Housing price indices for small areas

NBP International Workshop Recent trends in Real Estate Market ant its Analysis

M.J. Bárcena, C. González, P. Menéndez and F. Tusell

Zalesie Górne, December, 5, 2019

• Background and motivation

- Background and motivation
- Housing market price indices

- Background and motivation
- Housing market price indices
- Our previous work and evolution to local indices

- Background and motivation
- Housing market price indices
- · Our previous work and evolution to local indices
- Examples

 Ever since the housing market bubble burst in 2007, banks have been far more reluctant to grant credit.

- Ever since the housing market bubble burst in 2007, banks have been far more reluctant to grant credit.
- Depressed salaries do not help either to gain access to credit.

- Ever since the housing market bubble burst in 2007, banks have been far more reluctant to grant credit.
- Depressed salaries do not help either to gain access to credit.
- Many potential buyers have been diverted to the rental market.

- Ever since the housing market bubble burst in 2007, banks have been far more reluctant to grant credit.
- Depressed salaries do not help either to gain access to credit.
- Many potential buyers have been diverted to the rental market.
- As a result, rental prices have increased enormously.

- Ever since the housing market bubble burst in 2007, banks have been far more reluctant to grant credit.
- Depressed salaries do not help either to gain access to credit.
- Many potential buyers have been diverted to the rental market.
- As a result, rental prices have increased enormously.
- Increased information needs for policy making.

000

• 2011: index of housing prices in Bilbao. Sample: ≈ 7000 observations, web scrapped.

0.00

- 2011: index of housing prices in Bilbao. Sample: ≈ 7000 observations, web scrapped.
- 2017: index of housing prices for Basque Country: over 400,000 observations.

• 2011: index of housing prices in Bilbao. Sample: ≈ 7000 observations, web scrapped.

000

- 2017: index of housing prices for Basque Country: over 400,000 observations.
- 2019: propose methodology to produce rental price indices for very small areas.

- 2011: index of housing prices in Bilbao. Sample: ≈ 7000 observations, web scrapped.
- 2017: index of housing prices for Basque Country: over 400,000 observations.
- 2019: propose methodology to produce rental price indices for very small areas.
- Source of information: data on bails deposited and rental offers from real state webs; preliminary work is being done with sales data.

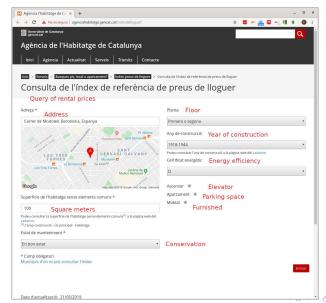
- 2011: index of housing prices in Bilbao. Sample: ≈ 7000 observations, web scrapped.
- 2017: index of housing prices for Basque Country: over 400,000 observations.
- 2019: propose methodology to produce rental price indices for very small areas.
- Source of information: data on bails deposited and rental offers from real state webs; preliminary work is being done with sales data.
- Spatial detail of price indices down to areas a few hundred meters wide..

- 2011: index of housing prices in Bilbao. Sample: ≈ 7000 observations, web scrapped.
- 2017: index of housing prices for Basque Country: over 400,000 observations.
- 2019: propose methodology to produce rental price indices for very small areas.
- Source of information: data on bails deposited and rental offers from real state webs; preliminary work is being done with sales data.
- Spatial detail of price indices down to areas a few hundred meters wide..
- A "rent advisor" needs to be programmed, feeding from the same data.



BACKGROUND

Interface of a web consulting tool





Laspeyres index

• Index at t base 0 defined as

$$I_0(t) = \frac{\sum_{i=1}^k p_{it} q_{i0}}{\sum_{i=1}^k p_{i0} q_{i0}} \times 100$$

Laspeyres index

Index at t base 0 defined as

$$I_0(t) = \frac{\sum_{i=1}^k p_{it} q_{i0}}{\sum_{i=1}^k p_{i0} q_{i0}} \times 100$$

• Compares the cost of the same basket of goods $\{q_{10}, \ldots, q_{k0}\}$ at prices at 0 and t.

Laspeyres index

• Index at t base 0 defined as

$$I_0(t) = \frac{\sum_{i=1}^k p_{it} q_{i0}}{\sum_{i=1}^k p_{i0} q_{i0}} \times 100$$

- Compares the cost of the same basket of goods $\{q_{10}, \ldots, q_{k0}\}$ at prices at 0 and t.
- We need to observe the prices of all products at 0 and t.

