SPATIAL WARSAW TEAM

UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

To ja czy sąsiedztwo?

Wpływ cech początkowych firm oraz absolutnych i relatywnych czynników lokalizacyjnych na wzrost startupów

dr Maria Kubara
Faculty of Economic Sciences,
Chair of Data Science
Spatial Warsaw Research Group
University of Warsaw
Supported by FNP (Fundacja na rzecz Nauki Polskiej)

Why location still matters?

Even for tech startups?

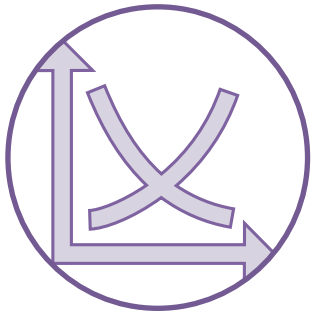
- **Startups** in the same city show different growth paths.
- Digitalisation did not erase spatial heterogeneity.
- Audretsch & Dohse (2013): 'Location is a neglected determinant of firm growth.'
- Within-city variation is largely unexplored in startup literature

Same city, different worlds

To zdjęcie, autor: Nieznany autor, licencja: CC BY-SA

The research puzzle: **Determinants of startup growth**

Initial endowments (assets, team, investors)



Other factors (luck, external conditions)

Location

- **Absolute -** general position in the city
- **Relative -** immediate neighbourhood context

How do these interact?

Which aspect dominates?

From classical to behavioural location theories

Period / School	Main idea	View of location
Classical – von Thünen (1826), Weber (1909), Lösch (1940)	Cost minimisation, profit optimisation	**Absolute**
New Economic Geography (Krugman, 1991)	Agglomeration, transport, scale economies	**Mixed**
Behavioural / Evolutionary (Pred 1967; Boschma 2009)	Bounded rationality, path dependence	**Relative**
Co-evolutionary Geography (Gong & Hassink, 2019)	Firms evolve with regional context	**Strongly relative**

Absolute vs relative location

- Absolute: fixed geography
  Distance to centre, airport, stable over time
- Relative: local amenities, accessibility, neighbourhood attractiveness, density

→ Both are theoretically relevant for **technological startups**, but in *different ways*.

Research goals & hypotheses

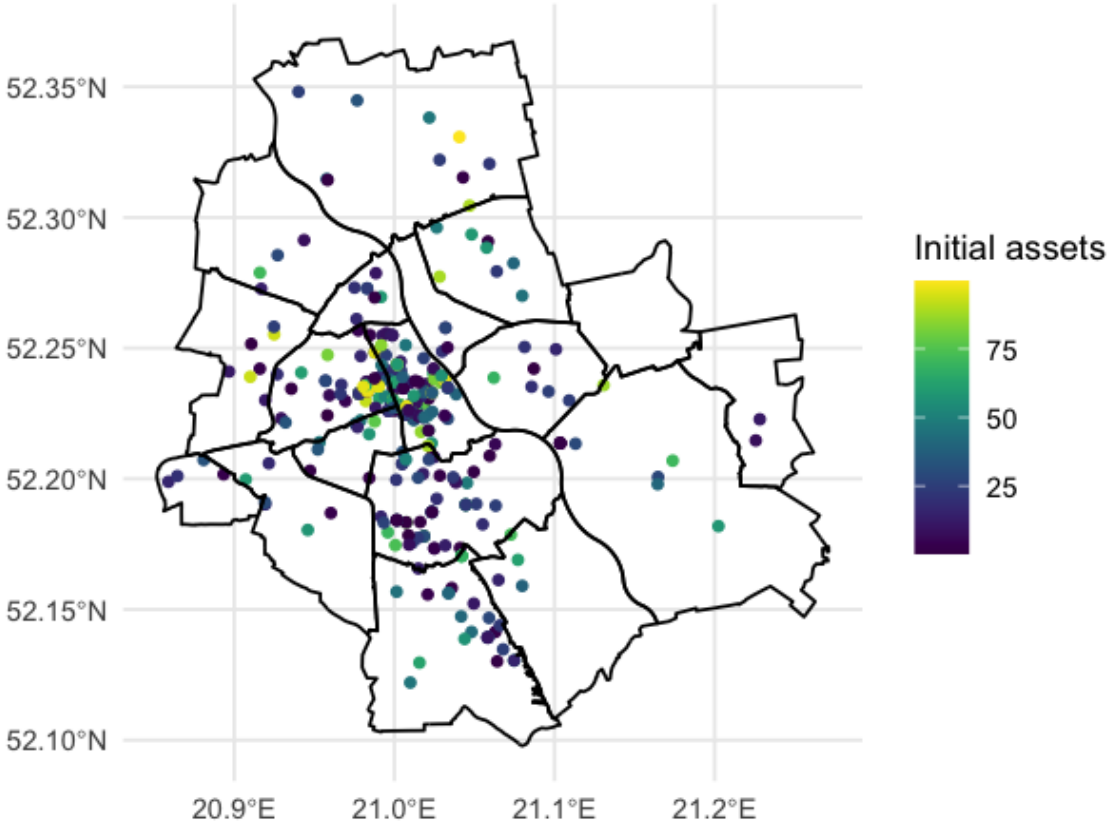
Combine startup-level data and spatial microdata

