Translating House Prices to Land Values - An exercise in Automated Land Valuation in Maricopa County, Arizona, USA

Risto Peltola

David C. Lincoln Fellowship Symposium: Measuring the Value of Land

January 28-29, 2021

Contents

Context

- Land leverage
- Land as a residual
- House price as an indicator of land values

The Maricopa exercise

- Data
- Models
- Maps
- Conclusions

Land leverage

P = L + C

LL = L/P = (P-C)/P

P = property price

L= land value

C = construction costs

D = depriciation

LL =land leverage (land share or property value)

LL varies a lot spatially and temporally.

Main implication: The higher LL the higher are cyclical variations in house prices.

Bourassa, Hoesli ea. 2010 refer the litterature:

Davis and Palumbo (2008) focus on 46 large metropolitan areas in the US from 1984 to 2004. For these areas, land leverage for single-family owner-occupied homes increased from an average of 32% in 1984 to 51% in 2004.

Case (2007) estimates land leverage for residential property in the US from 1975 through 2005.[2] Using data similar to those employed by Davis and Heathcote (2007), he produces quite different estimates of land leverage: about 14% in 1975 and 38% on 2005.

These and other calculated LL –figures are quite high. However Case calculates notably lower figures. Clearly it's not easy to make the calculation right.

Main focus of LL –litterature is analysis of housing markets, not the valuation of land as such.

Land as a residual

P=C−D+L

L = P-(C−D)

To my understanding the LL –figures were based on construction costs and depreceation. The methodology is very sensitive to the estimates of C and D.

Construction companies of course calculate their tender prices of land based on selling prices of houses and C.

In mass appraisal of land for taxation purposes the use of the residual method is complicated, and to my knowledge not widely used, if not used at all.

There is a danger of overvaluation of L given the uncertainties in C and D.

House price as an indicator of land values

- My approach is different. (This is also the approach when the Land Survey of Finland values 1 million housing lots for taxation purposes.)
- There is an implicit L in P.
- The implicit L is calculated stepwise: First all non-spatial components of P are controlled, a constant quality P is produced. Then CQP is used as an argument in an L model, and L as a function of P is obtained.
- A critical amount of land price transaction data is necessary. The price level depends solely on land price data.
- However the details of the value landscape are much more nuanced and intuitively convincing.
- This method seems to produce somewhat lower L values than the residual method.

Purpose of the paper

- The Maricopa real estate market and the institutional setting of real estate taxation in the county.
- The data, explored in tables and maps.
- Focus on land and residential real estate
- Hedonic models:
  - a land price model,
  - model for single family housing, and finally
  - a model of land price as a function of house prices.
- Maps of the value landscape.
- Candidates of value zones making use of clusters of price points of similar values

Maricopa real estate market and the institutional setting of real estate taxation

- Maricopa County is located in the south-central part of the U.S. state of Arizona. The U.S. Census Bureau estimated its population was 4,410,824 as of 2018, making it the state's most populous county, and the fourth most populous in the United States, containing more than half the population of Arizona.
- According to the U.S. Census Bureau, the county has a total area of 9,224 square miles (23,890 km2), mostly land.
- Maricopa County represents a textbook case of urban sprawl. This is an interesting case to value land, as the supply of land is rather elastic with plenty of land sales.

The tools

- For hedonic modelling SAS is used.
- For spatial analysis ArcGIS was used.

Translating house prices to land values

Home price is a function of spatial and non-spatial quality attributes

- $\text{Log (PRICE)} = \text{QUALITY}_{\text{SPATIAL}} + \text{QUALITY}_{\text{NON-SPATIAL}}$

$\text{QUALITY}_{\text{SPATIAL}}$ is the constant quality price of housing.

It is also an indicator to the value of land.

Step 1.

House price	=	Spatial components	+	Non-Spatial components

Step 2.

