Data science: cases and observations

…from personal and Reaktor's point of view

Janne Sinkkonen

Outline

Our company, shortly

Doing data science in practice, from Reaktor’s point of view
And some opinions about things

A case example: Yle recommendations

More case examples if you wish (haven’t prepared these but I have slides)

Please ask

We have plenty of time, and having some discussion is more entertaining and clarifying..

Chat (txt) or interrupt

Offices mainly in Helsinki, but also NY, Stockholm, Amsterdam, Dubai, and a bit elsewhere (Tokyo, Tampere etc.)

Reaktor	100% (past-)employee ownership	110M€ turnover 2020, est.	2000 Year founded	500–600 emloyees
Reaktor strategic partners Reaktor Ventures & Reaktor partners	12 companies	24 startups	>15 countries	>1200 people

Reaktor: less formal

20 years, data science 8 years

Growing maybe 10–20% per year, now also abroad

Flat organization

“Owned by employees”, strong culture

Emphasis on the human side: team, communication, wellbeing, …

Pioneer on agile methods; fits well with ds (empirical, iterative attitude)

Consultancy vs. products

Consultancy here means selling your work to customers who "own the product".
You are in the role of an expert, but it's still mostly development, not slides!

Consultancy: technologically wide
Products: technologically deep

You need some domain knowledge.
Context, as opposed to the tech. core, is emphasised in practice.
(Still you can't do it without understanding the technology.)

You are not in customer's organization, which is both good and bad.

Reaktor data science etc.

Includes: data science, machine learning, “AI”, data engineering

We don’t have definite roles, so it is hard to say how many we are, but roughly 20–35.

Not homogeneous: data engineering, statistics, sales, biz design, machine learning, NLP, machine vision.

We work mostly in teams, from two to several people.

Projects are pretty organic, often start with small and the grow up. Some last weeks, some years.

AI, ML, ds

AI is a recent hype term (since ≈2015), harmful to markets.
- Used to mean almost human-level intelligence.
- Now wide in scope, includes linear regression in an operational context.

Data science is a cloudy concept as well.

A useful division maybe, in practice:
- Inferential work, ≈ statistics, data science
- Operational systems, ≈ machine learning
- Infrastructure, cloud etc., ≈ data engineering
- Conceptual work around the core and before, ≈ AI/biz design

Beware of data

Data is a useful term when it refers to **storage or transfer**.
- In storage and transfer, semantics of bits mostly don't matter.

When data refers to **measurements** supposed to be operational or increase understanding, context is fundamental!

You need to understand in detail where the data originates.

Preferably, you should decide how it originates.

Compare to science: observational data vs. controlled experiments.

Data and then what?

Report? Ok… where does it lead? What are the decisions, actions?
- You need to know (estimate) the effects of those actions.

Operational system? It needs to act as well, i.e., make choices.
- Again, you need causal inference (often implicit).

How can you know the effects of the actions without doing the actions?
- Theory? Usually it is weak or nonexistent on many domains (sales, recommendations)
- Or, you need actions/interventions, and this **contradicts the idea of passive data**.

You need controlled, experiments, reinforcement learning, etc.

Cornerstones of an “AI” solution

ML / DS

- Design of measurements and interventions
- Integrity of data
- Models
- Interfaces of models (APIs)
- Implementation: efficiency, scalability etc.

Data engineering

- Infrastructure
- Data flow, storage
- Security
- Transparency
- Correctness

Biz design

- The goal: what it fulfills
- Possible?
- Who does it?
- Who needs to be involved at the customer
- “Operationalization”
  - Design of measurements (data) and interventions
  - Overall system architecture

Example: Yle recommendations

Yle: Finnish national newsmedia (compare to BBC, or NRK in Norway)
Was: Radio and then TV
Now: yle.fi and Areena, a Netflix-like streaming service

Both the news web site and Areena need a recommendation system

We have been doing this for about 6–7 years now.

It is not all about a core algorithm. ;)

Example of “data”

Data is a byproduct of UI’s, no-one has ever used it -> broken by default

Heterogeneity of UIs (clients), with legacy: heterogeneity of data

Heterogeneity of users: bots are involved, but no clearly separated

“Item has been viewed”: no clear definition

Non-causal: co-occurrences A&B, but do they imply A -> B or B -> A or both?

Yle: environment

Use of the service is preferred to non-use (reading time, viewing time, click rate)

Yle wants to have young customers as well

Yle also has somewhat lofty goals of education etc.

Journalists, editors, they want to have a say -> hybrid system (ML & human)

Huge long tail of content, popular content changing rapidly

Diverse client software (including Elisa etc., Apple, etc.)

Algorithmic solutions

Components, iterative development, lots of trial and error.

A/B testing? Yeah, but a long story.

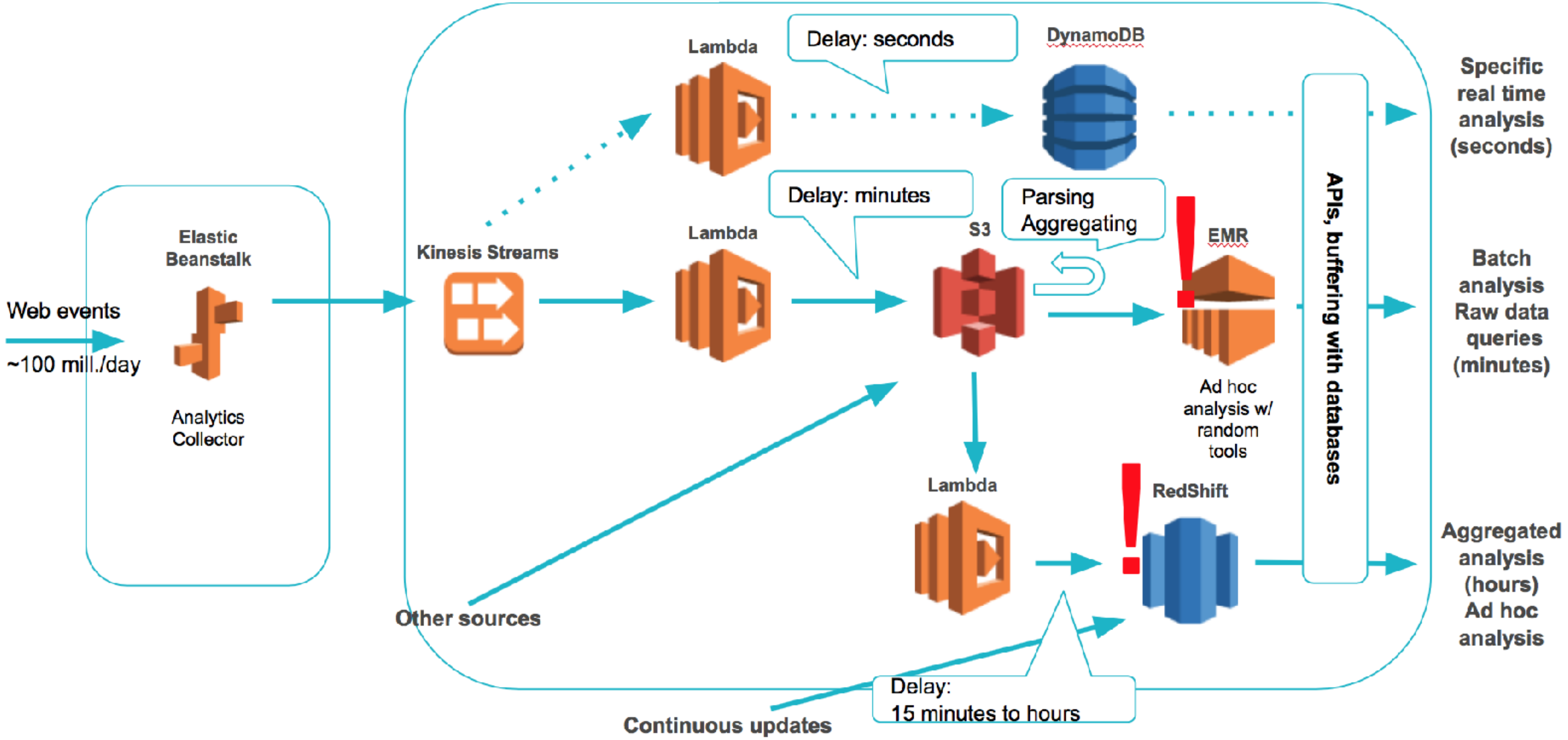
Matrix factorisation: good for genres etc.