Housing price indices

 Above approach clearly unfeasible: traded houses are non-replicable, essentially unique.

Housing price indices

- Above approach clearly unfeasible: traded houses are non-replicable, essentially unique.
- No "basket" of houses is repeatedly traded at all times to allow computation of a Laspeyres-type index.

Housing price indices

- Above approach clearly unfeasible: traded houses are non-replicable, essentially unique.
- No "basket" of houses is repeatedly traded at all times to allow computation of a Laspeyres-type index.
- Plus: opaque market, prices difficult to observe, expensive and slow to survey.

• Use only **repeated sales** of same houses

- Use only **repeated sales** of same houses
- For houses i = 1, ..., n for which two sales are observed, fit

$$\log(P_{it_k}) - \log(P_{it_0}) = \sum_{i=0}^k \gamma_{t_i} + \epsilon$$

- Use only repeated sales of same houses
- For houses i = 1, ..., n for which two sales are observed, fit

$$\log(P_{it_k}) - \log(P_{it_0}) = \sum_{j=0}^k \gamma_{t_j} + \epsilon$$

• $\sum_{j=0}^{k} \gamma_{t_j}$ can be replaced for any function $\gamma(t)$.

- Use only repeated sales of same houses
- For houses i = 1, ..., n for which two sales are observed, fit

$$\log(P_{it_k}) - \log(P_{it_0}) = \sum_{j=0}^k \gamma_{t_j} + \epsilon$$

- $\sum_{j=0}^{k} \gamma_{t_j}$ can be replaced for any function $\gamma(t)$.
- Same houses, same quality, same location: difference of prices attributable to time.

- Use only repeated sales of same houses
- For houses i = 1, ..., n for which two sales are observed, fit

$$\log(P_{it_k}) - \log(P_{it_0}) = \sum_{j=0}^k \gamma_{t_j} + \epsilon$$

- $\sum_{j=0}^{k} \gamma_{t_j}$ can be replaced for any function $\gamma(t)$.
- Same houses, same quality, same location: difference of prices attributable to time.
- Index is then

$$I_{t_0}(t_k) = \exp\left(\sum_{j=0}^k \hat{\gamma}_{t_j}
ight) imes 100$$

$$\log(P_{it}) = \sum_{k} \beta_k x_{ik} + \gamma(t) + \epsilon$$

• Account for differences in quality with **hedonic model**:

$$\log(P_{it}) = \sum_{k} \beta_k x_{ik} + \gamma(t) + \epsilon$$

• x_{ik} is attribute k of house i; β_k its implicit price base period.

$$\log(P_{it}) = \sum_{k} \beta_k x_{ik} + \gamma(t) + \epsilon$$

- x_{ik} is attribute k of house i; β_k its implicit price base period.
- $\gamma(t)$ is the effect of time, can be used to construct index.

$$\log(P_{it}) = \sum_{k} \beta_k x_{ik} + \gamma(t) + \epsilon$$

- x_{ik} is attribute k of house i; β_k its implicit price base period.
- $\gamma(t)$ is the effect of time, can be used to construct index.
- But the price of an attribute (e.g. garage) is not the same everywhere...

$$\log(P_{it}) = \sum_{k} \beta_k x_{ik} + \gamma(t) + \epsilon$$

- x_{ik} is attribute k of house i; β_k its implicit price base period.
- $\gamma(t)$ is the effect of time, can be used to construct index.
- But the price of an attribute (e.g. garage) is not the same everywhere...
- ...and we need much more information (attributes of each hose)

Geographically Weighted Regression (GWR)

• We can fit the model **locally**, with implicit prices different at each location (u_i, v_i) :

$$\log(P_{it}) = \sum_{k} \beta_k(u_i, v_i) x_{ik} + \gamma(t) + \epsilon$$

Geographically Weighted Regression (GWR)

• We can fit the model **locally**, with implicit prices different at each location (u_i, v_i) :

$$\log(P_{it}) = \sum_{k} \beta_k(u_i, v_i) x_{ik} + \gamma(t) + \epsilon$$

• Now, the price of attribute k is $\beta_k(u_i, v_i)$, dependent on the coordinates (u_i, v_i) .

