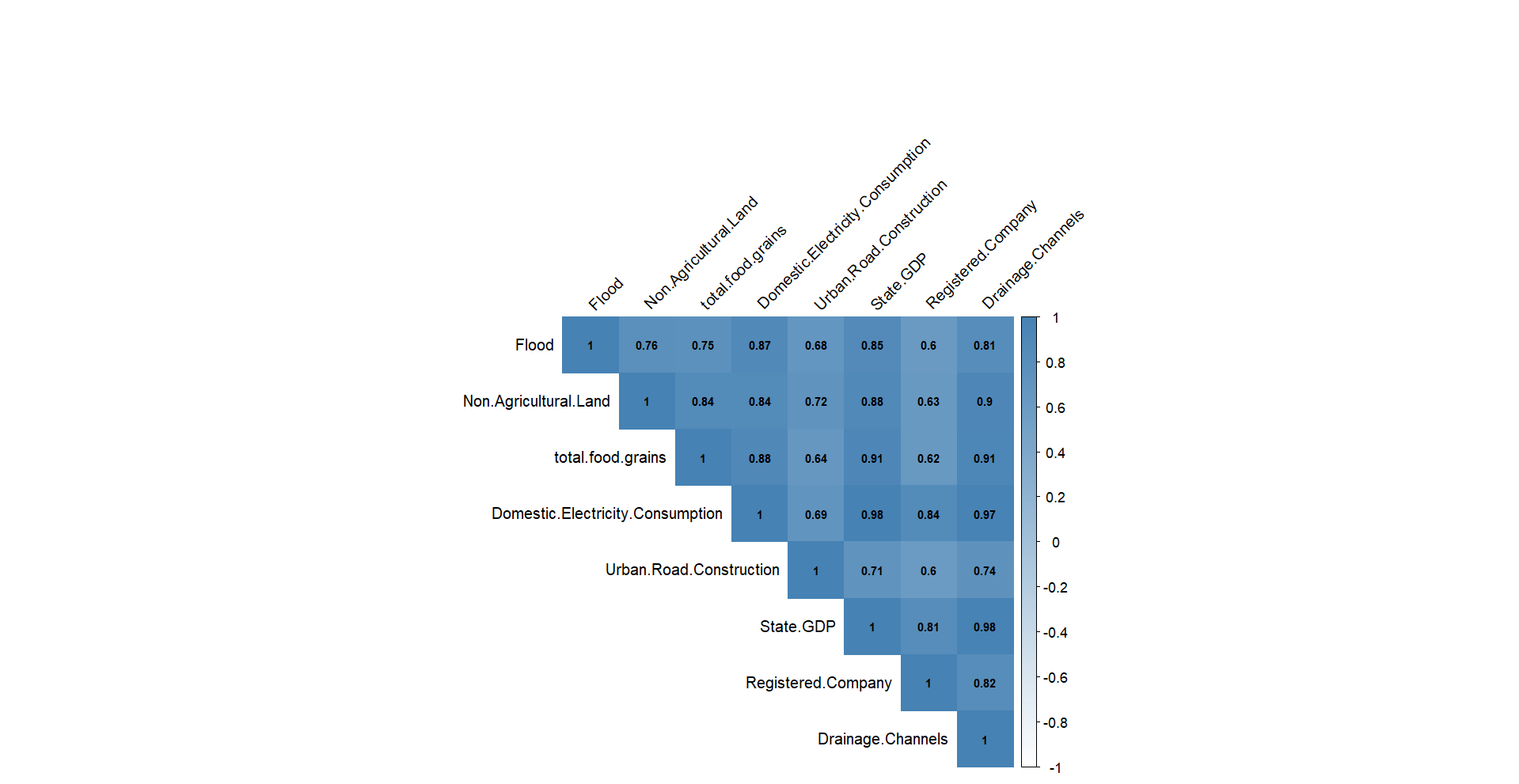
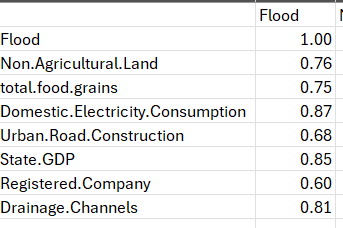
# Exploratory data analysis

## Correlation Analysis



|  |  |  |
| --- | --- | --- |
| **Variable** | **Corr. with Flood** | **Statistical Note** |
| Domestic Electricity Consumption | **0.87** | Strongest linear correlation with flood; tightly aligned trend |
| State GDP | 0.85 | Also very strong; GDP is a proxy for overall urban & economic expansion |
| Drainage Channels | 0.81 | Surprisingly high; could indicate reactive infrastructure post-flood |
| Non Agricultural Land | 0.76 | High positive correlation; strong evidence of land-use change |
| Total food grains | 0.75 | Agricultural productivity rises with development but may co-occur with land conversion |
| Urban Road Construction | 0.68 | Moderate to strong; potential influence of impervious surfaces |
| Registered Company | 0.6 | Moderate; likely delayed effects from formal business expansion |