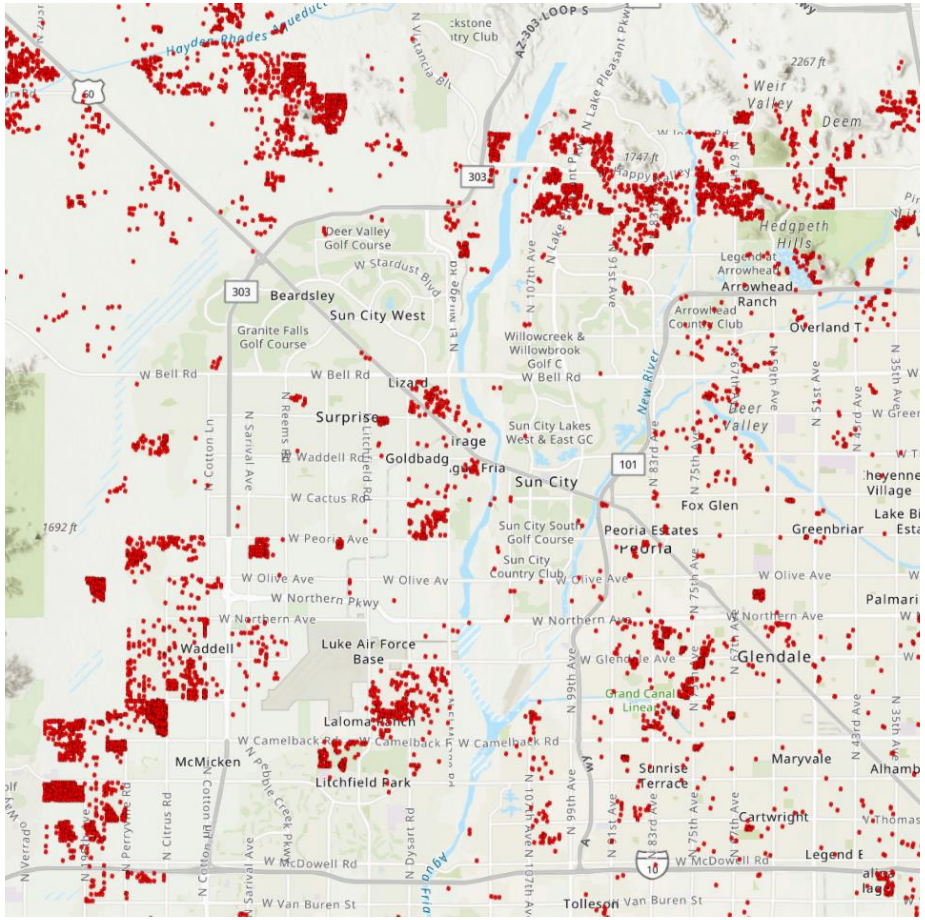
Non-Spatial components	=	House price	-	Spatial components (crude estimate)

Step 3.

House price As an indicator of land value	=	House price	-	Non-Spatial components

The data (land sales)

- A rich and high quality set of sales and attribute data is available on both land sales and sales of improved property.
- As a rule, there is no shortage of sales of improved property. The supply of land sale information varies a lot.

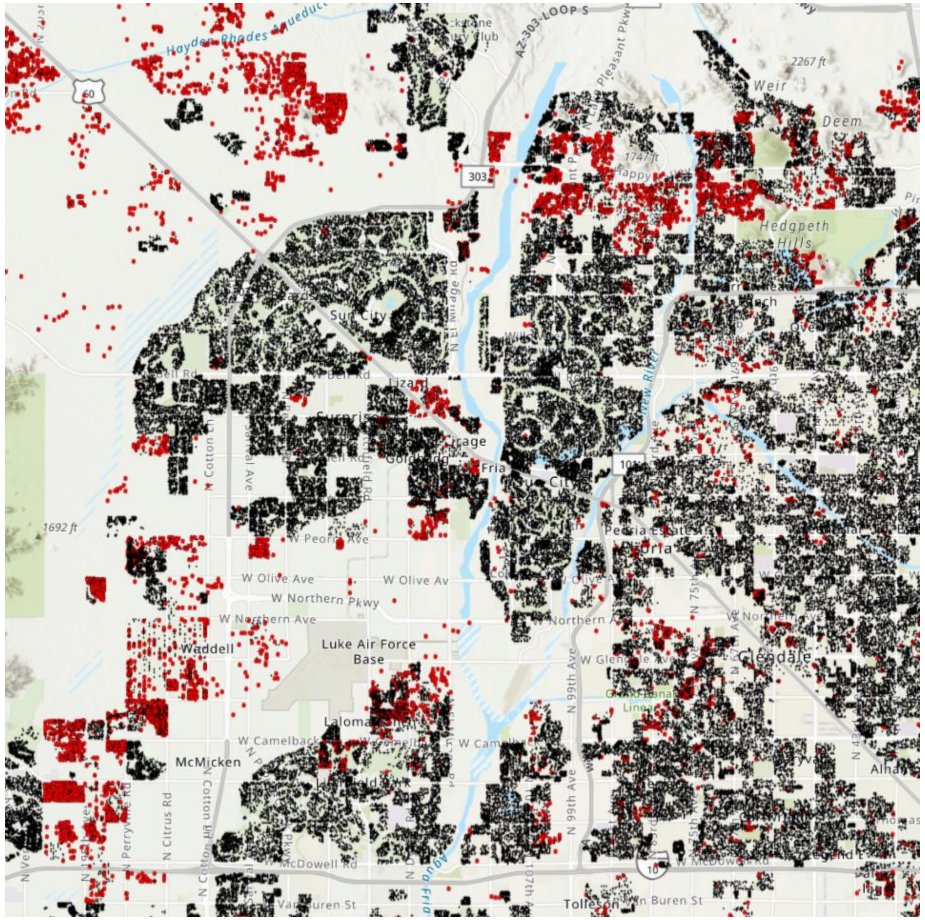


The data (land sales, improved sales)