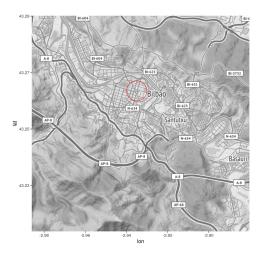
Geographically Weighted Regression (GWR)

• We can fit the model **locally**, with implicit prices different at each location (u_i, v_i) :

$$\log(P_{it}) = \sum_{k} \beta_k(u_i, v_i) x_{ik} + \gamma(t) + \epsilon$$

- Now, the price of attribute k is $\beta_k(u_i, v_i)$, dependent on the coordinates (u_i, v_i) .
- With a spline replaced for $\gamma(t)$ this was the model in Barcena et al.(2013).

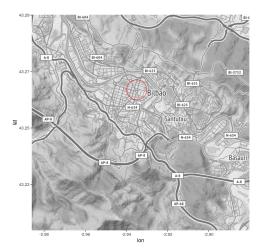
How do we fit the model? (I)



 Pretend you have prices at t = 0 and fit the GWR part:

$$\log(P_{it}) = \sum_{k} \beta_k(u_i, v_i) x_{ik} + \epsilon$$

How do we fit the model? (I)



000000000

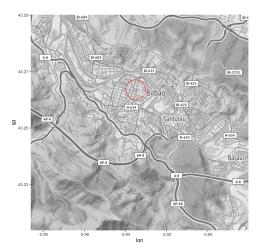
Pretend you have prices at t = 0 and fit the GWR part:

$$\log(P_{it}) = \sum_{k} \beta_k(u_i, v_i) x_{ik} + \epsilon$$

Compute residuals

$$\hat{\epsilon}_{it} = \log(P_{it}) - \sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}$$

How do we fit the model? (I)



 Pretend you have prices at t = 0 and fit the GWR part:

$$\log(P_{it}) = \sum_{k} \beta_k(u_i, v_i) x_{ik} + \epsilon$$

Compute residuals

$$\hat{\epsilon}_{it} = \log(P_{it}) - \sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}$$

Compute

$$\hat{\gamma}(t) = \arg\min\sum_{i,t} \left(\hat{\epsilon}_{it} - \gamma(t)\right)^2$$

How do we fit the model? (II)

• Iterate now (1)–(2)–(3) until convergence:

$$\log(P_{it}) - \hat{\gamma}(t) = \sum_{k} \beta_k(u_i, v_i) x_{ik} + \epsilon$$
 (1)

$$\hat{\epsilon}_{it} = \log(P_{it}) - \sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}$$
 (2)

$$\hat{\gamma}(t) = \arg\min \left[\sum_{i,t} (\hat{\epsilon}_{it} - \gamma(t))^2 \right]$$
 (3)

How do we fit the model? (II)

• Iterate now (1)–(2)–(3) until convergence:

$$\log(P_{it}) - \hat{\gamma}(t) = \sum_{k} \beta_k(u_i, v_i) x_{ik} + \epsilon$$
 (1)

$$\hat{\epsilon}_{it} = \log(P_{it}) - \sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}$$
 (2)

$$\hat{\gamma}(t) = \arg\min \left[\sum_{i,t} (\hat{\epsilon}_{it} - \gamma(t))^2 \right]$$
 (3)

• In all cases we have tried the iteration invariable converges.

• A fitted model at each location (u_i, v_i) of the form

$$\log(P_{it}) = \underbrace{\sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}}_{\text{Base price}} + \hat{\gamma}(t) + \hat{\epsilon}$$

• A fitted model at each location (u_i, v_i) of the form

$$\log(P_{it}) = \underbrace{\sum_{k} \hat{eta}_{k}(u_{i}, v_{i}) x_{ik}}_{ ext{Base price}} + \hat{\gamma}(t) + \hat{\epsilon}$$

• The $\hat{\beta}_k(u_i, v_i)$ vary in space, not in time; $\hat{\gamma}(t)$ is global.

• A fitted model at each location (u_i, v_i) of the form

$$\log(P_{it}) = \underbrace{\sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}}_{ ext{Base price}} + \hat{\gamma}(t) + \hat{\epsilon}$$

- The $\hat{\beta}_k(u_i, v_i)$ vary in space, not in time; $\hat{\gamma}(t)$ is global.
- Index can be calculated as

$$I_{t_0}(t) = \frac{\exp[\hat{\gamma}(t)]}{\exp[\hat{\gamma}(t_0)]} \times 100$$

• A fitted model at each location (u_i, v_i) of the form

$$\log(P_{it}) = \underbrace{\sum_{k} \hat{eta}_{k}(u_{i}, v_{i}) x_{ik}}_{ ext{Base price}} + \hat{\gamma}(t) + \hat{\epsilon}$$

- The $\hat{\beta}_k(u_i, v_i)$ vary in space, not in time; $\hat{\gamma}(t)$ is global.
- Index can be calculated as

$$I_{t_0}(t) = \frac{\exp[\hat{\gamma}(t)]}{\exp[\hat{\gamma}(t_0)]} \times 100$$

• By-product: shadow prices $\hat{eta}_k(u_i,v_i)$ at all locations and $t=t_0$

The gory details

$$\log(P_{it}) = \underbrace{\sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}}_{\text{Base price}} + \hat{\gamma}(t) + \hat{\epsilon}$$

• GWR. Radius around (u_i, v_I) from 200m to 500m. Can be set by cross-validation (but quite expensive).