Association rule -like heuristics: good for long tail, kind of local over content and users

Popularity separately

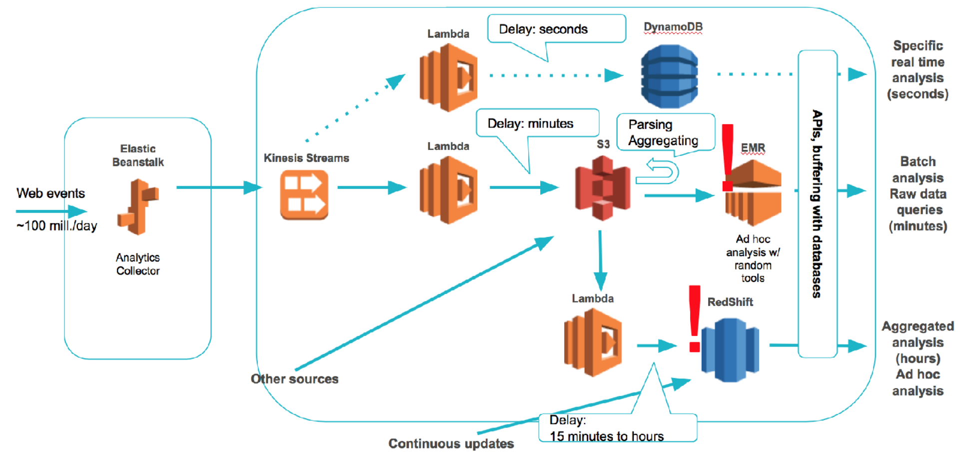
Some NLP

Yle Analytics Pipeline



Immediate architecture of recommendation

Yle internal APIs
- programs (title, availability,...)
- users



"AI"

model computation

db for models

Auto-replicated containers
- dot products $u^T v$ and sorting
- filtering, rules

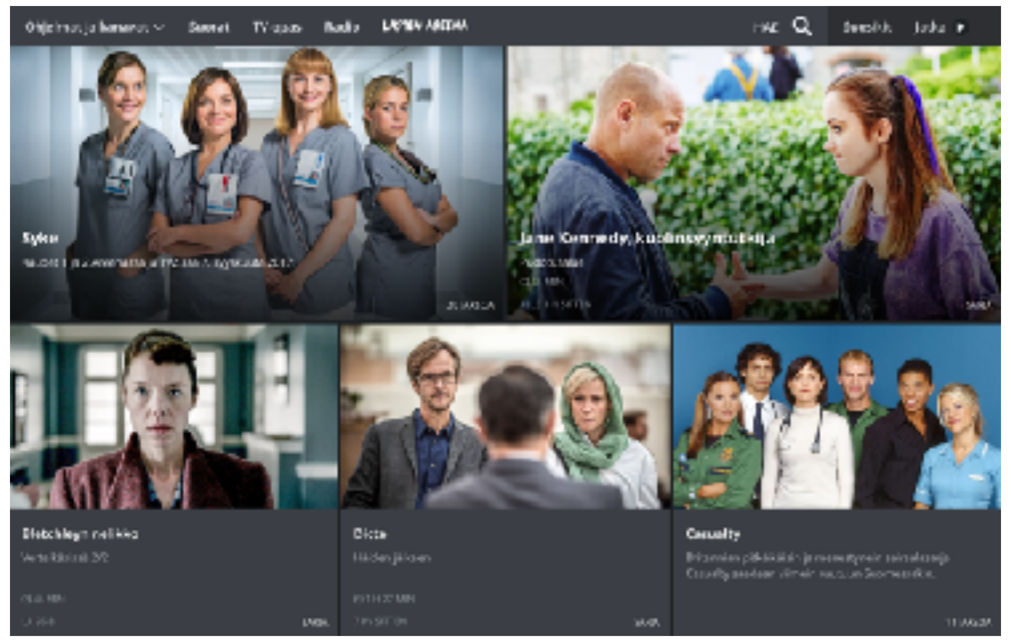
"AI"

"AI"

"AI"

"AI"