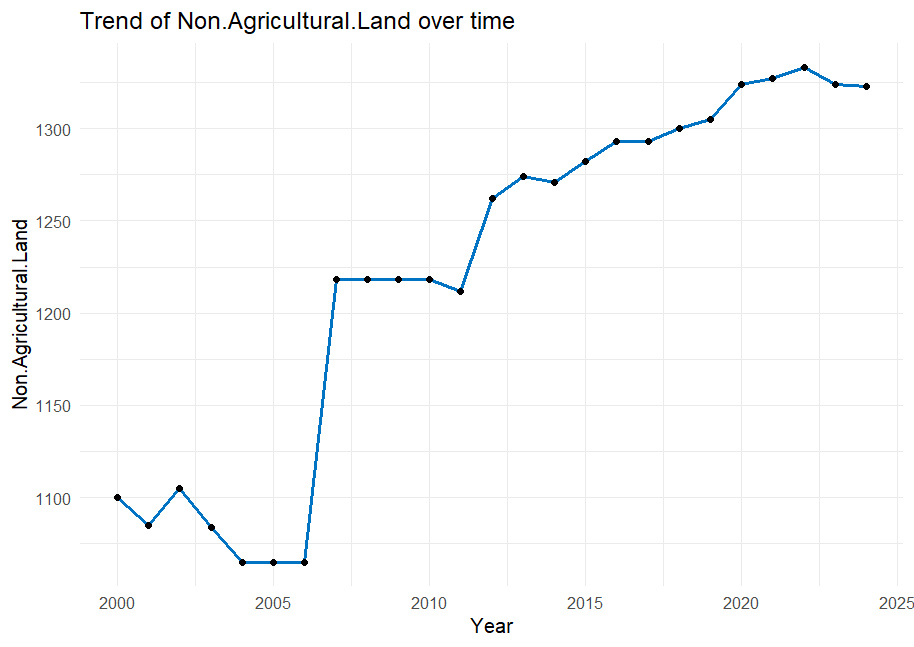
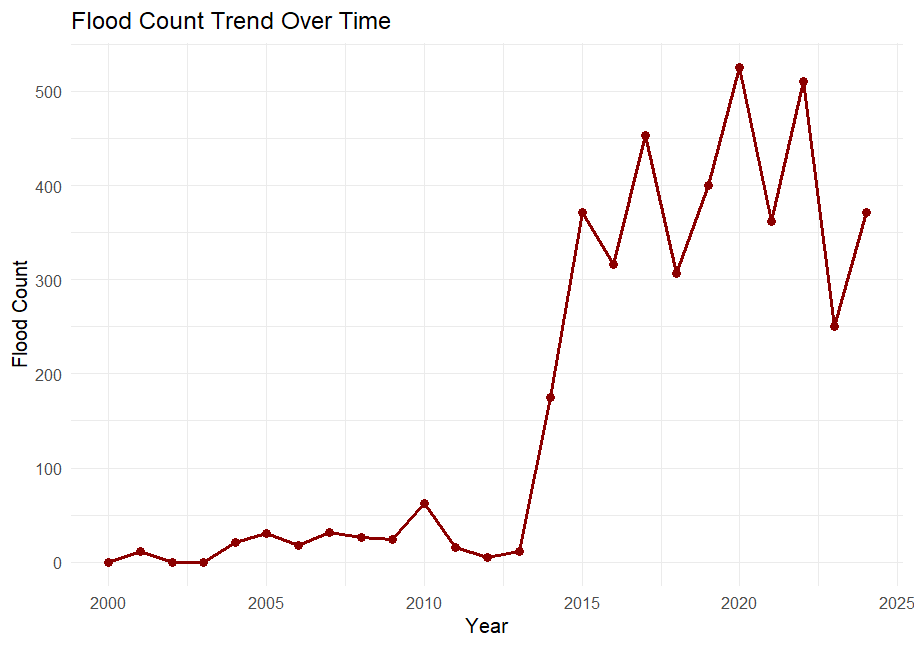
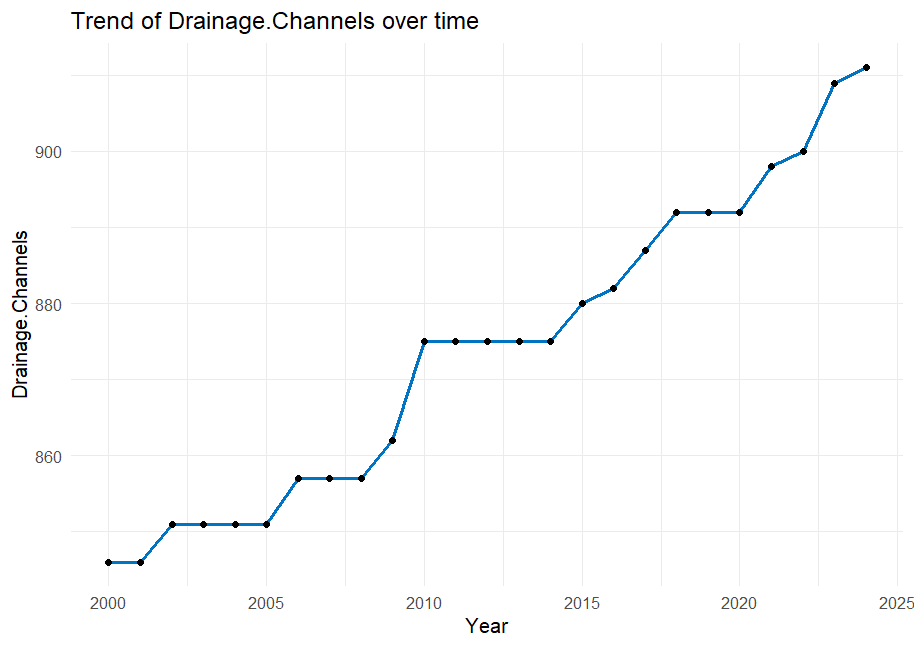