The gory details

$$\log(P_{it}) = \underbrace{\sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}}_{ ext{Base price}} + \hat{\gamma}(t) + \hat{\epsilon}$$

- GWR. Radius around (u_i, v_I) from 200m to 500m. Can be set by cross-validation (but quite expensive).
- Alternative: variable radius, k nearest neighbours.

The gory details

$$\log(P_{it}) = \underbrace{\sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}}_{\text{Base price}} + \hat{\gamma}(t) + \hat{\epsilon}$$

- GWR. Radius around (u_i, v_I) from 200m to 500m. Can be set by cross-validation (but quite expensive).
- Alternative: variable radius, k nearest neighbours.
- $\hat{\gamma}(t)$ "smooth"; in our use, a cubic spline (set of cubic polynomials which connect smoothly). Using 1 df per year of data, more than enough.

• Ideally, we might fit GWR's for $t = t_0, \dots, t_k$:

$$\log(P_{it}) = \sum_{k} \hat{\beta}_{k,t}(u_i, v_i) x_{ik} + \hat{\epsilon}$$

• Ideally, we might fit GWR's for $t = t_0, \dots, t_k$:

$$\log(P_{it}) = \sum_{k} \hat{\beta}_{k,t}(u_i, v_i) x_{ik} + \hat{\epsilon}$$

• Price a "median house" at i in times t₀ and t and compute:

$$I_{i,t_0}(t) = \frac{\sum_k \hat{\beta}_{k,t}(u_i, v_i) x_{ik}}{\sum_k \hat{\beta}_{k,t_0}(u_i, v_i) x_{ik}} \times 100$$

• Ideally, we might fit GWR's for $t = t_0, \dots, t_k$:

$$\log(P_{it}) = \sum_{k} \hat{\beta}_{k,t}(u_i, v_i) x_{ik} + \hat{\epsilon}$$

• Price a "median house" at *i* in times *t*₀ and *t* and compute:

$$I_{i,t_0}(t) = \frac{\sum_k \hat{\beta}_{k,t}(u_i, v_i) x_{ik}}{\sum_k \hat{\beta}_{k,t_0}(u_i, v_i) x_{ik}} \times 100$$

• ...i.e. Laspeyres' index with homogeneous attributes (rather than houses) priced in location i at t_0 and t.

• Ideally, we might fit GWR's for $t = t_0, \dots, t_k$:

$$\log(P_{it}) = \sum_{k} \hat{\beta}_{k,t}(u_i, v_i) x_{ik} + \hat{\epsilon}$$

• Price a "median house" at *i* in times *t*₀ and *t* and compute:

$$I_{i,t_0}(t) = \frac{\sum_k \hat{\beta}_{k,t}(u_i, v_i) x_{ik}}{\sum_k \hat{\beta}_{k,t_0}(u_i, v_i) x_{ik}} \times 100 \quad I_0(t) = \frac{\sum_k p_{kt} q_{k0}}{\sum_k p_{k0} q_{k0}} \times 100$$

• ...i.e. Laspeyres' index with homogeneous attributes (rather than houses) priced in location i at t_0 and t.

• Ideally, we might fit GWR's for $t = t_0, \dots, t_k$:

$$\log(P_{it}) = \sum_{k} \hat{\beta}_{k,t}(u_i, v_i) x_{ik} + \hat{\epsilon}$$

• Price a "median house" at *i* in times *t*₀ and *t* and compute:

$$I_{i,t_0}(t) = \frac{\sum_k \hat{\beta}_{k,t}(u_i, v_i) x_{ik}}{\sum_k \hat{\beta}_{k,t_0}(u_i, v_i) x_{ik}} \times 100 \quad I_0(t) = \frac{\sum_k p_{kt} q_{k0}}{\sum_k p_{k0} q_{k0}} \times 100$$

- ...i.e. Laspeyres' index with homogeneous attributes (rather than houses) priced in location i at t_0 and t.
- Heavy computation and data requirements.