Typical combinations are the following:

- Development outside the city (plenty of land sales, no improved sales)
- Development at the edge of the city (plenty of sales of both types)
- Large development area inside the city
- Scattered land sales inside the city (a few land sales, plenty of improved sales)
- No land sales at all in a large area

The last two types are most challenging



Land price model

The model is based on 87800 land sales and has 303 variables, R2= 0,84.

Land price
= quality (9 variables)
+ temporal control variables (30)
+ spatial control variables (264)

Land quality function

quality=
+llandsqft*-**0.288962703**
+golf***0.566198931**
+noroad*-**0.197607740**
+trans_ln*-**0.252460935**
+maj_int***0.098347290**
+flt_no*-**0.004538302**
+corner***0.223122837**
+arterial***0.031394566**
+grnbelt***0.332331214** ;

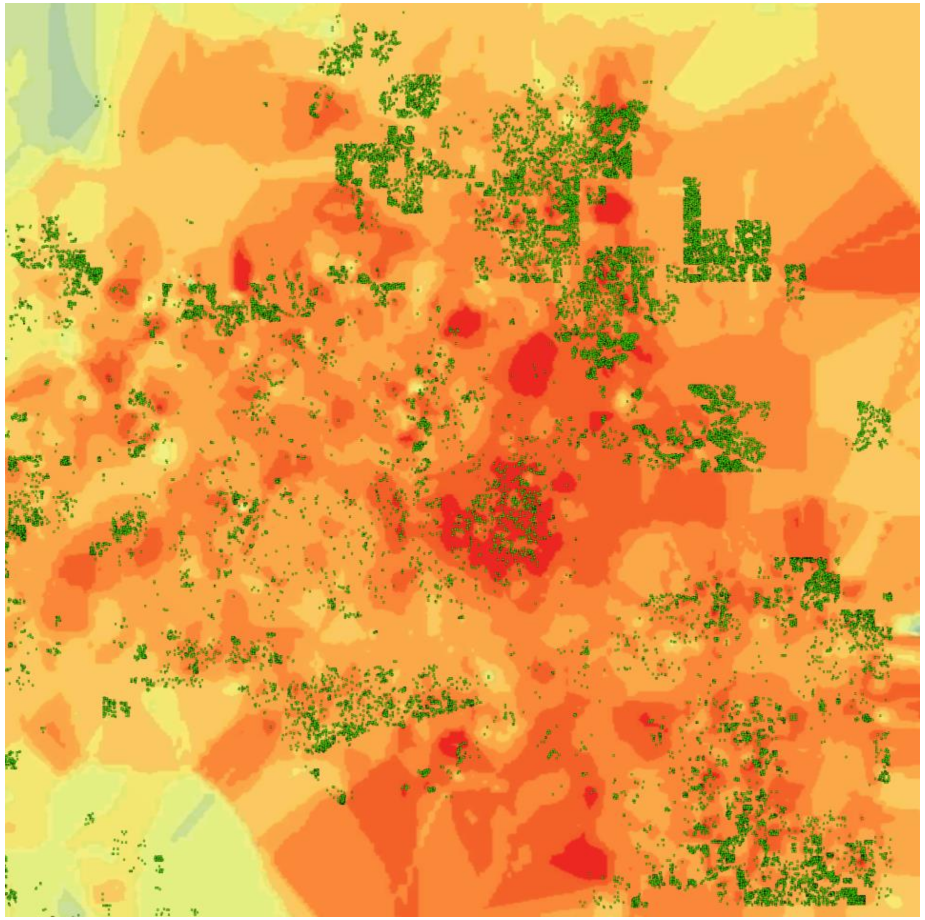
Constant land quality prices

Constant quality prices are simply the difference between unit price and quality, in logs.

CQ Land price:

luspricecq_land =lusprice _land - quality _land;

Figure 1. CQ land prices plotted on a map produces a first view of the value landscape.



Single family house price model

The model is based on 487592 house sales and has 548 variables, R2= 0,38.