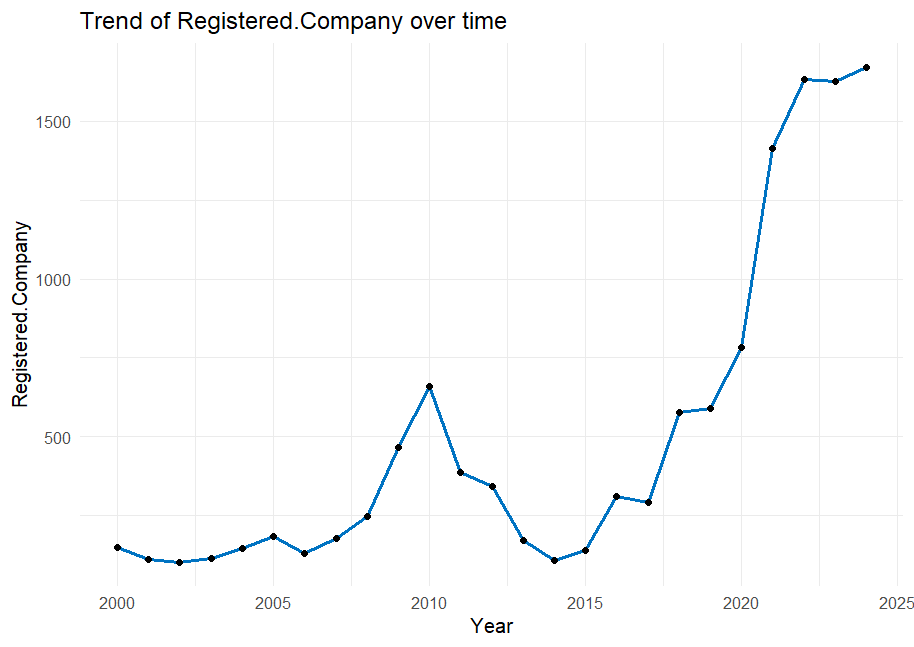
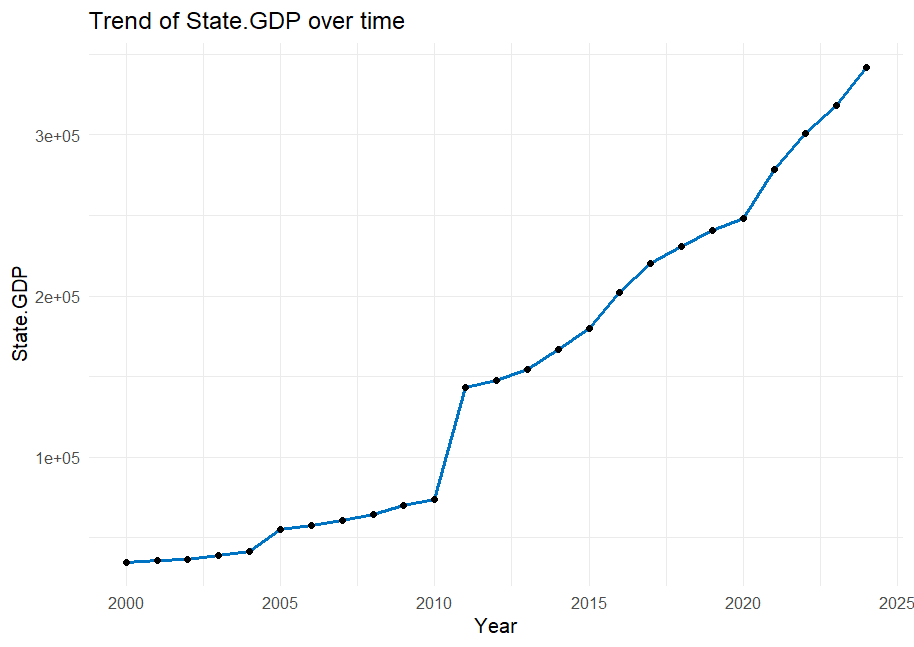
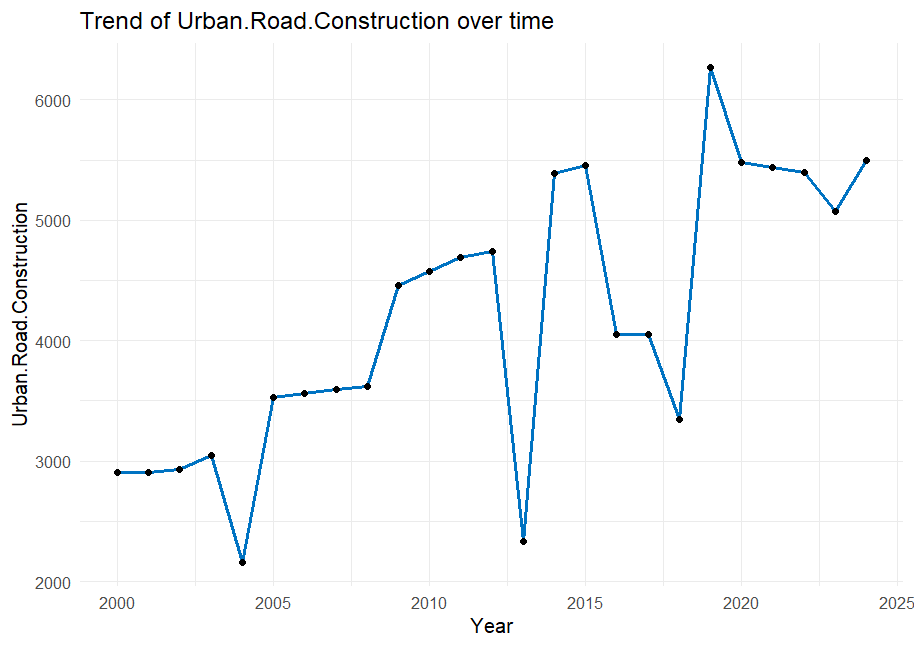