Can our model be adapted to do something similar?

$$\log(P_{it}) = \underbrace{\sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}}_{ ext{Base price}} + \hat{\gamma}(t) + \hat{\epsilon}$$

• Want to have $\hat{\gamma}_i(t)$ at each chosen location, rather than a single $\hat{\gamma}(t)$.

Can our model be adapted to do something similar?

$$\log(P_{it}) = \underbrace{\sum_{k} \hat{\beta}_{k}(u_{i}, v_{i}) x_{ik}}_{ ext{Base price}} + \hat{\gamma}(t) + \hat{\epsilon}$$

- Want to have $\hat{\gamma}_i(t)$ at each chosen location, rather than a single $\hat{\gamma}(t)$.
- Could repeat the process for each area, but this would neglect information from neighbouring areas.

• Define $w_{ij} = \left(1 - \frac{d_{ij}}{b}\right)^2$ (or zero if $d_{ij} > b$ for chosen b).

• Define $w_{ij} = \left(1 - \frac{d_{ij}}{b}\right)^2$ (or zero if $d_{ij} > b$ for chosen b).

$$\log(P_{it}) - \hat{\gamma}(t) = \sum_{k} \beta_k(u_i, v_i) x_{ik} + \epsilon$$

$$\hat{\epsilon}_{it} = \log(P_{it}) - \sum_{k} \hat{\beta}_k(u_i, v_i) x_{ik}$$

$$\hat{\gamma} = \arg\min \left[\sum_{i} \sum_{t} (\hat{\epsilon}_{it} - \gamma(t))^2 \right]$$

• Define $w_{ij} = \left(1 - \frac{d_{ij}}{b}\right)^2$ (or zero if $d_{ij} > b$ for chosen b).

$$\log(P_{it}) - \hat{\gamma}(t) = \sum_{k} \beta_{k}(u_{i}, v_{i})x_{ik} + \epsilon$$

$$\hat{\epsilon}_{it} = \log(P_{it}) - \sum_{k} \hat{\beta}_{k}(u_{i}, v_{i})x_{ik}$$

$$\hat{\gamma}_{j}(t) = \arg\min\left[\sum_{i} \sum_{t} w_{ij} (\hat{\epsilon}_{it} - \gamma(t))^{2}\right]$$

• Define $w_{ij} = \left(1 - \frac{d_{ij}}{b}\right)^2$ (or zero if $d_{ij} > b$ for chosen b).

$$\log(P_{it}) - \hat{\gamma}(t) = \sum_{k} \beta_{k}(u_{i}, v_{i})x_{ik} + \epsilon$$

$$\hat{\epsilon}_{it} = \log(P_{it}) - \sum_{k} \hat{\beta}_{k}(u_{i}, v_{i})x_{ik}$$

$$\hat{\gamma}_{j}(t) = \arg\min\left[\sum_{i} \sum_{t} w_{ij} (\hat{\epsilon}_{it} - \gamma(t))^{2}\right]$$

• In computing $\hat{\gamma}_i(t)$, residuals close in space weighted more.

• The radius (bandwidth) around (u_i, v_j) is the distance within which we assume the hedonic $\beta_k(u, v)$ fairly stable.

- The radius (bandwidth) around (u_i, v_j) is the distance within which we assume the hedonic $\beta_k(u, v)$ fairly stable.
- The bandwidth b in $w_{ij} = \left(1 \frac{d_{ij}}{b}\right)^2$ is the radius of the region where we assume the time trend to be common.

- The radius (bandwidth) around (u_i, v_j) is the distance within which we assume the hedonic $\beta_k(u, v)$ fairly stable.
- The bandwidth b in $w_{ij} = \left(1 \frac{d_{ij}}{b}\right)^2$ is the radius of the region where we assume the time trend to be common.
- No need for both to be equal!

- The radius (bandwidth) around (u_i, v_j) is the distance within which we assume the hedonic $\beta_k(u, v)$ fairly stable.
- The bandwidth b in $w_{ij} = \left(1 \frac{d_{ij}}{b}\right)^2$ is the radius of the region where we assume the time trend to be common.
- No need for both to be equal!
- Computational effort similar for both the single trend and each local trend models.

- The radius (bandwidth) around (u_i, v_j) is the distance within which we assume the hedonic $\beta_k(u, v)$ fairly stable.
- The bandwidth b in $w_{ij} = \left(1 \frac{d_{ij}}{b}\right)^2$ is the radius of the region where we assume the time trend to be common.
- No need for both to be equal!
- Computational effort similar for both the single trend and each local trend models.
- ...but trying to set the two bandwidths by cross-validation very expensive.