cache

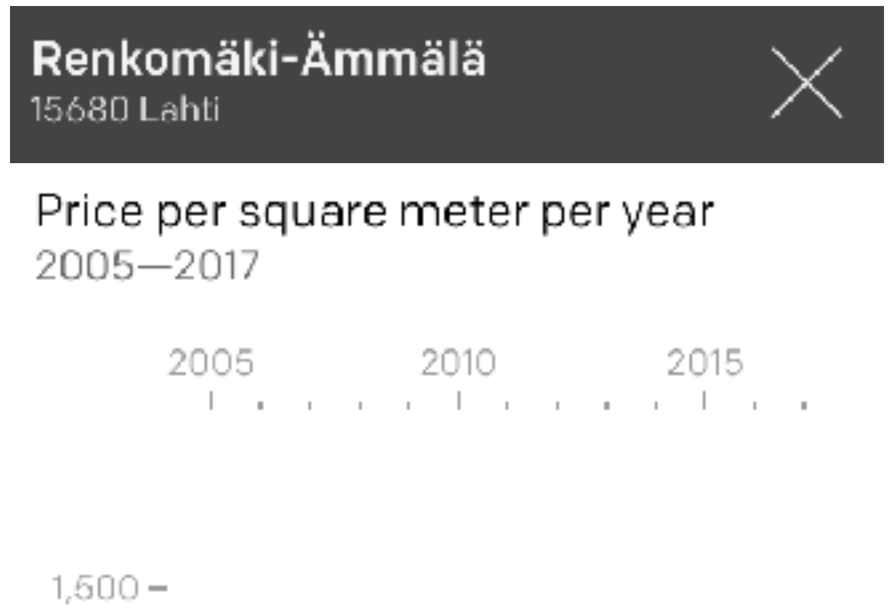


Areena UI

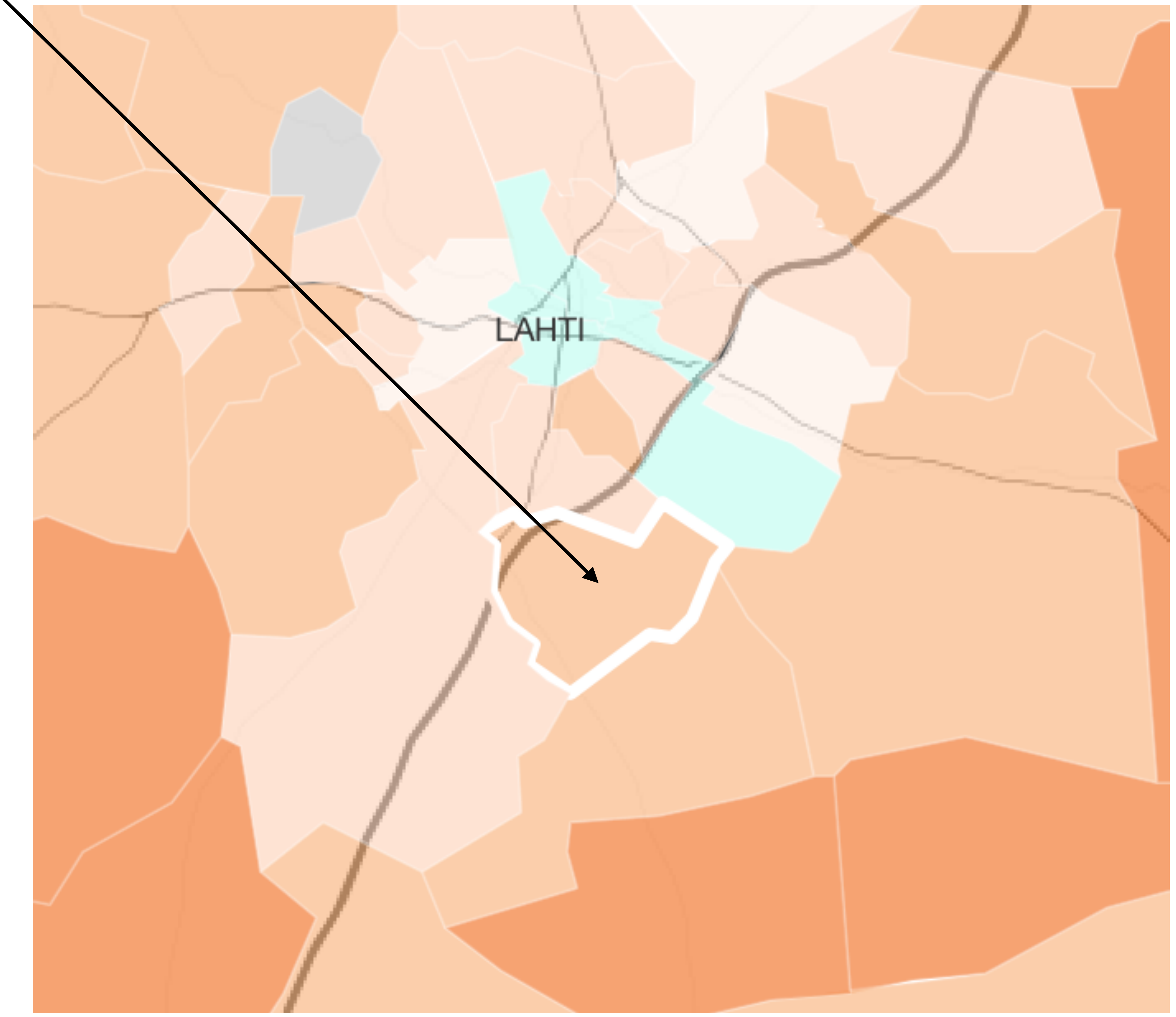
user

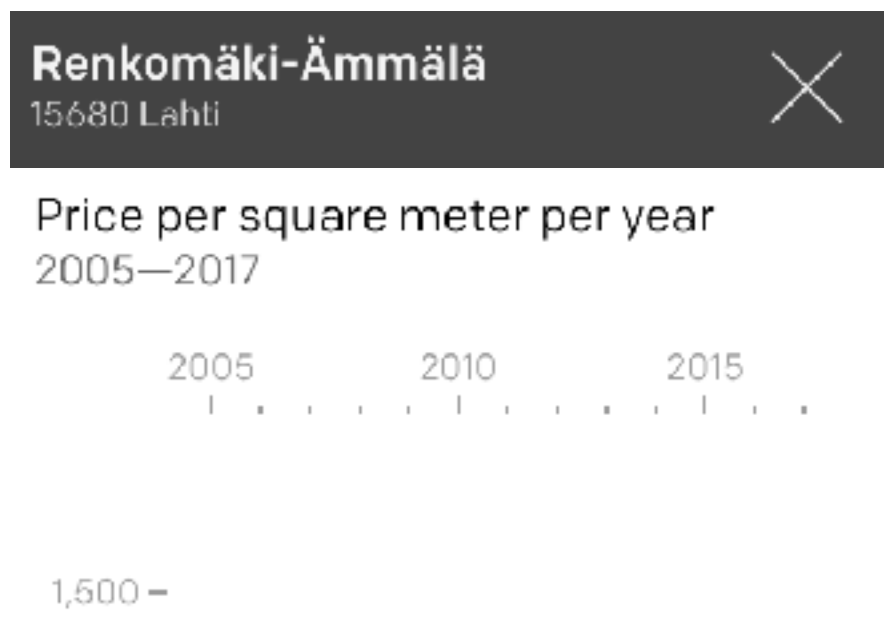
Estate price level estimation

Estate price level estimation



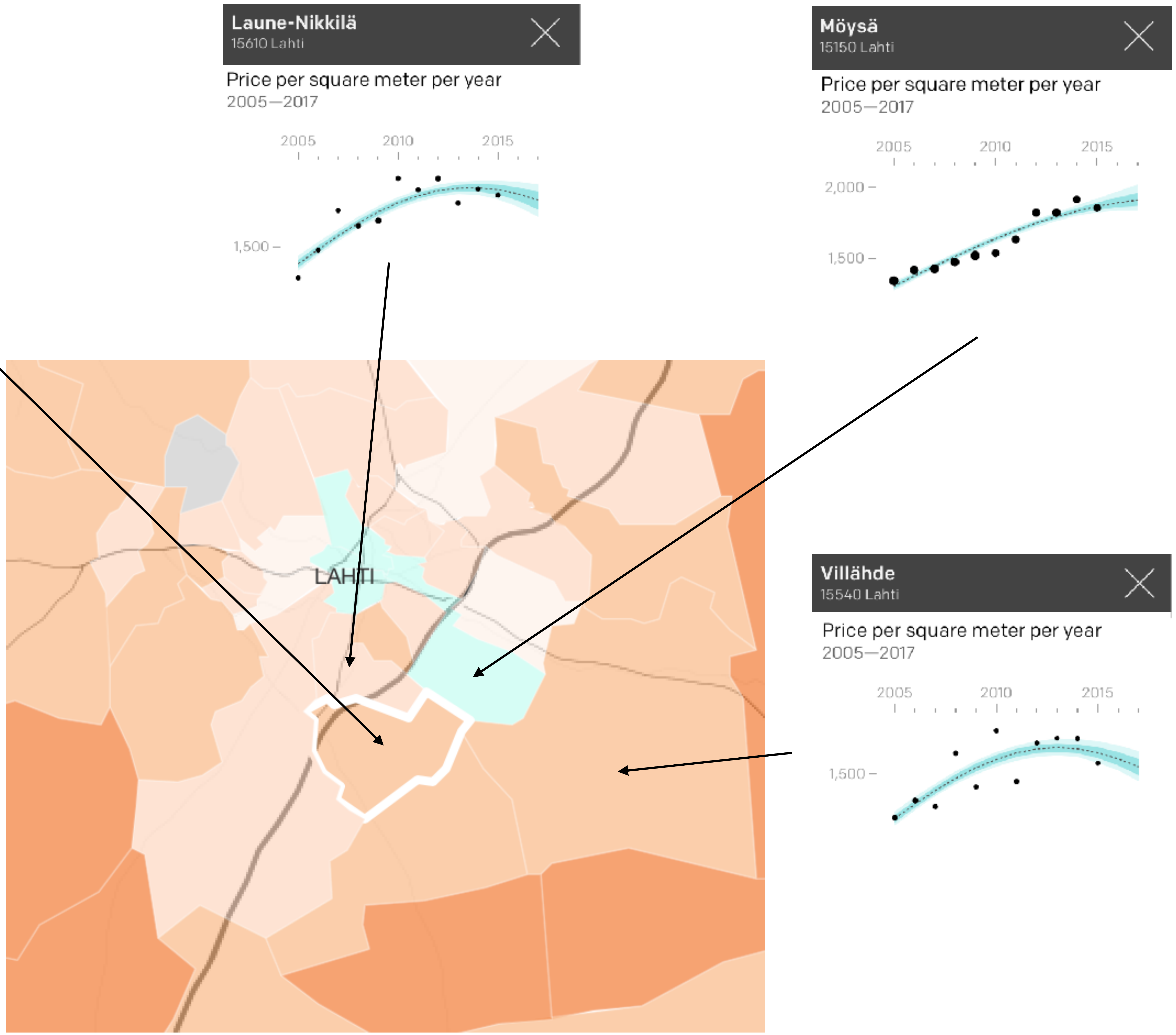
How to estimate the yearly prices for Renkomäki-Ämmälä?