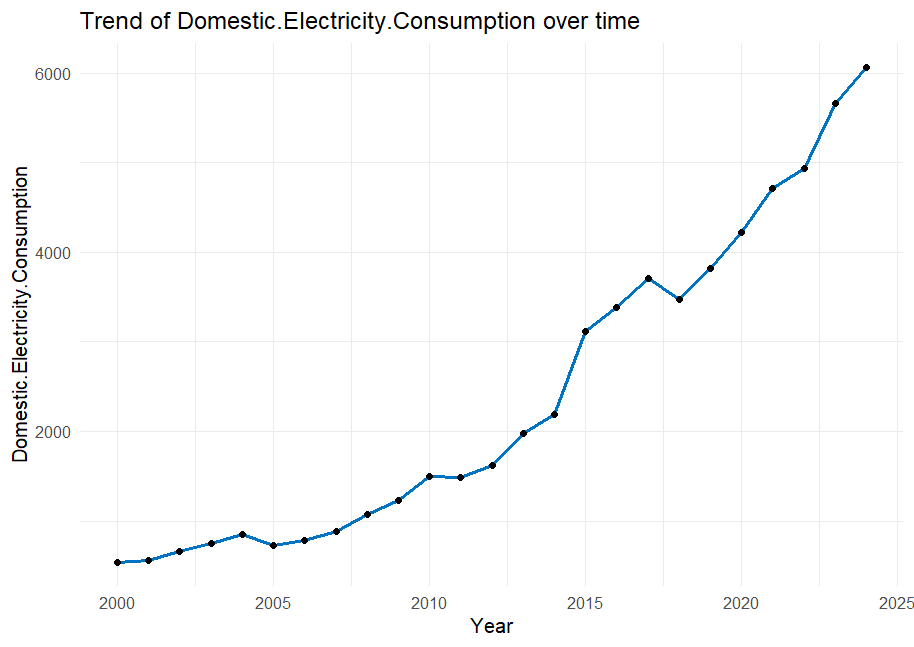
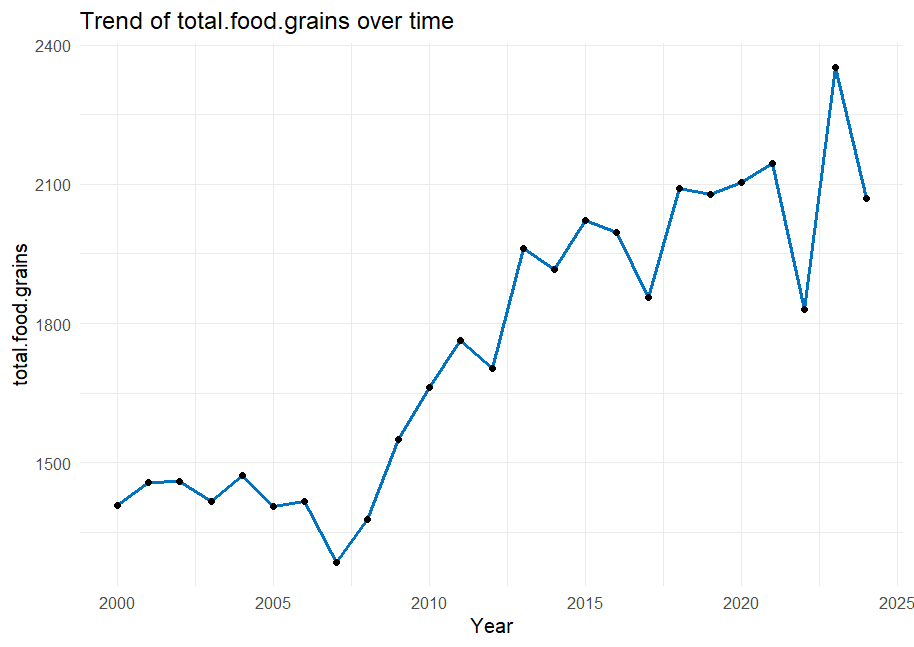
**# Multicollinearity Red Flags**