Test the predictive importance of absolute vs relative factors

Examine heterogeneity: how initial endowments modify spatial effects

Capture non-linear, interaction-driven relations

- H1: Spatial features influence startup growth
- H2: Relative location > Absolute location
- H3: Effects depend on initial firm features
- H4: Relations are non-linear and heterogeneous



Dataset

Technological startups in Warsaw
(founded 2016 – observed for 5 years)
455 firms (software, IT, digital services)

Sources:

- KRS business registry (firm data)
- ORBIS (financials, ownership)
- Statistics Poland + OpenStreetMap (spatial context)

Variables: firm-level + 1 km spatial buffers

Outcome: total assets after 5 years

Variable groups

1. Initial characteristics	Assets, Employees, Ownership (foreign), No. of investors
2. Absolute location	Distance to center, to airport, Central Business District dummy
3. Relative – agglomeration context	Local population density, business concentration, specialisation index
4. Relative – neighbourhood amenities	Accessibility and attractiveness indexes (transport, cafés, restaurants, shopping)

Outcome: total assets after 5 years (proxy for growth)

PCA feature	Variable	Description
Outcome	Total assets after five years of operation	Dependent variable in the model. Total assets of a startup after five years of operation (in hundreds of EUR). If a startup has been closed before this date, this value equals to zero
Initial characteristics	Total assets first	Total assets at the first year of operation (in hundreds of EUR)
	Employees first	Number of employees in the first year of operation
	Ownership Foreign	Dummy variable, whether a company is owned in majority by foreign capital (above 51% in the initial total assets)
Absolute location	Investors	Count variable, how many investors a company has (people or corporate investors)
	Distance to centre	Continuous variable, distance to centre in km
	CBD	Dummy variable, whether a company is located in the Central Business District
	Distance to the airport	Continuous variable, distance to Warsaw's Chopin Airport in km
Relative location features connected to local agglomeration (economic background)	LocalAgglomerationIndex	Relative historical agglomeration index, calculated according to the formula below (index based on the data including the total population of companies operating in Warsaw in 2012, "business census"); variable showing the benchmark of local business competition in an area
	PopulationIndex	Relative historical population density index, calculated according to the formula below (index based on the census data from 2011, population census); variable showing the benchmark of local population density in an area
	LocLQ	Local industrial specialisation, measured by the Location Quotient, calculated on the "business census" data from 2012 (values above 1 show local business specialisation; calculated in small 500m radius). LQ is calculated as the relation between the local dominance of a specific business sector, comparing to the levels of specialisation within the whole NUTS2 Mazovian region
Relative location features describing local neighbourhood characteristics	TransportIndex	Public transport accessibility index, calculated according to the formula below (index based on the OSM data from 2016)
	CafeIndex	Café's accessibility index, calculated according to the formula below (index based on the OSM data from 2016)
	RestaurantIndex	Restaurant's accessibility index, calculated according to the formula below (index based on the OSM data from 2016)
	ShoppingIndex	Shopping's accessibility index, calculated according to the formula below (index based on the OSM data from 2016)

Research goal

Goal: capture *complex, heterogeneous spatial effects* on startup growth

Approach: combine interpretability with predictive strength

How do individual factors vs relative & absolute location features influence startup growth?

$$Total\ assets_{t=5} = Initial\ characteristics_{t=1} + Absolute\ location\ features + Relative\ location\ features\ related\ to\ loagglomeration + Relative\ location\ features\ related\ to\ neighbourhood\ traits$$

Research design and methodological choices

Steps:

1 **Dimensionality reduction** – *Principal Component Analysis (PCA)* to summarise feature groups

2 **Prediction model** – *XGBoost* (gradient boosting) to handle non-linearity & interactions (in two versions)

3 **Interpretation layer** – *Explainable AI (DALEX)*:

- Feature importance → key drivers
- SHAP / PD / CP → direction & heterogeneity of effects

Why this combination:

- Handles multicollinearity and spatial heterogeneity
- Avoids linearity & homogeneity assumptions
- Still interpretable through XAI

Specificities of spatial data & training

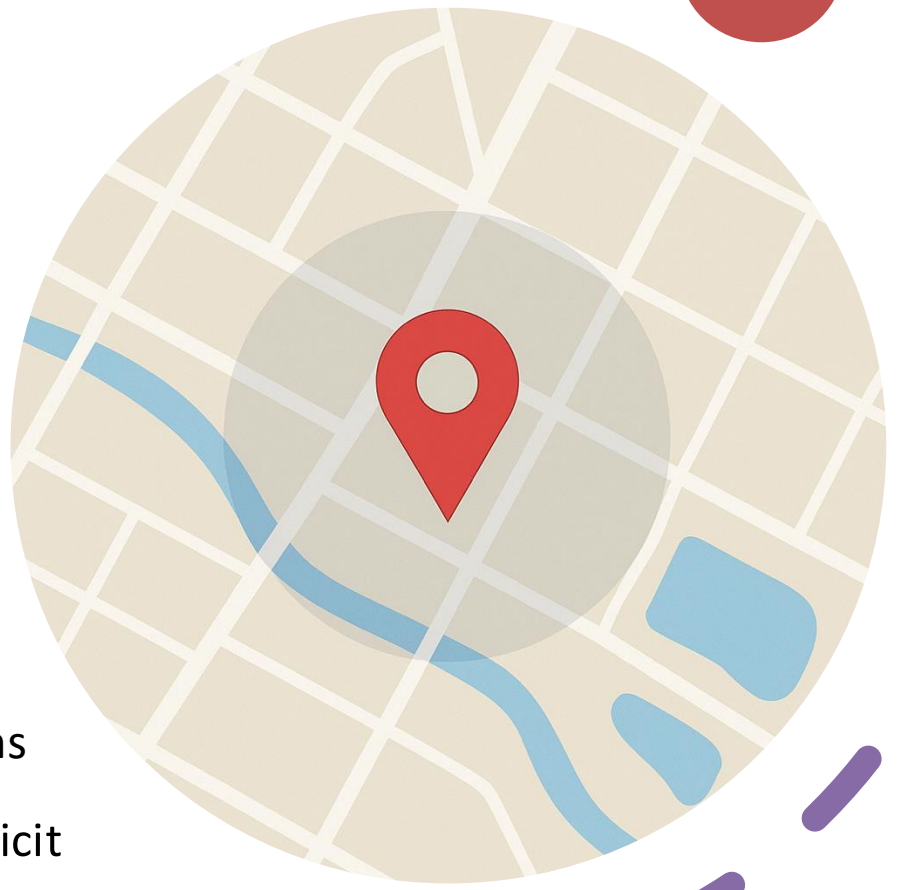
Challenges:

- Uneven spatial density of observations
- Correlated local features (spatial dependence)
- Different spatial scales of influence
- Risk of overfitting to dense areas (e.g. city centre)