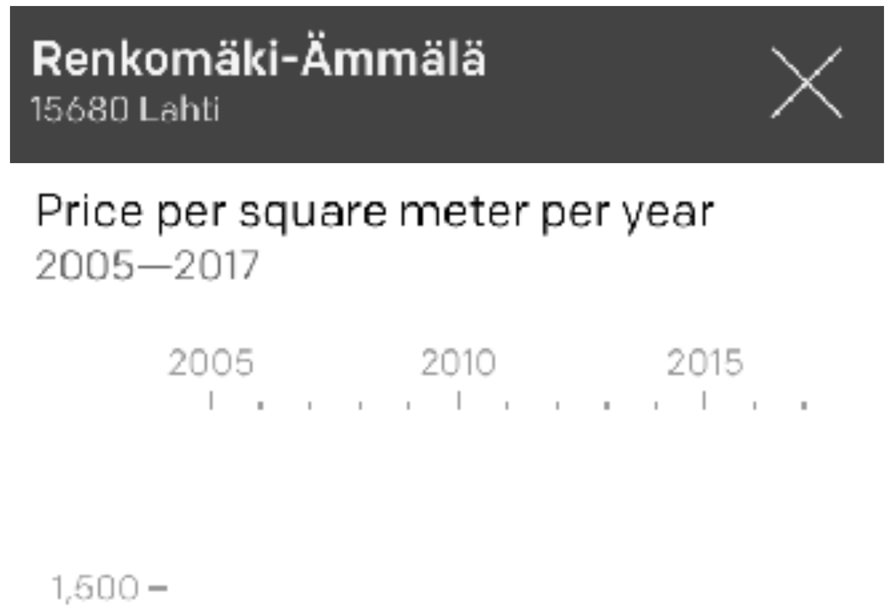


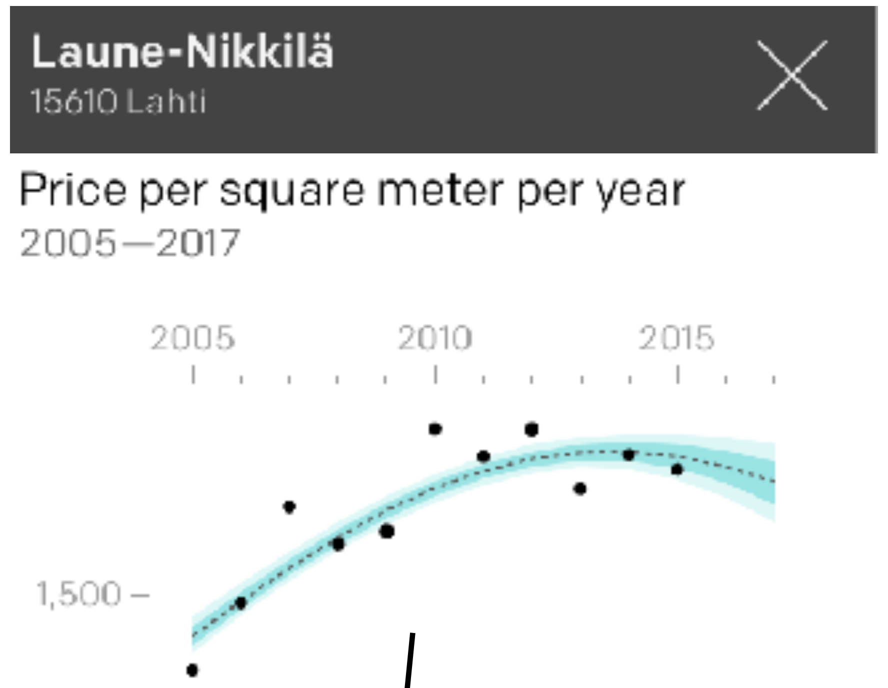


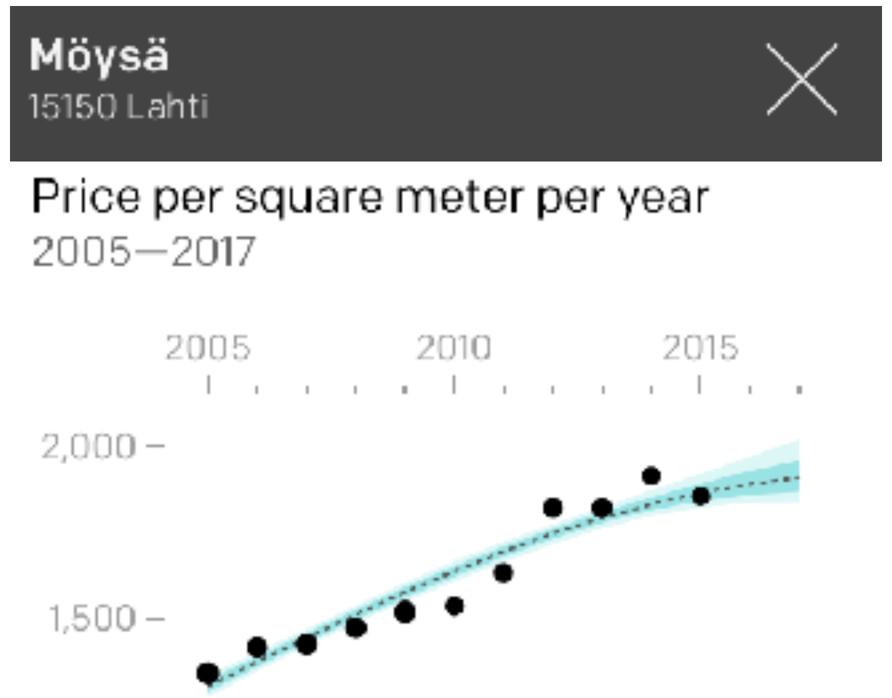
How to estimate the yearly prices for Renkomäki-Ämmälä?

Could borrow information from the adjacent ZIP codes.