* Electricity, GDP, and Drainage Channels are all **very highly intercorrelated** (0.91–0.97).
* Suggests **strong multicollinearity**, we will explore Ridge/Lasso Regression Models

Insights

* Urbanization proxies (like Electricity, GDP & non-agricultural land) show strong correlation with floods – supporting the hypothesis that urban expansion is linked with increased flood counts.
* Drainage channels also correlate highly — suggesting either : Post-flood infrastructure response, OR that poorly planned drainage expansion is inadequate to offset flood risk.
* The lower correlation with **registered companies** (0.60) suggests that **formal business growth alone** may not directly cause floods – maybe their impact takes years to emerge. In light of this, we might consider time-lagged effects for these variables (Registered companies & road construction).

## Time Series plot of each variable

**Pending** : Find news articles/research papers/govt reports to support these observations.

* **Non-agricultural land**

A noticeable jump between 2005–2006, followed by a steady climb until 2020. Growth plateaus around 2020. Sudden increase may reflect rapid land conversion, possibly due to urban development policies or industrialization. Conversion from agricultural/forested land → built-up area likely increased surface runoff, contributing to urban floods.

* **Total food grains**

Generally increasing trend after 2008 with fluctuations post-2016. Peaks around 2023, then dips. Suggests expansion of cultivated land or improved agricultural productivity. May lead to loss of natural flood buffers (wetlands/forests), indirectly elevating flood risk. Could also mean irrigation infrastructure expansion, increasing impermeable zones.

* **Domestic Electricity Consumption**

Monotonic and steep growth from ~2009 onward. Near-exponential rise, especially after 2015. Strong proxy for urban household growth and electrification. Supports hypothesis that increasing population density and housing concentration have amplified flood vulnerability, especially in low-lying urban pockets.

* **Urban Road Construction**

Jagged upward trend with sharp spikes and dips (e.g., 2006, 2011, 2017). Post-2012 shows generally high values (~5000–6000 kms). Rapid road expansion likely led to increased impervious surfaces, limiting natural absorption. Fluctuations may align with policy cycles or funding allocations. Could explain periodic spikes in flood counts.

* **State GDP**

Clear exponential growth, especially after 2009. Sharp jump between 2009–2011. Indicates macroeconomic boom — correlates well with other urban growth indicators (electricity, roads). Suggests a period of aggressive development, which may have outpaced ecological safeguards like drainage, zoning, and environmental buffers.

* **Registered Company**

Relatively flat till 2010, slight bump in 2011, then sharp rise post-2016. Nearly tripled between 2016 and 2022. Indicates delayed but strong formalization of the private sector. Increased commercial land use and office construction may have intensified flood risk zones in urban centres (e.g., Guwahati).

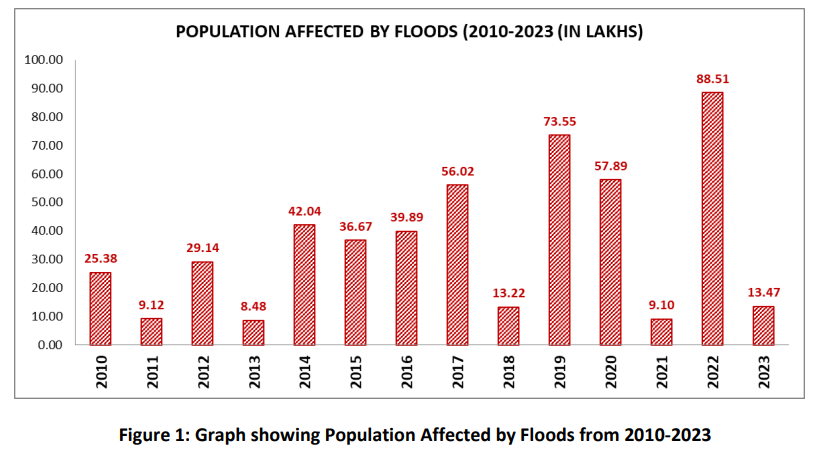
* **Drainage Channels**

Gradual, linear growth from 2000–2024. Slight acceleration post-2010. Although there’s some investment in flood mitigation, the slower growth of drainage channels compared to roads, land use, and GDP suggests infrastructure mismatch. Drainage development may be reactive, lagging behind urban expansion and unable to fully mitigate flood risk.