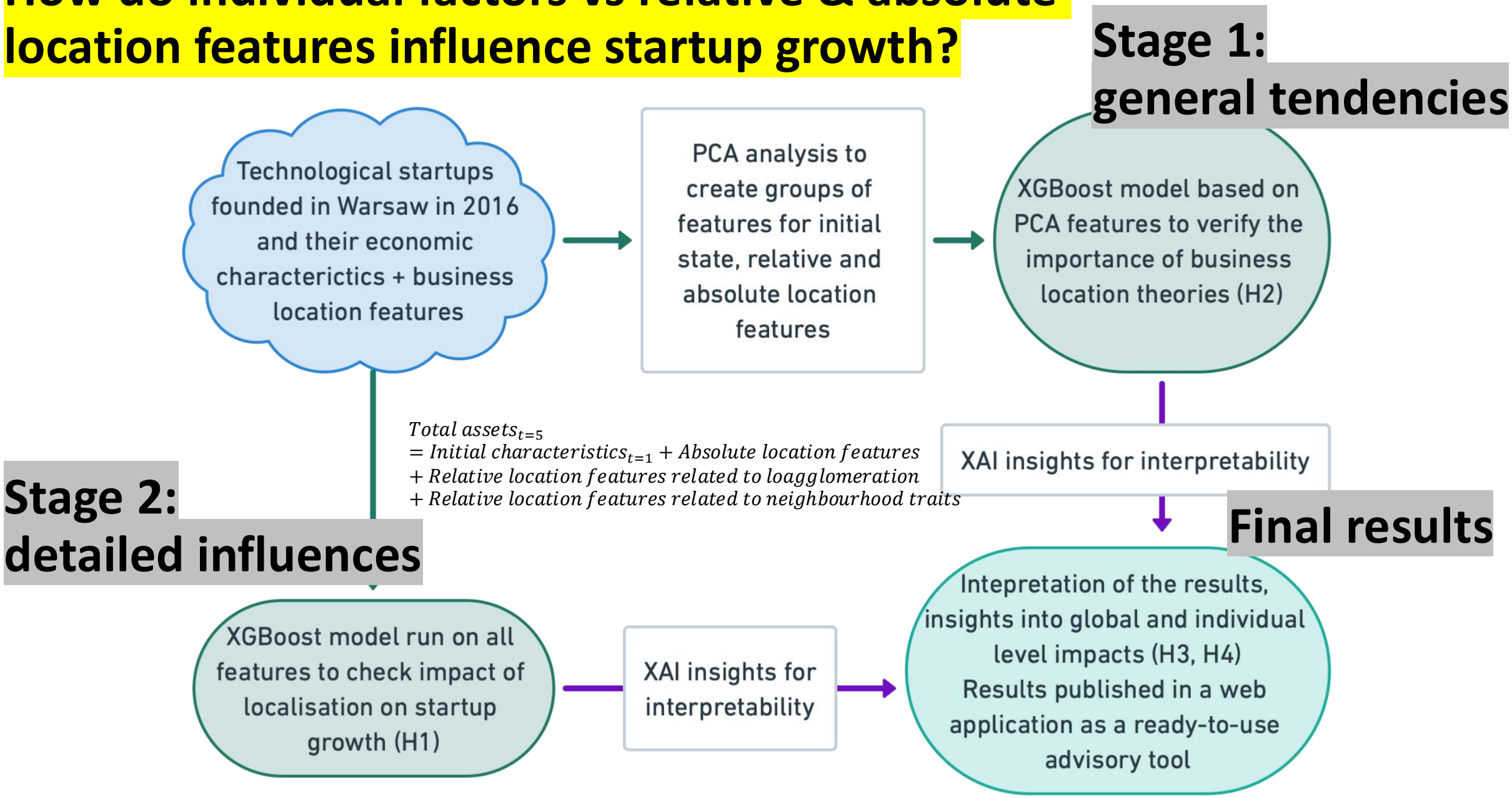
Solutions applied:

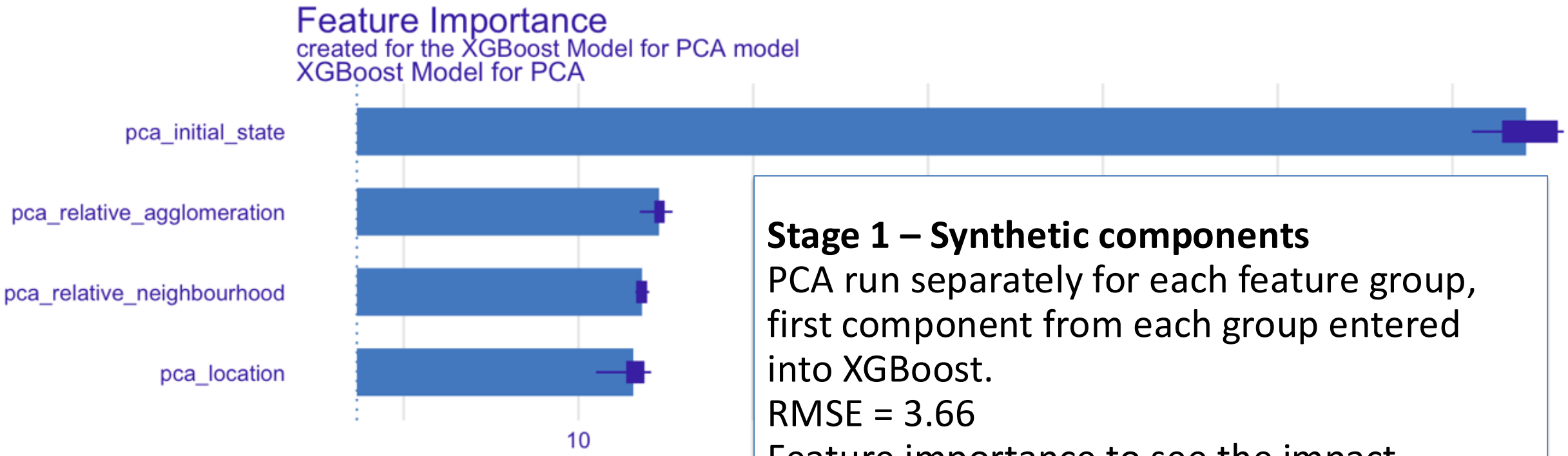
- Micro-geographic variables aggregated in **1 km buffers**
- Normalised & rescaled indices (0–100) for comparability
- Separate **absolute vs. relative** feature groups
- **Cross-validation** and stability check across Warsaw districts to verify model generalisation
- XGBoost handles interactions & non-linear (spatial) patterns