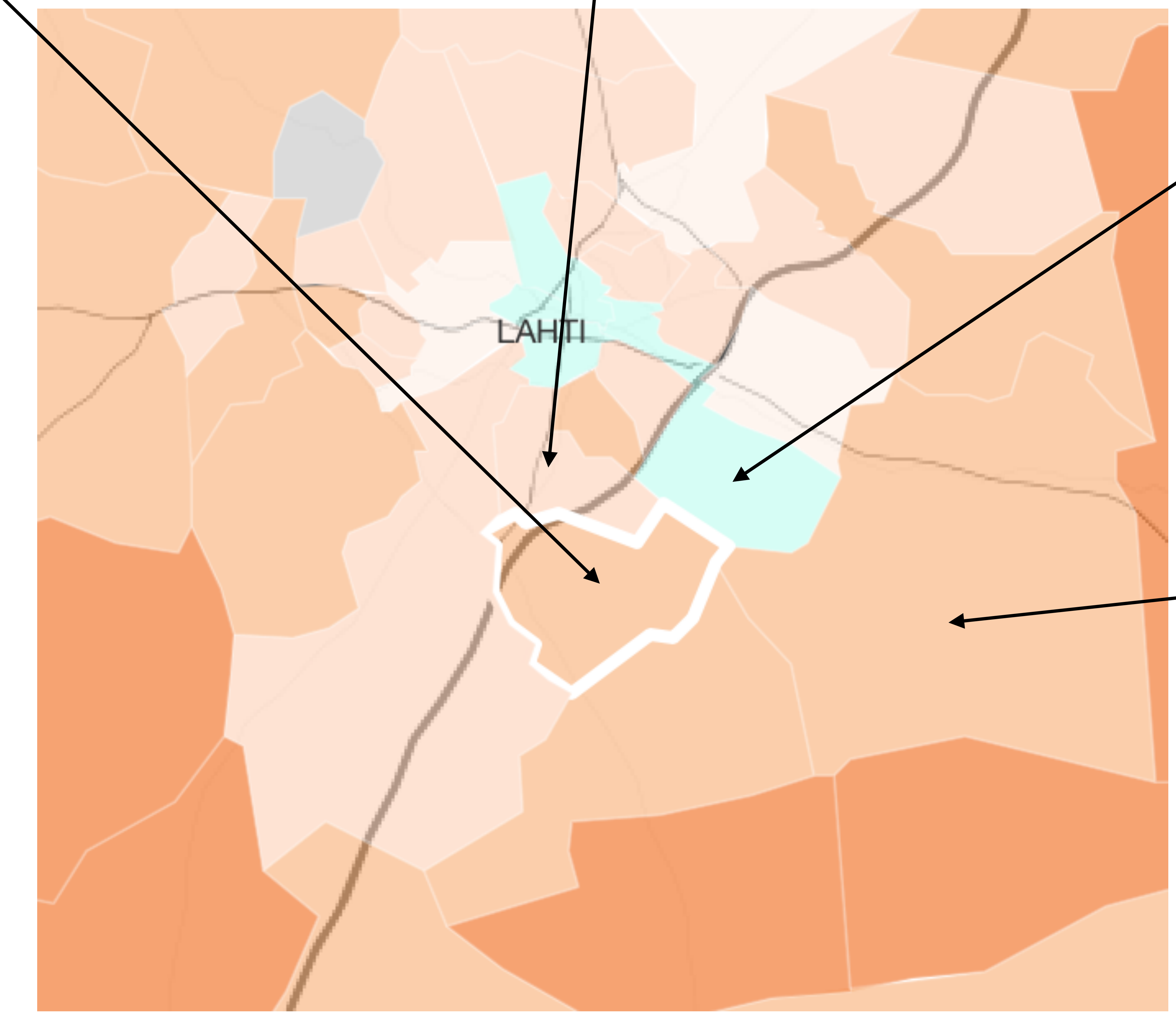


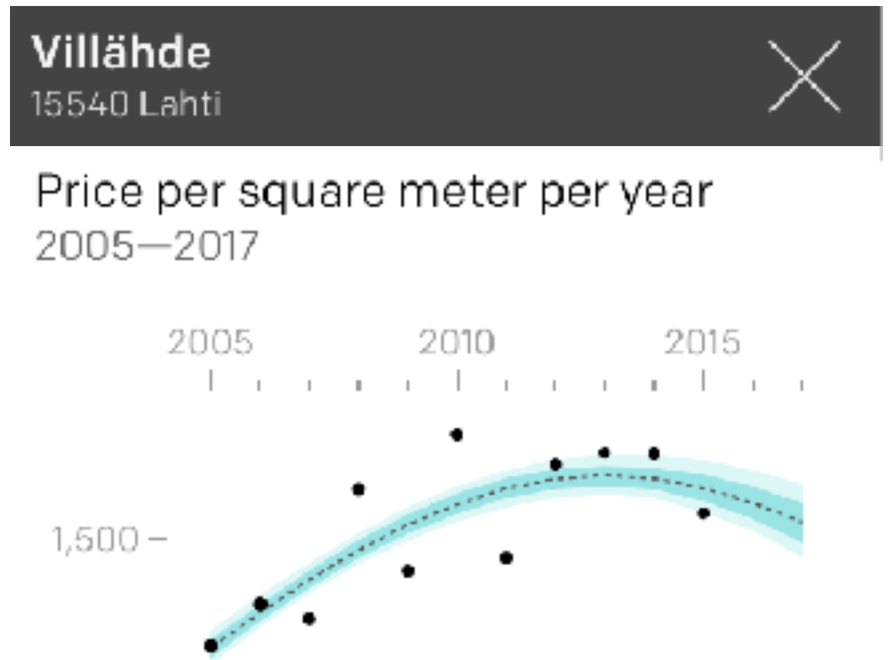


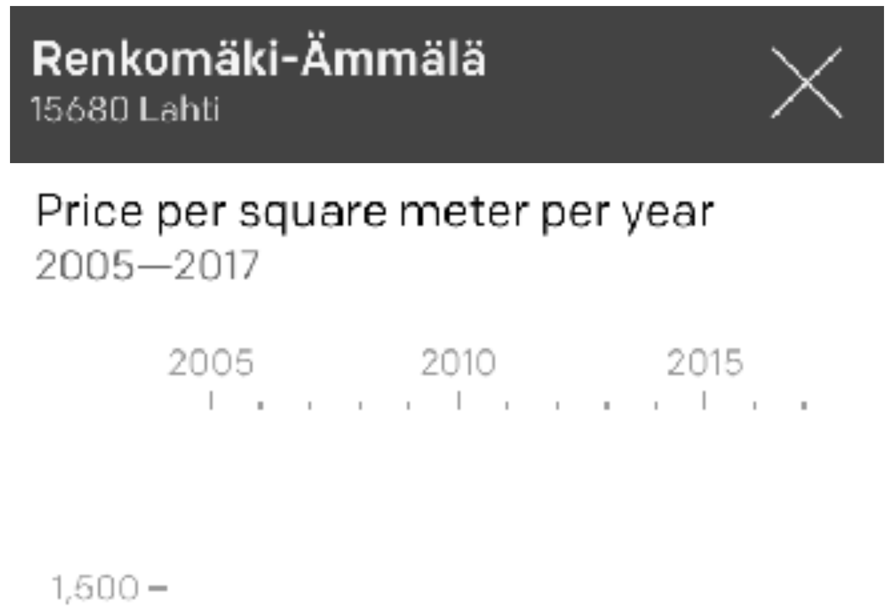
How to estimate the yearly prices for Renkomäki-Ämmälä?

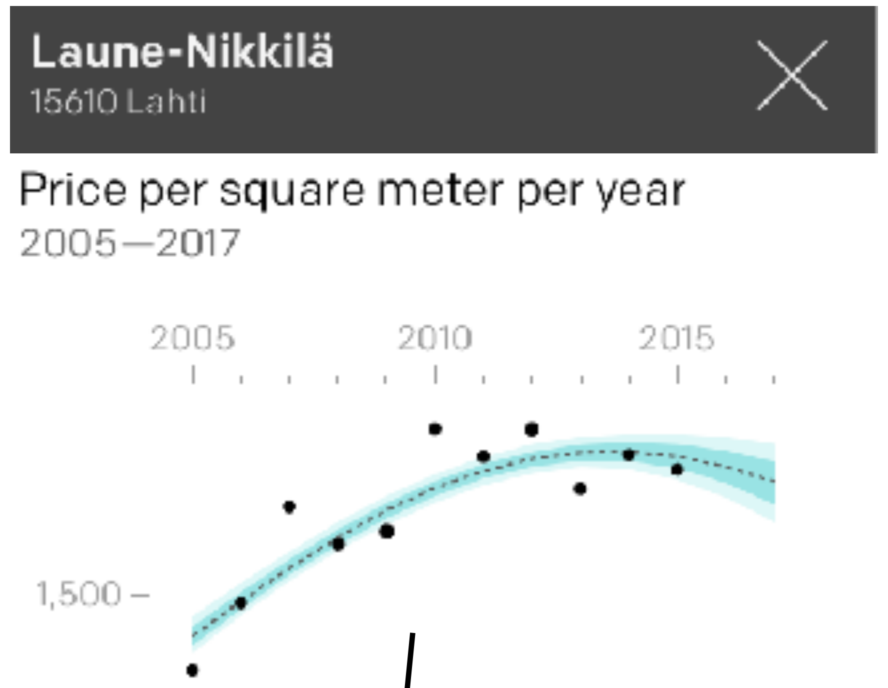