House price

= quality (29 variables)

+ temporal control variables (14)

+ spatial control variables (505)

The model is much less efficient than the land price model. Part of the reason is the fact that in the period 2014-2018 the cyclical variation is not dominant. This may be a problem and its main reasons are the heterogeneity of the houses, missing variables and the fact that even inside certain neighborhoods the prices differ a lot.

The accuracy of the house price model, measured as root mean square error (RSE), is however better (lower) than in the land price models. This is typical with these models.

Single family house quality function

quality=	+ PADSITE * -0,044
+ age * -0,006	+ LAKE * 0,241
+ AGE40 * -0,01	+ ADJ_APT * 0,054
+ AGE20 * 0,001	+ MAJ_INT * -0,286
+ AGE10 * -0,005	+ UT_SEPTC * 0,007
+ lR_IMPTOTSQFT * -0,546	+ PRESERVE * 0,049
+ llot_ratio * -0,050	+ C_PERCENT_COMPLETE * -0,072
+ PERIMP * -4,576	+ AIR_PARK * 0,365
+ LANDSQFT * 0,000	+ ADJ_CM * 0,054
+ LONGITUDE * 0,242	+ UT_NONE * 0,003
+ LATITUDE * -0,001	+ GOLF * 0,163
+ GATED * 0,108	+ NOROAD * -0,015
+ UT_GAS * -0,021	+ TRANS_LN * -0,055
+ PREMIUM * 0,020	+ PERLAND * -0,092
+ MTN * 0,012	+ CULDESAC * -0,011

Constant quality prices

Constant quality prices are simply the difference between unit price and quality, in logs.

CQ Land price:

luspricecq_land =lusprice _land -

quality _land;

CQ House price:

luspricecq_improved

=lusprice _improved - quality _improved;

Translating house prices to land values

- I apply a simple method to translate house prices to land prices:
- I choose a subdivision where good quality price information is available on both land and house prices,
- Then I model the relationship assuming both prices accurately represent values at the same location.
- I neglect RNBHDs, I use grids.