Outcome: Robust model capturing local variation without explicit spatial weights



How do individual factors vs relative & absolute location features influence startup growth?





Stage 1: PCA reduced model

Stage 1 – Synthetic components
PCA run separately for each feature group, first component from each group entered into XGBoost.
RMSE = 3.66
Feature importance to see the impact.

Findings:
All groups contribute to growth prediction
Relative neighbourhood and *relative agglomeration* ≈ strongest among spatial factors
Confirms spatial influence even in aggregated form

Stage 2: Full XGBoost model

Stage 2 – Full model with all variables
RMSE = 2.88

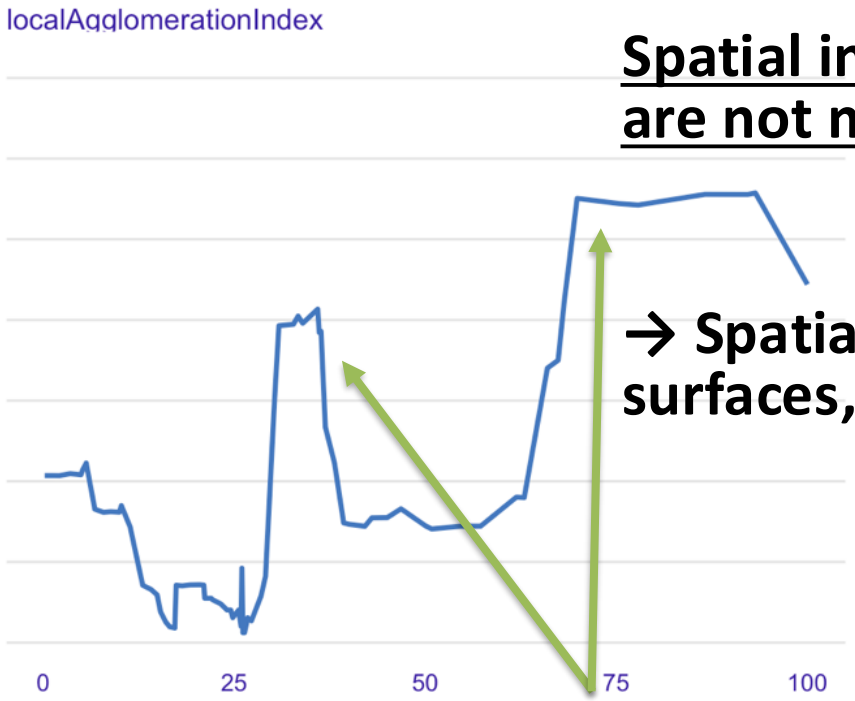
Top features (Feature Importance):

1. Initial assets
2. Employees in first year
3. Distance to centre / airport
4. Local agglomeration, neighbourhood indices

→ The full model reproduces intuitive patterns — initial capital and team size remain crucial — but spatial variables follow closely.
Location retains strong explanatory power.

Non-linear & contextual patterns

Distance to airport: two optima (~2 km and > 11 km)



Agglomeration index: benefits in both dense and low-competition areas
Negative zone in mid-range → congestion / noise / competition trade-offs

<u>Spatial impacts are not monotonic</u>

→ Spatial effects = complex surfaces, not straight lines.