Could borrow information from the adjacent ZIP codes.

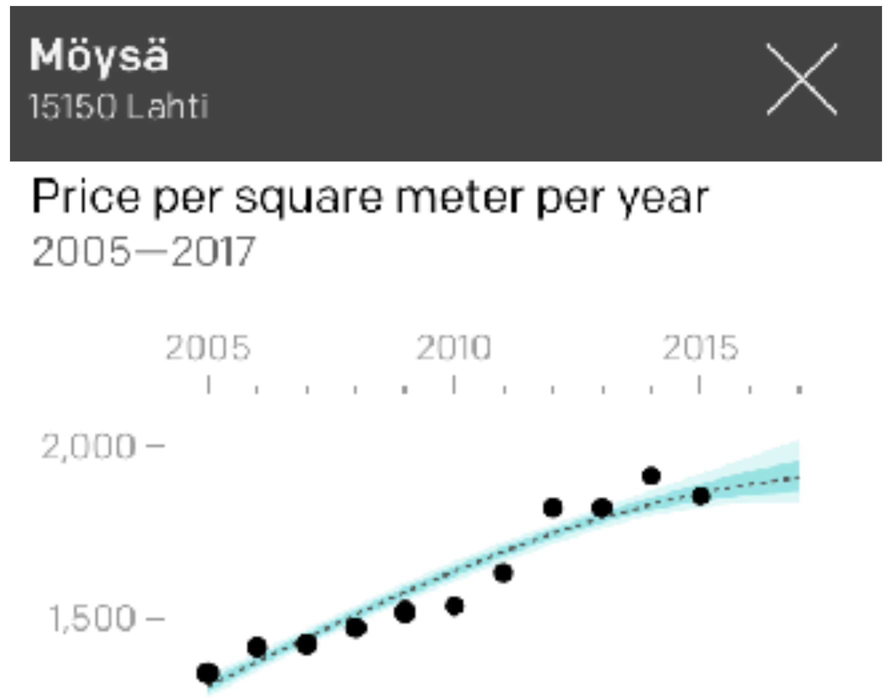
But how to choose from contradicting information (e.g. increasing prices in Möysä, and decreasing prices in Laune-Nikkilä)?









