this method [link](https://actuariesclimateindex.org/explore/guided-tour/)) – consider percentiles for 5-year blocks.
3. Annualise the daily flood count.

**Data Dictionary**

* ***Year*** - 2000 to 2024
* ***Flood*** - annual flood count (output variable)
* ***Non-Agricultural Land (1000's hectares)*** - area of land used for non-agricultural purposes captures the shift in land use pattern, if increases then we are reducing forest cover and increasing urbanisation
* ***Total food grains (yield/hectare)*** - measure for crop production indicative of how the agriculture is expanding.
* ***Domestic Electricity Consumption (MU)*** - yet another indicator for urbanization capturing how well the electrification of domestic households are happening
* ***Urban Road Construction (kms)*** - another indicator for urbanization capturing the infrastructural development aspect
* ***State GDP (Rs. in crore)*** - measure of goods and services produced by the state
* ***Registered Company*** - captures the growth of formal business sector
* ***Drainage Channels (kms)*** - captures the drainage channels built, it contributes towards waterlogging

**Missing Value Imputation**



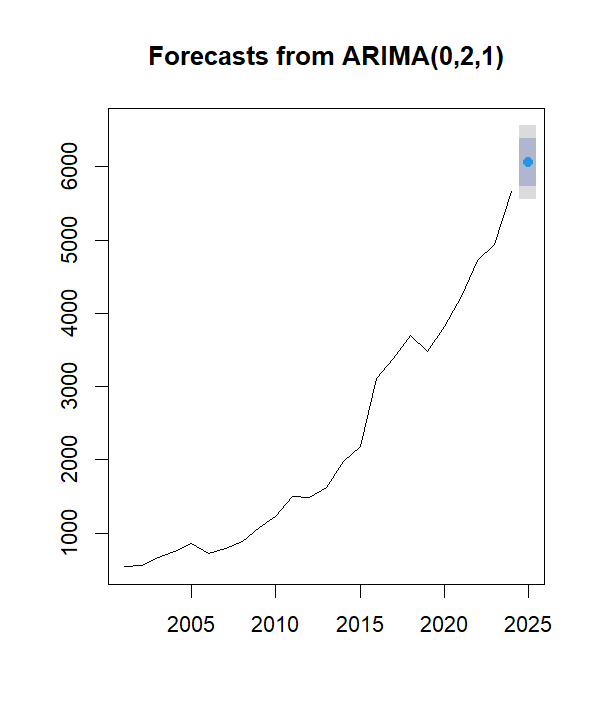
Above is a snapshot of our entire dataset – highlighted cells are the missing values. Instead of simply deleting missing value related entries, we explore various Data Imputation techniques spanning from statistical to machine learning.

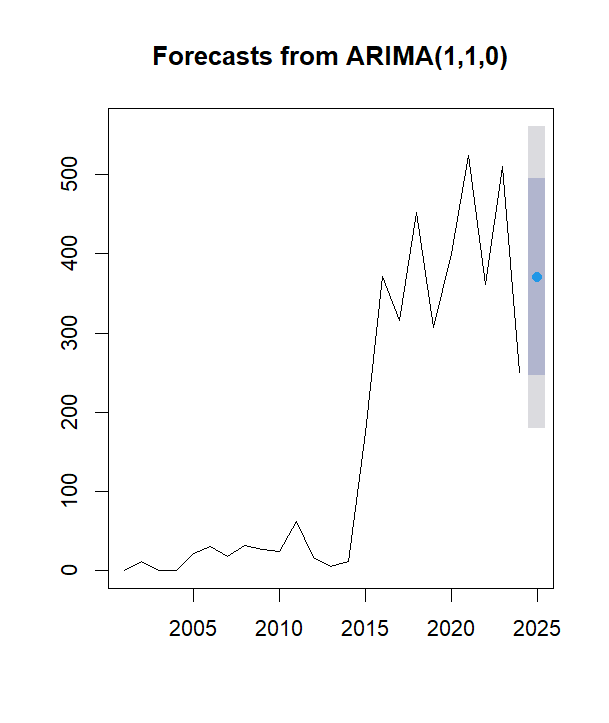
**The 1st layer of data imputation** for variables like *Flood*, *Total Food Grains*, *Domestic Electricity Consumption* and *Drainage Channels* – for these variables, only the last value pertaining to 2024 was missing. So, a simple ARIMA model was fitted to forecast the values for 2024.

In R, we use *auto.arima()* command which finds the best model on the basis of AIC.

|  |  |  |
| --- | --- | --- |
| **Variable** | **ARIMA Model** | **Forecasted Value** |
| **Total Food Grains** | ARIMA(1, 1, 0)  Where, | 2069.65 |
| **Domestic Electricity Consumption** | ARIMA(0, 2, 1)  Where, | 6060.417 |
| **Drainage Channels** | ARIMA(0, 1, 0)  Where, | 911.76 |
| **Flood** | ARIMA(1,1,0)  Where, | 370.85 |

A graph with numbers and lines

AI-generated content may be incorrect.A graph with numbers and a line

AI-generated content may be incorrect.

**The 2nd layer of data imputation** for variables – *Non-Agricultural Land* and *Urban Road Construction*. Here, we explored the following techniques – MICE, K-nearest neighbour and Random forest. Among the three methods, Random forest yielded the most sensible results and was thus selected.

**1. MICE (Multiple Imputation by Chained Equations)**

**Concept:**  
MICE imputes missing values by iteratively modelling each variable with missing data as a function of other variables using regression. It generates multiple plausible datasets (imputations), runs analyses separately on each, and pools the results for robustness.

**How it works:**

* Each incomplete variable is imputed conditionally based on other variables using predictive models (e.g., linear regression, logistic, PMM).
* The process cycles through all variables multiple times (chained equations).
* Produces multiple imputed datasets for uncertainty estimation.

**Best for:**

* Mixed data types (numeric, categorical)
* When relationships between variables are approximately linear
* When missingness is not monotonic

**Reference Paper:**

*van Buuren, S., & Oudshoorn, C. G. M. (1999). Flexible multivariate imputation by MICE.* TNO Prevention and Health.  
van Buuren (2011) Journal of Statistical Software

**2. MissForest (Random Forest Imputation)**

**Concept:**  
MissForest uses a non-parametric approach where missing values are imputed using predictions from a Random Forest model. It's particularly good at capturing complex interactions and non-linearities.

**How it works:**

* Initially fills in missing values with mean/mode.
* Trains Random Forests on observed values for each variable.
* Iteratively updates missing values using predictions from the Random Forests until convergence or minimal change.

**Best for:**

* Datasets with complex non-linear relationships
* Mixed-type data
* Small to medium datasets

**Reference Paper:**

*Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data.* Bioinformatics, 28(1), 112–118.  
DOI:10.1093/bioinformatics/btr597

**3. KNN (K-Nearest Neighbors Imputation)**

**Concept:**  
KNN imputation estimates missing values by finding the *k* most similar (non-missing) observations and imputing based on their values (usually using a weighted or unweighted mean/mode).

**How it works:**

* Computes distance between observations (Euclidean for numeric).
* For each missing entry, selects *k* nearest observations based on available features.
* Uses average (or majority vote) of the *k* nearest values to impute the missing value.

**Best for:**

* When local similarity is meaningful
* Numeric data (after normalization)
* Small amounts of missing data

**Reference Paper:**

*Troyanskaya, O., et al. (2001). Missing value estimation methods for DNA microarrays.* Bioinformatics, 17(6), 520–525.  
DOI:10.1093/bioinformatics/17.6.520