Heterogeneity & interpretability

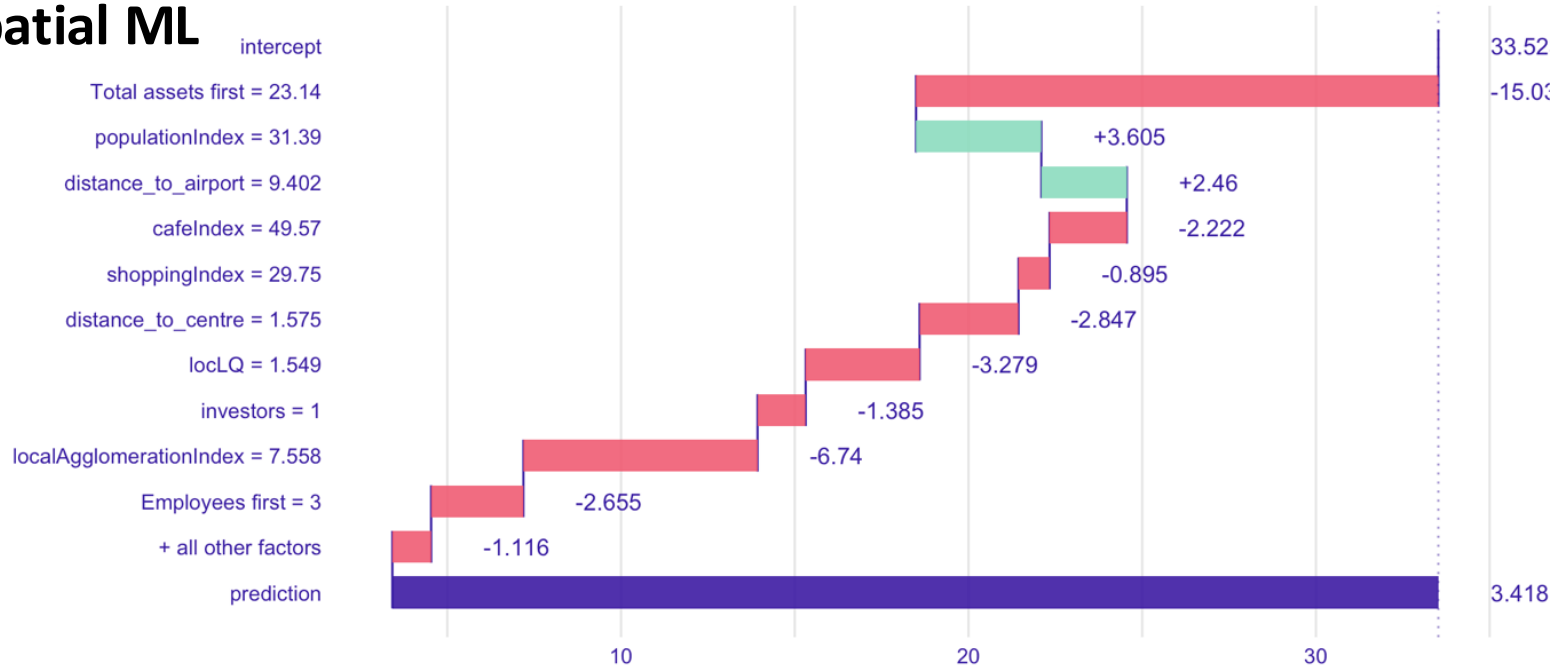
Individual-level explanations (XAI)

- SHAP values → variable contributions for each startup
- Ceteris Paribus profiles → “what if” scenarios

Example startup:

- close to centre, low agglomeration → predicted decline in assets
- potential improvement: different neighbourhood / extra investor