 Have tested method on offered prices for sale, as data for bails not yet available with geocoding.

- Have tested method on offered prices for sale, as data for bails not yet available with geocoding.
- Area studied is Great Bilbao, a conurbation with ≈ 1 million people, almost 50% of Basque Country.

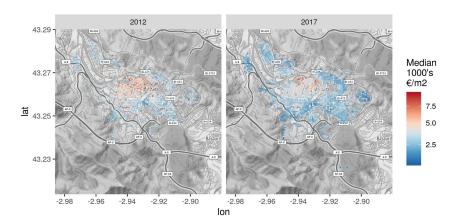
- Have tested method on offered prices for sale, as data for bails not yet available with geocoding.
- Area studied is Great Bilbao, a conurbation with ≈ 1 million people, almost 50% of Basque Country.
- Offered price data shown next for 2017; there are 34337 obs binned in squares about 65×65 meters.

- Have tested method on offered prices for sale, as data for bails not yet available with geocoding.
- Area studied is Great Bilbao, a conurbation with ≈ 1 million people, almost 50% of Basque Country.
- Offered price data shown next for 2017; there are 34337 obs binned in squares about 65×65 meters.
- Median offered price in thousands of euros color coded.

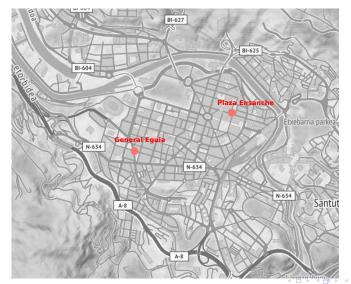
Great Bilbao. Offered prices (1000's €/m2)



Bilbao: 2012 vs 2017



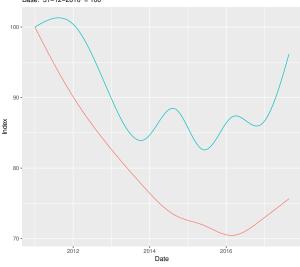
Choice of centroids to investigate different trends





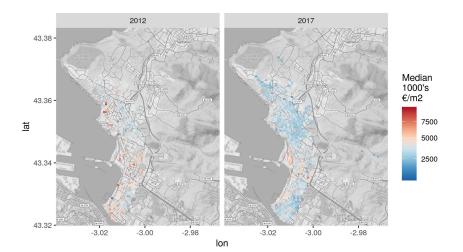
Local trends





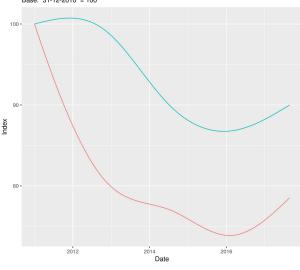
Location — GeneralEguia — PlazaEnsanche

Getxo: 2012 vs 2017



Local trends





Location - LasArenas - Neguri

 In R, several packages implement GWR. Among them, spgwr, GWmodel, gwrr, ScaGWR.

- In R, several packages implement GWR. Among them, spgwr, GWmodel, gwrr, ScaGWR.
- In R, several packages provide non-parametric smoothing, including mgcv and gam.

- In R, several packages implement GWR. Among them, spgwr, GWmodel, gwrr, ScaGWR.
- In R, several packages provide non-parametric smoothing, including mgcv and gam.
- Using above packages, it doesn't take much to implement the method described. It has been done in package ipv, in GitHub, https://github.com/FernandoTusell/ipv.git.

- In R, several packages implement GWR. Among them, spgwr, GWmodel, gwrr, ScaGWR.
- In R, several packages provide non-parametric smoothing, including mgcv and gam.
- Using above packages, it doesn't take much to implement the method described. It has been done in package ipv, in GitHub, https://github.com/FernandoTusell/ipv.git.
- Currently work in progress and mostly untranslated, but will evolve and improve.

- In R, several packages implement GWR. Among them, spgwr, GWmodel, gwrr, ScaGWR.
- In R, several packages provide non-parametric smoothing, including mgcv and gam.
- Using above packages, it doesn't take much to implement the method described. It has been done in package ipv, in GitHub, https://github.com/FernandoTusell/ipv.git.
- Currently work in progress and mostly untranslated, but will evolve and improve.
- Associated paper and slides also available, https://github.com/FernandoTusell/NBP2019.git.