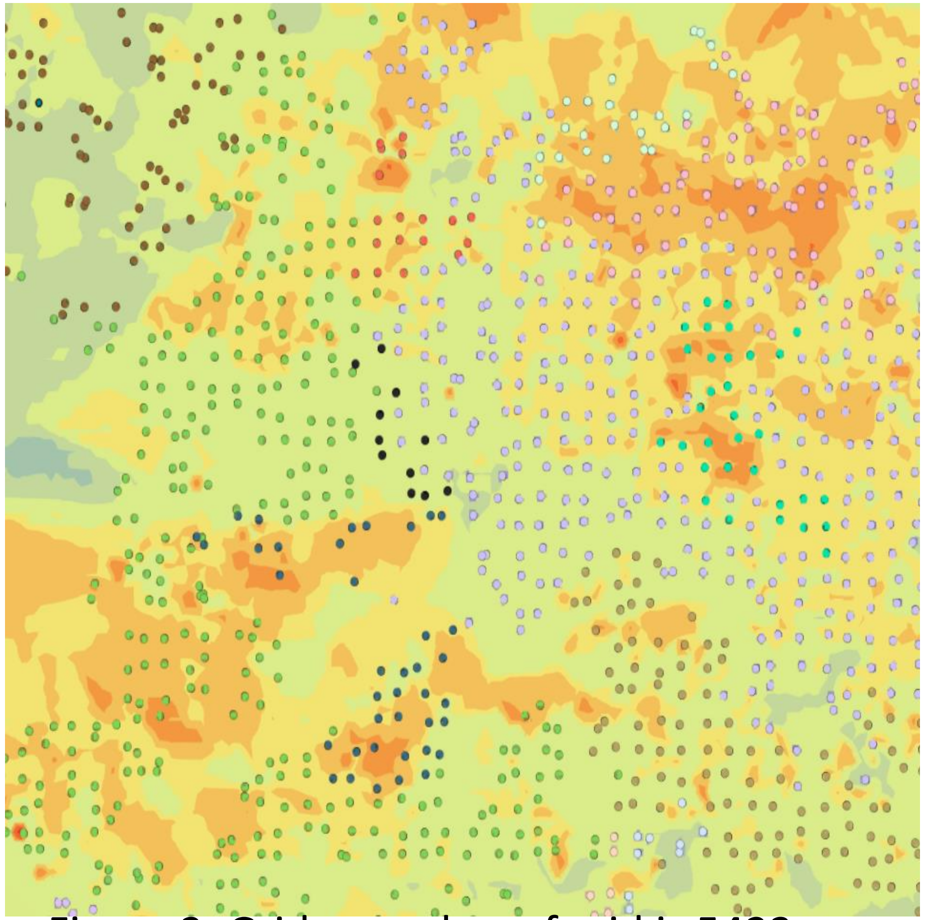


Figure 2. Grids, number of grid is 5423

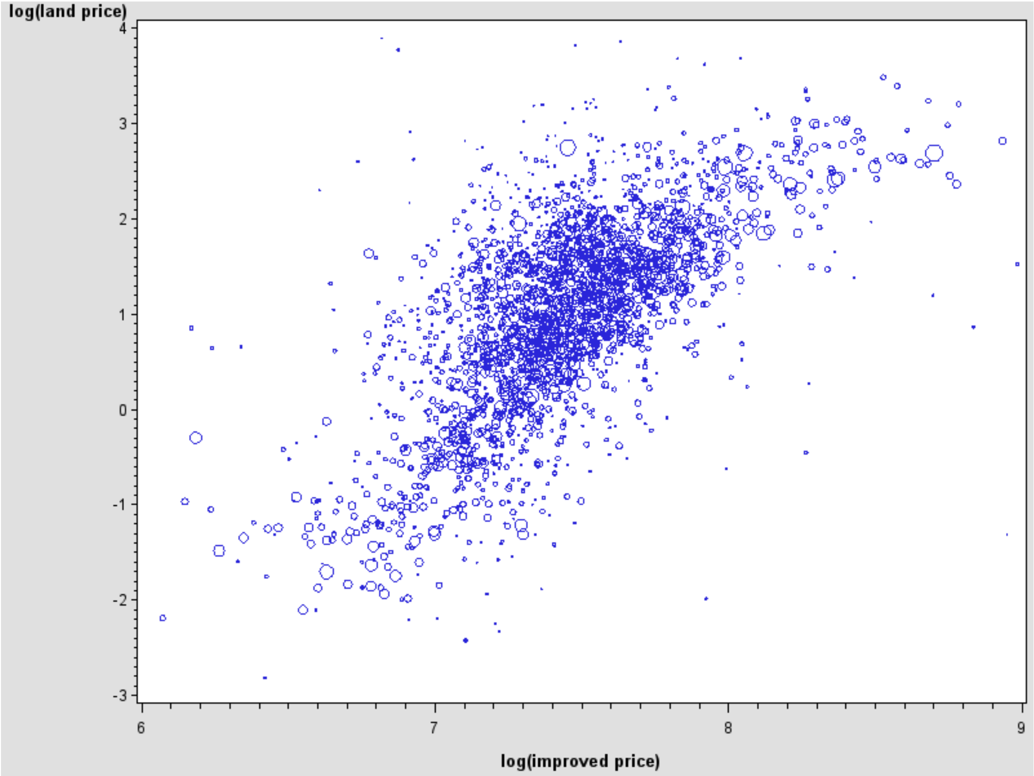
Descriptive statistics of grids where land and/or house values are available

Variable	Label	N	Min	Median	Mean	Std Dev	Max	Sum
CQ House price	unit price of a house,constant quality	3816	4,5	7,4	7,4	0,4	11,3	
CQ Land price	unit price of a land,constant quality	4142	-4,6	0,5	0,2	1,5	4,1	
LAND_SHARE	land share of a house price (%)	2583	0	3	8	30	1046	
lot_ratio	lot ratio (%)	3938	0	22	27	57	1038	
lLAND_SHARE	log(land share)	2583	-11	-3,4	-3,9	1,8	2,3	
llot_ratio	log(lot ratio)	3816	-2,8	3,2	2,8	1,1	6,6	
COUNT_IMPROVED	number of improved sales	3938	1	94	129	125	1445	507192
COUNT_LAND	number of land sales	4209	1	7	17	27	298	72593

CQ Land price as function of CQ house price.