XAI = interpretable micro-level spatial ML



Practical application

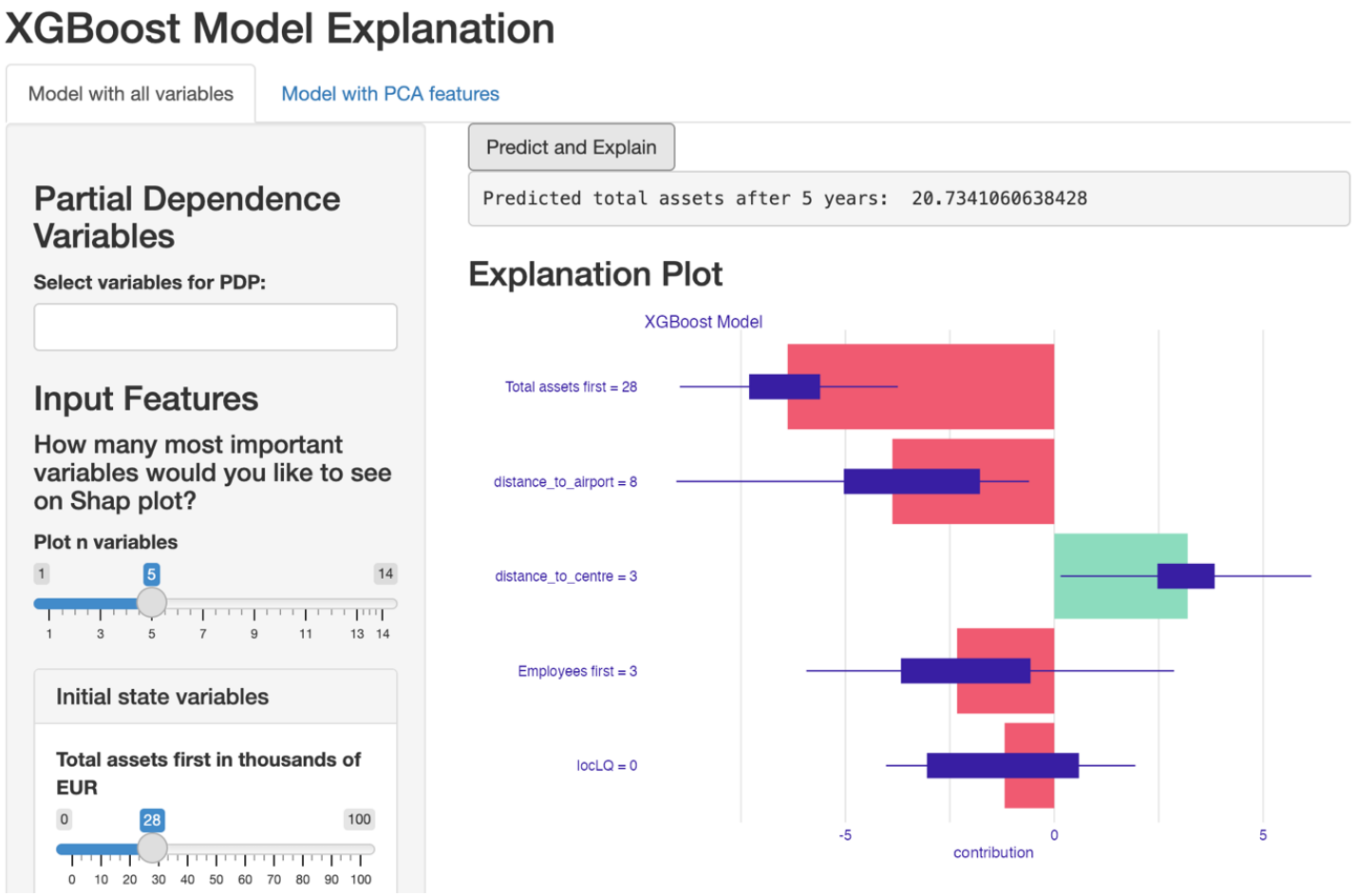
Interactive web tool (Shiny app)

- Open-source model deployed for exploration

Users can:

- input startup characteristics
- adjust spatial indices
- view predicted 5-year growth & feature effects

A starting point for possible policy guidance tool



tinyurl.com/appStartupXAI

Main conclusions

- ✅ Location matters — even for digital startups
- ✅ Relative > Absolute location for intra-urban variation
- ✅ Startup–location interactions are **heterogeneous**
- ✅ Relations are **non-linear**
- ✅ XGBoost + XAI = interpretable, data-driven bridge between ML & regional science

Policy relevance:

- tailor startup support to local micro-contexts
- inform spatial planning & advisory services

Next steps

- Extend analysis to other CEE cities
- Combine with other datasets
- Integrate temporal dynamics (LSTM / spatio-temporal ML)
- …

maria.kubara@uw.edu.pl
@KubaraMaria

SPATIAL WARSAW TEAM

UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

To zdjęcie, autor: Nieznany autor, licencja: CC BY-SA