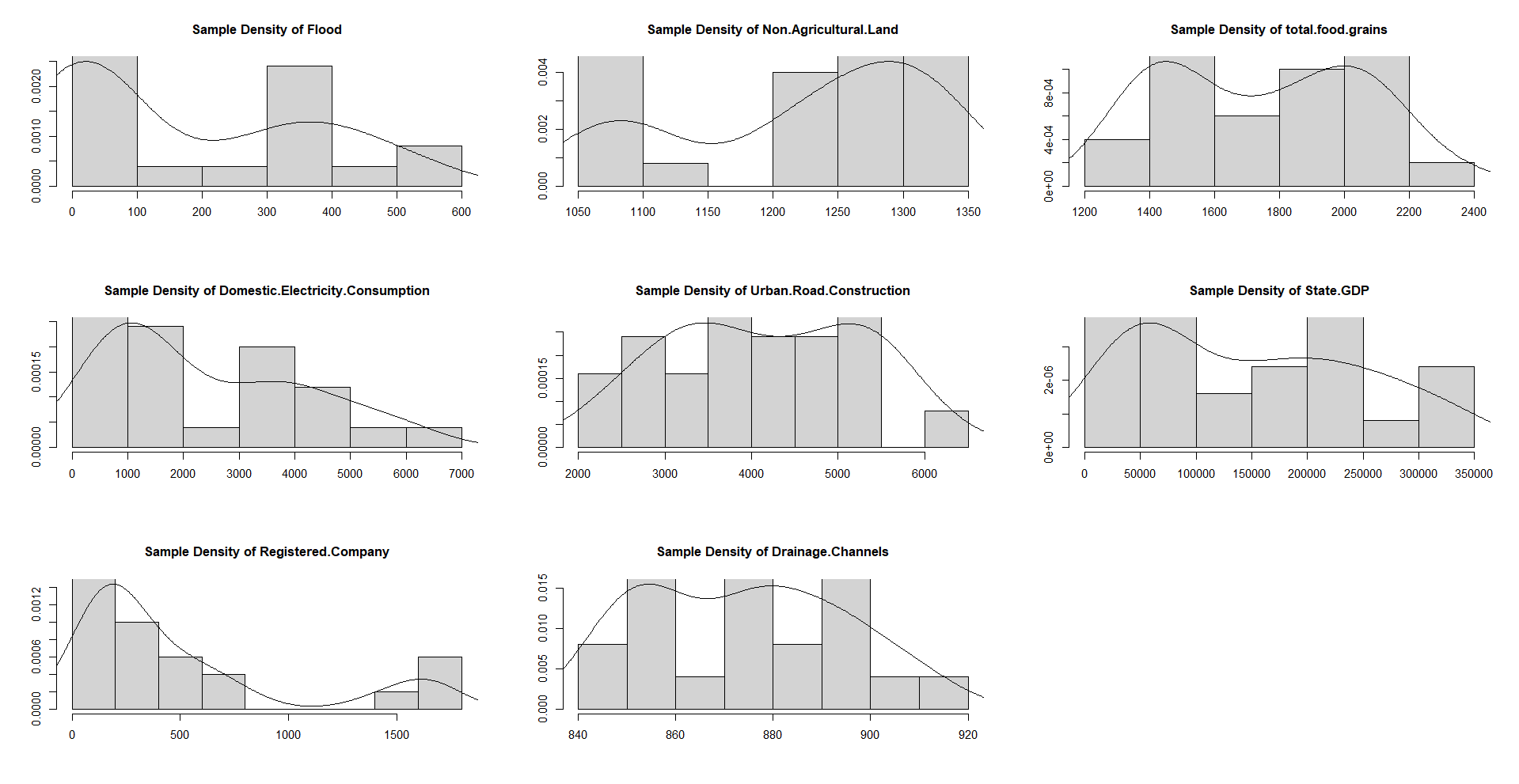
* **Flood Count**

Stable and low until 2013, then a dramatic spike from 2014 onward (persistently high through 2024). Most values post-2014 are above 300. 2014 marks a critical turning point — possibly tied to policy lapses, land conversion thresholds being crossed, or inadequate urban planning. Strong temporal alignment with boom in GDP, electricity consumption, and road construction. Supports argument that urbanization-induced changes, not just natural events, are driving flood trends.

Assam Flood Memorandum 2023 [link](https://asdma.assam.gov.in/sites/default/files/swf_utility_folder/departments/asdma_revenue_uneecopscloud_com_oid_70/menu/document/assam_flood_memorandum_2023.pdf) – Page 7 (Graph confirms the above trend observation)



## Sample Density Plots



|  |  |  |
| --- | --- | --- |
| **Variable** | **Distribution Pattern** | **Statistical Note** |
| **Flood** | Right-skewed, clustered below 400 with outliers | Most years have moderate flood counts, but 1–2 extreme years (e.g., 2014) inflate the tail |
| **Non Agricultural Land** | Slight right-skew | Mostly linear increase — reflects a steady shift rather than volatility |
| **Total food grains** | Bi modal, with humps around 1600 and 2200 | Indicates two distinct agricultural productivity regimes (pre-2010 vs post-2016) |
| **Domestic Electricity Consumption** | Strong right-skew | Rapid growth in recent years, but long tail due to exponential pattern |
| **Urban Road Construction** | Fairly uniform with slight central hump | No strong skew — construction occurred at various magnitudes across years |
| **State GDP** | Right-skewed, long tail | Follows expected economic expansion; few high GDP years pulling mean up |
| **Registered Company** | Strong right-skew | Explains why mean can misrepresent central tendency; large jump post-2016 |
| **Drainage Channels** | Close to normal, slight left skew | Most stable and regularly increasing variable — consistent policy action |

|  |  |
| --- | --- |
| **Variable** | **Interpretation** |
| **Flood** | The skew confirms presence of rare but severe flood years — supports your decision to investigate structural breaks and outliers (e.g., 2014) |
| **Non Agricultural Land** | Steady growth implies continuous urban expansion — supports long-term land-use impact on flooding |
| **Total food grains** | Bi-modal pattern could reflect agricultural transformation, possibly linked to irrigation projects and tech adoption (which reduce natural permeability) |
| **Domestic Electricity Consumption** | Exponential distribution mirrors urban population density growth — aligns with flood vulnerability in expanding cities |
| **Urban Road Construction** | Even distribution suggests no single dominant road boom, but continuous infrastructure buildup contributing cumulatively to impervious cover |
| **State GDP** | Indicates that economic surges are concentrated in a few recent years, which may correlate with policy-driven urban projects |
| **Registered Company** | The skew points to a delayed but sudden burst in commercial formalization, coinciding with flood uptick — potentially through construction pressure on low-lying land |
| **Drainage Channels** | Near-normality confirms gradual infrastructure buildup, but its slow pace vs other variables supports the “infrastructure mismatch” theory in your study |