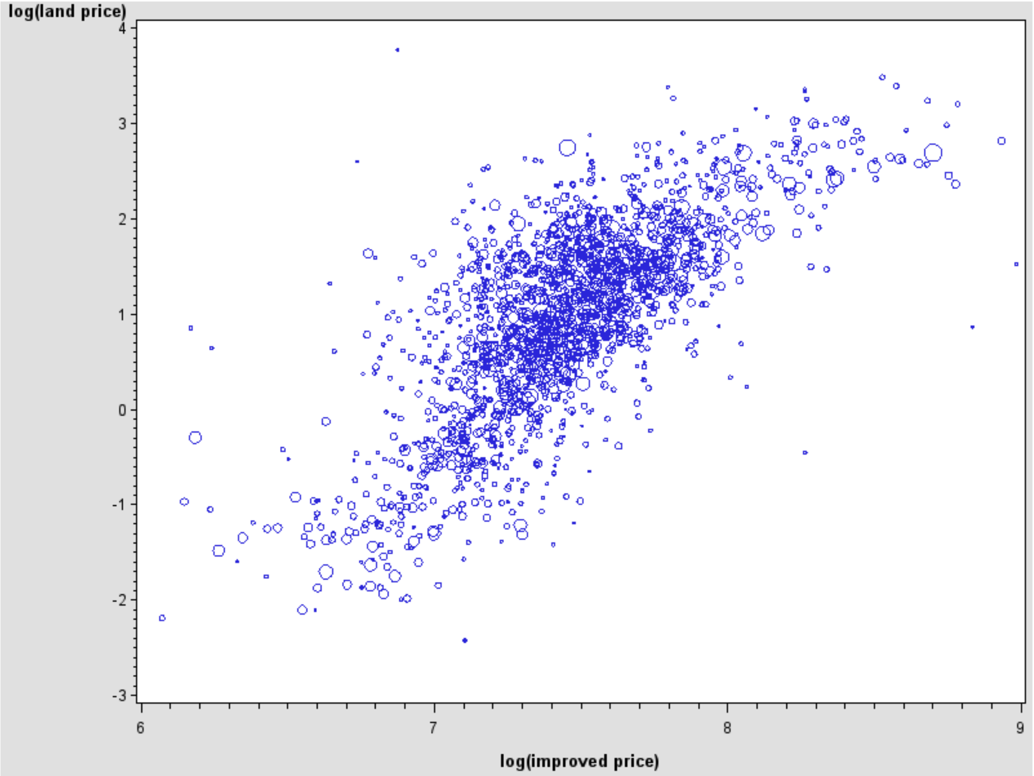
All grids with at least one sale

N=2568, R2= 0.39, slope=1.76



Grids with at least four land sales

N=1947, R2= 0.51, slope=1.98



A function that translates house prices to land values

CQ Land price= α+ β x CQ House price

The parameter 1,76 in model 1 is the elasticity of land values to house price. As house prices increase 1 %, the land values increase 1,76 %.

In model 2 there is no fixed elasticity level, but the elasticity depends in location. The higher the house prices, the lower the effect on land values.

Model 2 is theoretically justified and simple enough, but the gains in accuracy may be small. The choice is between models 1 and 3. If we want to keep lot ratios fixed we should choose model 1. However, we know the density depends on location.

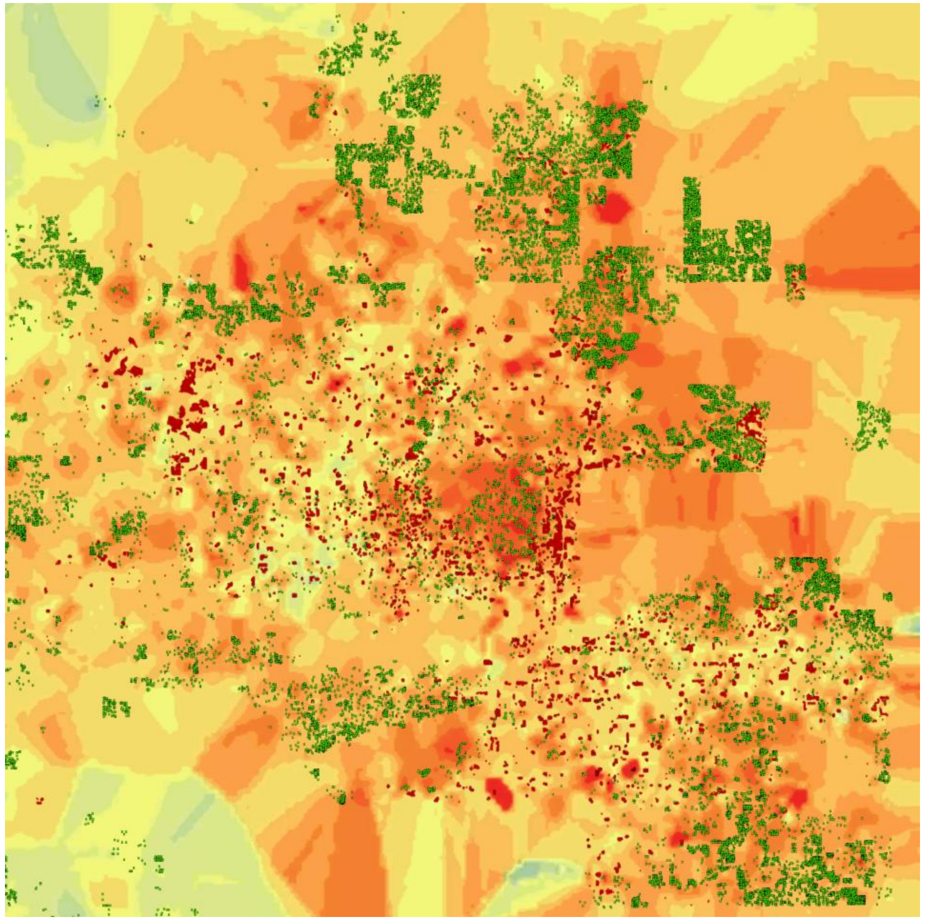
Even if we only have an indirect data on density we know the approximate effect of density to lot prices. In this exercise we use the land value estimates produced by model 1 for both reasons of simplicity and theory