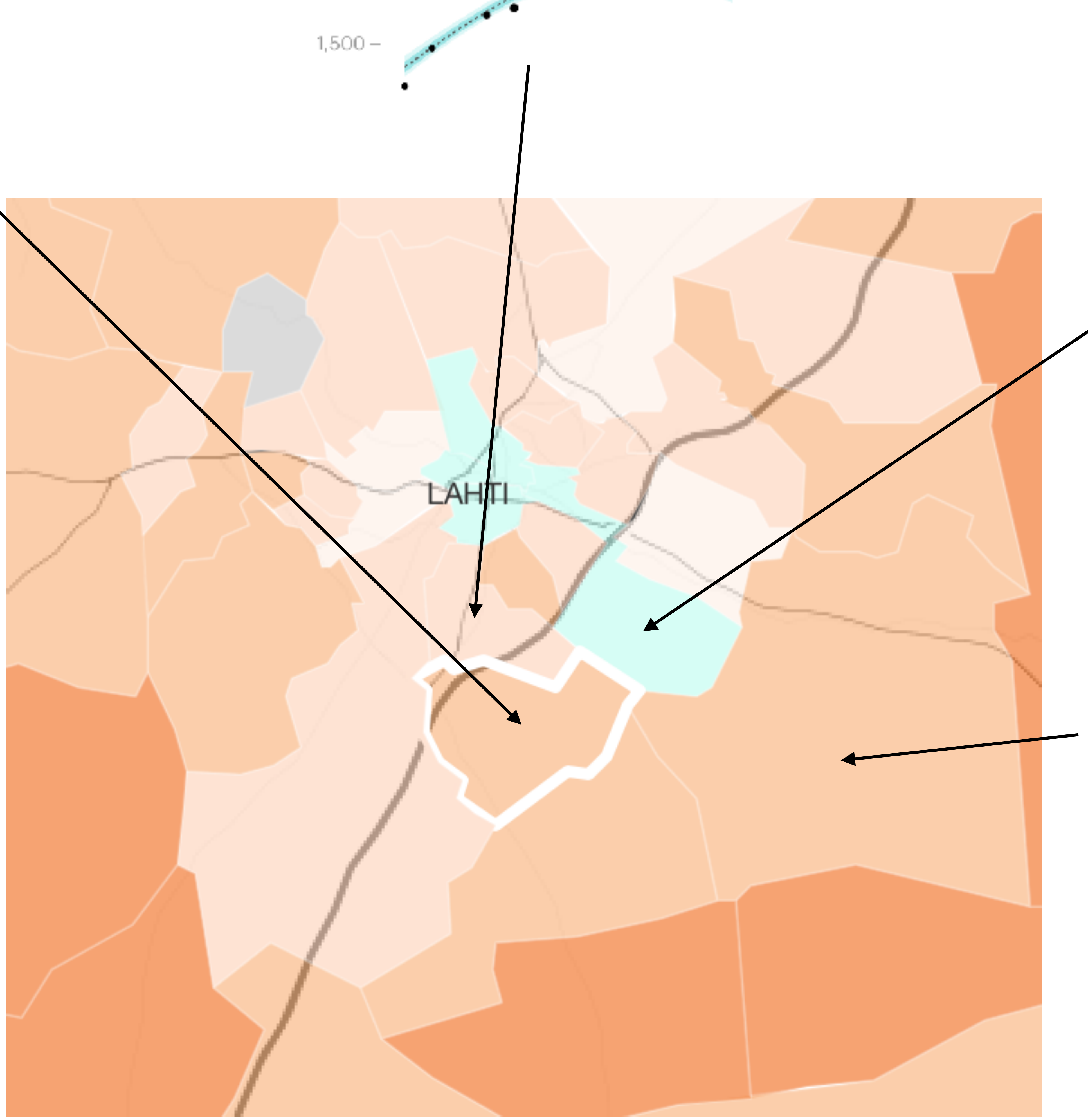


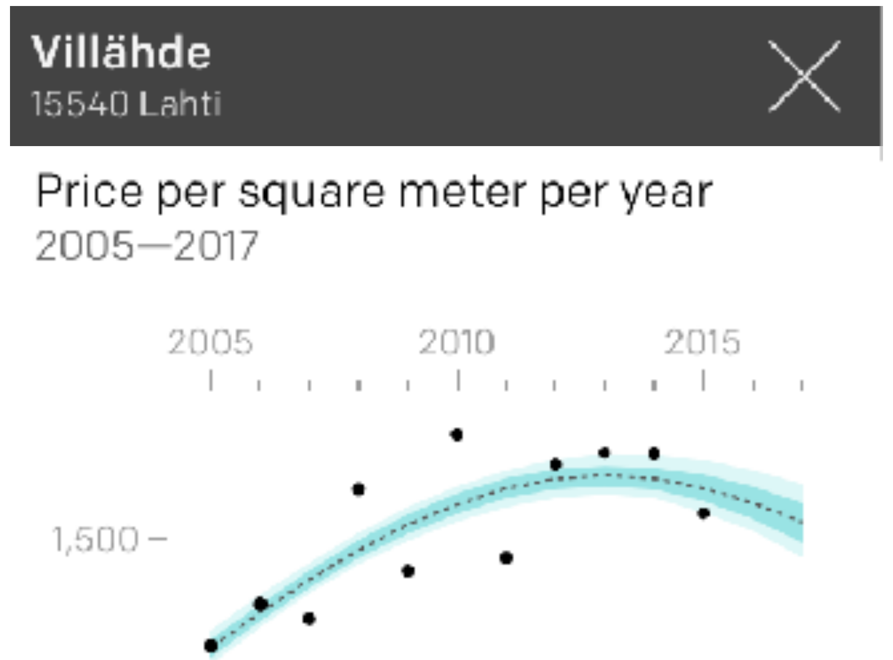
How to estimate the yearly prices for Renkomäki-Ämmälä?

Could borrow information from the adjacent ZIP codes.

But how to choose from contradicting information (e.g. increasing prices in Möysä, and decreasing prices in Laune-Nikkilä)?

Solution: use additional information (population density) about the similarity of the ZIP codes.





The model

Use a quadratic model for time.

Use three levels of hierarchies for the postal code:

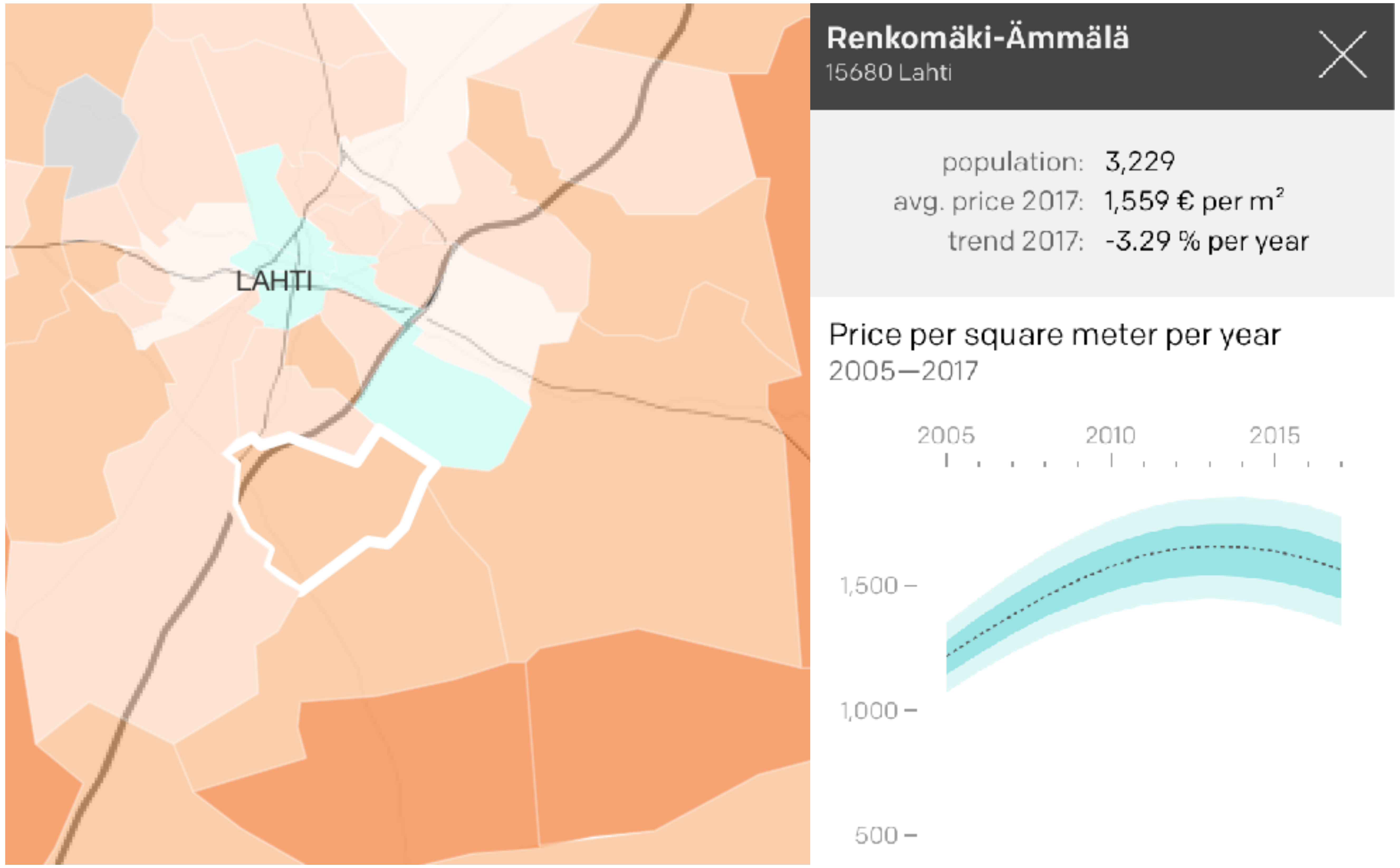
- 15680 (Renkomäki-Ämmälä)
- 156XX (Renkomäki-Ämmälä, Laune-Nikkilä)
- 15XXX (Renkomäki-Ämmälä, Laune-Nikkilä, Lahti, …)