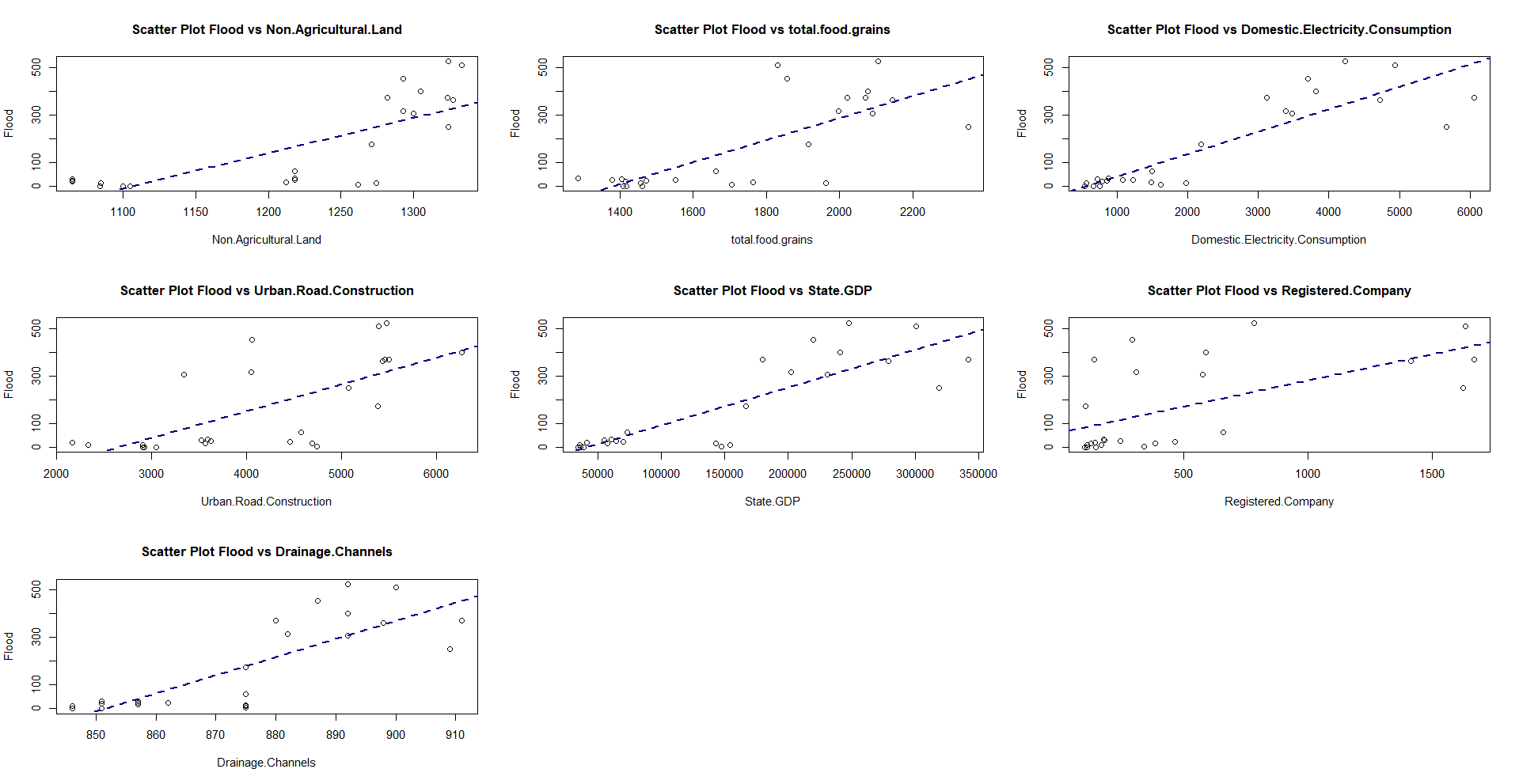
**Overdispersion – Mean & Variance**

* Mean flood count = 171.84
* Variance of flood count = 36023.89
* Variance > Mean – clearly there is over dispersion
* Modelling suggestion – when building GLM models, the ideal choice of family would be Quasi-Poisson or Negative Binomial.

**Suggestions for Modelling**

* Normalize or log-transform right-skewed variables like Flood, Electricity, GDP, Registered Company before regression.
* Consider dummy variables or breakpoints for bi-modal distributions (like food grains).
* Validate outlier robustness with models like robust regression or quantile regression.