Number of Observations Read	**5423**
Number of Observations Used	2583
Number of Observations with Missing Values	2840

	Model 1	Model 2	Model 3
R-Square	0,404	0,405	0,635
Intercept	-12,1	-17,0	-10,9
	(-38,56)	(-7,44)	(-44,11)
CQ House price	1,76	3,11	1,45
	(41,83)	(5,2)	(-42,67)
llot_ratio			0,5
			(-40,4)
CQ House price 2		-0,09	
		(-2,18)	

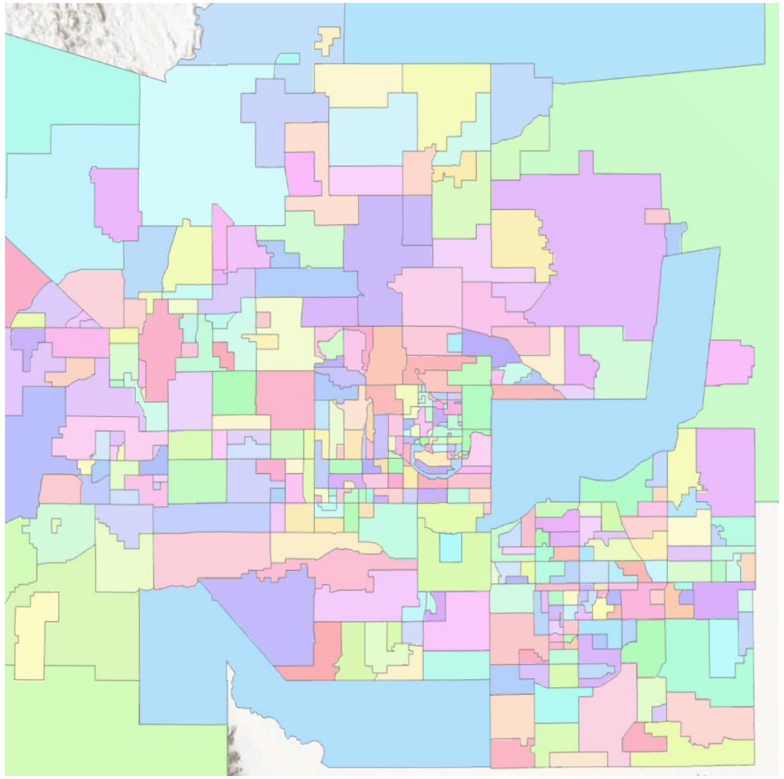
Second view of the price landscape

- Figure 3. Land value based on both land sales and improved sales produced by Empirical Bayesian Kriging (EBK)
- A more nuanced view of price landscape compared to figure 1. The number of data points is much larger, in many locations orders of magnitudes larger.



Possible areas of same land value

Figure 4. Number of RNBHDs: 377.
The colors only indicate the presence of a RNBHD