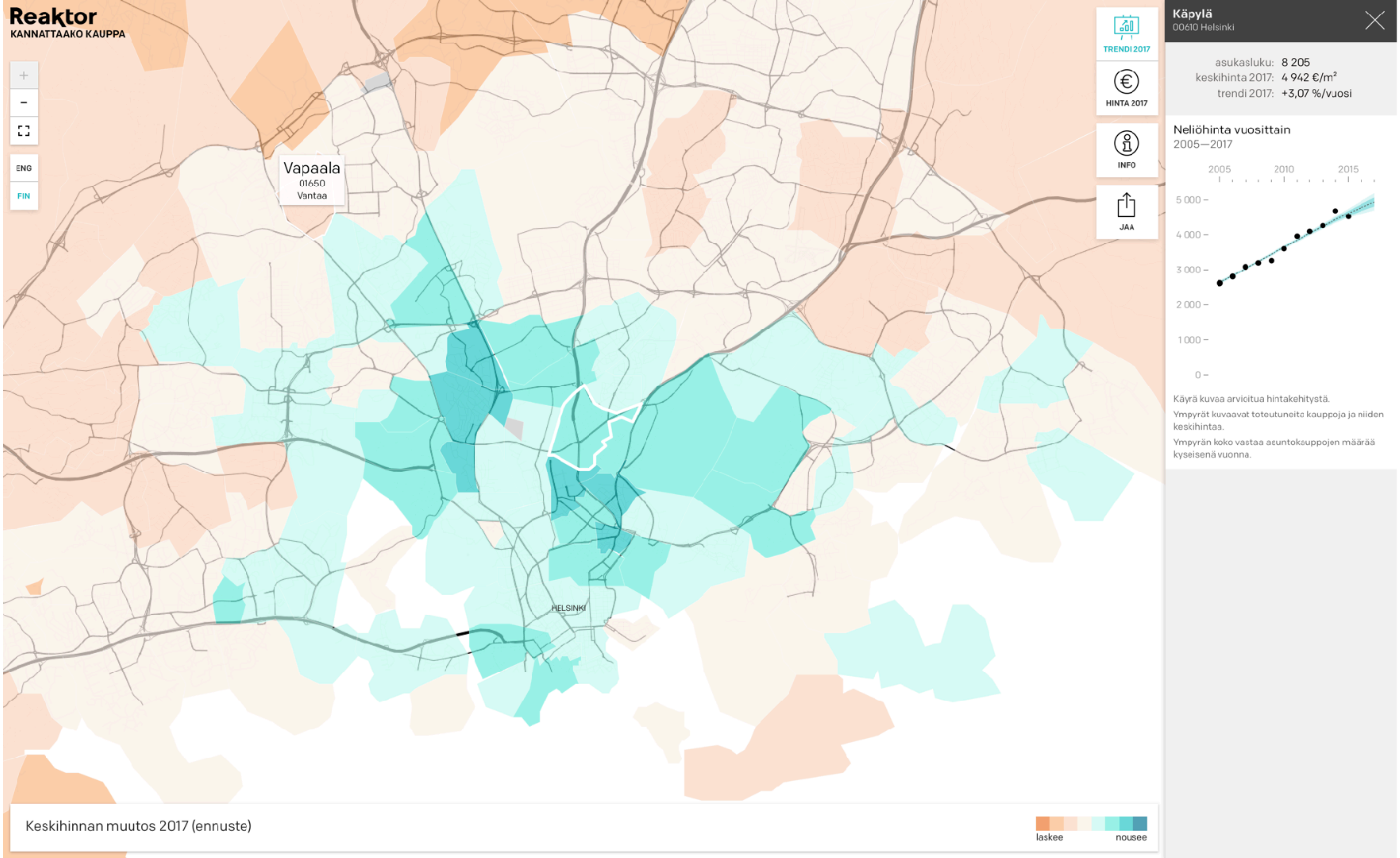
Use the population density of the 156XX.

The model on the lowest level of hierarchy:

$$\log h_{it} = \beta_{i1} + \beta_{i2} t + \beta_{i3} t^2 + \beta_{i'4} d_i + \beta_{i'5} d_i t + \beta_{i'6} d_i t^2,$$

$$\log y_{it} \sim \mathrm{t}\left(\log h_{it}, \sqrt{\sigma_y^2 + \frac{\sigma_w^2}{n_{it}}}, \nu\right),$$





Reaktor
KANNATTAAKO KAUPPA
ENG
FIN
Vapaala
01650
Vantaa
HELSINKI
TRENDI 2017
HINTA 2017
INFO
JAA
Käpylä
00610 Helsinki
asukasluku: 8 205
keskihinta 2017: 4 942 €/m²
trendi 2017: +3,07 %/vuosi
Neliöhinta vuosittain
2005—2017
2005
2010
2015
5 000 –
4 000 –
3 000 –
2 000 –
1 000 –
0 –
Käyrä kuvaa arvioitua hintakehitystä.
Ympyrät kuvaavat toteutuneita kauppoja ja niiden keskihintaa.
Ympyrän koko vastaa asuntokauppojen määrää kyseisenä vuonna.
Keskihinnan muutos 2017 (ennuste)
laskee
nousee