## Scatter Plots

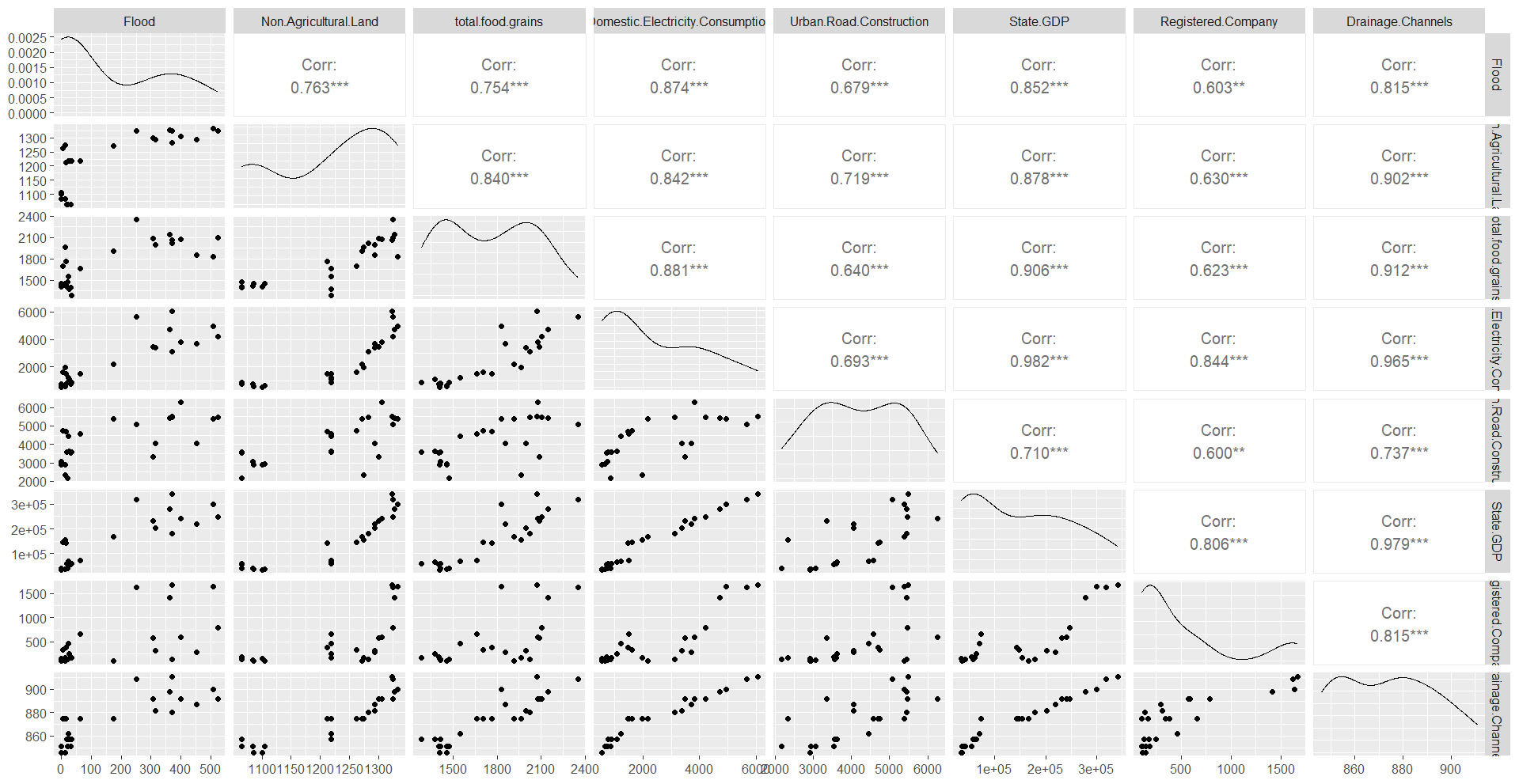


|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Scatter Trend** | **Statistical Comment** | **Interpretation** |
| **Non Agricultural Land** | Moderate positive linear trend | Visibly aligned with flood increase, especially in upper ranges | As land transitions away from agriculture, floods increase — reinforces the role of land-use change in degrading natural drainage |
| **Total food grains** | Linear with widespread | Positive trend but considerable noise | Possibly due to agricultural encroachment on ecologically sensitive zones or intensified irrigation reducing natural permeability |
| **Domestic Electricity Consumption** | Strong positive linear relationship | Clean upward pattern, low scatter in higher range | Strong urban proxy — supports the hypothesis that **urban densification** = **increased flood risk**, especially in settlements without proper drainage |
| **Urban Road Construction** | Linear with slight curvature | Positive but noisy — outliers weaken linearity | Reflects the role of **impervious surfaces**, though noise suggests other co-occurring variables (like drainage, elevation) influence the relationship |
| **State GDP** | Strong linear | Tight alignment — likely a powerful predictor | Confirms that **macroeconomic growth is not flood-safe** — aligns with theory that growth has outpaced infrastructure planning |
| **Registered Company** | Weak positive | High scatter and slight trend — less predictive alone | Weaker relationship may reflect **delayed** or **indirect impact** — businesses concentrate after basic infrastructure is built, so flood pressure is lagged |
| **Drainage Channels** | Clear upward trend | Strong linear path — surprising since it should reduce floods | The positive correlation may indicate **reactive investment** — higher flood years lead to more drainage construction, not vice versa |

**Additional Notes**

* **Electricity and GDP** are likely the strongest predictors of flood counts, statistically.
* **Drainage Channels** behaving like a predictor (rather than a solution) suggests **policy lag**.
* For modelling, consider:
  + **Interaction effects** (e.g., Roads × Drainage)
  + **Time-lag variables** (e.g., previous year's GDP, company growth)

## Pairs plot



**Multicollinearity Concerns**

* Variables like **GDP, electricity, and drainage** have correlations **>0.90** with each other:
  + GDP & Electricity: 0.982\*\*\*
  + GDP & Drainage: 0.979\*\*\*
  + Electricity & Drainage: 0.965\*\*\*
* This implies **high multicollinearity**, which will:
  + Inflate standard errors in linear regression
  + Make coefficient interpretation unstable

|  |  |
| --- | --- |
| **Variable** | **What It Suggests** |
| **Electricity, GDP, and Floods** | Urban expansion (more housing, commerce, services) closely tracks flood count — reinforcing your claim that **urbanisation is a key driver** |
| **Drainage Channels & Floods** | The **unexpected positive correlation** with flood suggests drainage is **reactive** (built after flood years), not preventive |
| **Land & Food** | Conversion of land and increased agricultural footprint (likely irrigated land) reduce permeability and raise **hydro-pressure** |
| **Company Registrations** | Less directly tied — supports the idea that **informal growth and infrastructure precede formal sector growth** |