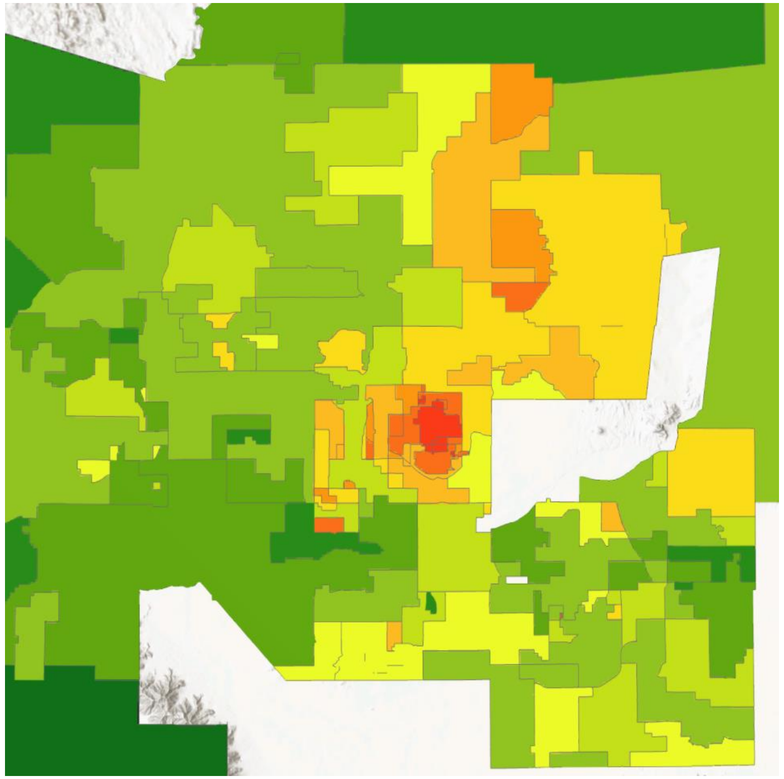


- Candidates for value zones were produced by spatially constrained multivariate clustering (SCMC). The initial subdivision which SCMC was based on was RNBHD, the number of which was less than 380. Some of the RNBHDs are very large. I would have needed a more detailed subdivision of several thousands, such as blocks of lots.

Possible areas of same land value

- Neighboring RNBHDs were united, if their land value estimates were close enough. Three different sets of clusters were produced based on the number of clusters: 55, 111 and 222. Figure 5 shows the result with 111 clusters.
- The tool spatially constrained multivariate clustering (SCMC) is offered by ArcGIS and has been in the commercial software since 2019. To my knowledge it has not been used so far in the assessment community.

Figure 5. Value of land based on land and house sales. Number of clusters: 111



Conclusions

- These exercises illustrate in simple non-sophisticated terms the tools that are used to produce residential land values for mass appraisal purposes. As to assessment of non-residential land, these tools are less useful and are certainly more challenging. These tools are useful also in mass valuation of improved residential property, but the approach is different and is not tested here.

The data

- The data itself is rich and high quality.
- However, as my approach is based on subdivisions offered in the data, I missed one or two very detailed levels of subdivision, such as blocks of lots.
- A lot ratio variable would have been welcome.

Conclusions / Hedonic modelling

- Hedonic models are general models. When fitted to local markets the outcome is more accurate and useful.
- Hedonic models are general also in terms of time. Especially the land price model comprises 19 years of sales. Some of the price factors must have changed a lot. The spatial-temporal interaction must have played a role, as some locations must have been winners and other losers.
- Other interactions were neither tested.
- The choice of arguments in quality function, produced by hedonic function, is very briefly done. Some important arguments may be missing.
- The choice of control variables in hedonic function, is very briefly done. However, this may not be critical. The control variables may be good enough.
- The connection between house prices and land values is the Achilles heel of the method. The modelling of that connection needs to be done with extreme caution. In this exercise the elasticity was estimated using grid averages. Some other subdivision may be more useful, perhaps some subdivision with prior knowledge of homogenous value zones.
- The method is critical to accurate measure of constant quality land and house values. Even if the method works well in most locations, in some locations it may be badly misleading.
- The calculated elasticity figure, 1,76, which measures the effect of house prices to land prices, is in line with my previous research.

Conclusions / Spatial analysis

- Kriging, namely EBK, was used to produce the value landscape. However, so far this was only used as a background information, to guide the way. The calibration of kriging parameters was minimal. So far I was convenient with parameters offered automatically by the tool. If value zones were tried to produce EBK-based, the calibration needed would be a critical and perhaps time consuming task.
- Candidates for value zones were produced by spatially constrained multivariate clustering (SCMC). I missed a detailed level of subdivision, such as blocks of lots. The initial subdivision which SCMC was based on was RNBHD, the number of which was less than 400. Some of the RNBHDs are very large. I would have needed a more detailed subdivision of several thousands, such as blocks of lots.
- The calibration of SCMC was not done, apart from setting the number of clusters. Several parameters can be chosen to produce an optimal clustering.

Different types of land

- The method has its best potential in land for single-family housing, by far the most important land use type in United States and many other countries, by land area and very often in land value too.
- As to other residential land, the method has limited potential, or the method is more challenging.
- As to commercial land use, the method may not be useful at all
- As to CBDs and high rise areas around them: the method may have some potential as a starting point, if there are sales prices available, mainly sales prices of land and residential property, including condominiums.
- Finnish Land Survey has used it successfully in densely built residential land.

Conclusions / The results

- Probably the value zones are far too large, and the number of them too low (max 200 in my exercise). A far larger number of value zones could be produced, of course, if a more detailed initial subdivision were available, or were artificially produced as a part of the project. As the number of data points is ca. 2 million, around one thousand value zones is not unreasonably large.
- At this point there may be any number of inconsistencies and small errors and in this exercise. Some of them will be easy to correct. Many of them will be challenging and some of them will be unsolvable.