Maxime Colomb
4GC - Promotion 2008/2013
2ème semestre 2011 - 2012

ÉCOLE NATIONALE D'INGÉNIEURS DE SAINT-ÉTIENNE
58 rue Jean Parot – 42023 Saint-Étienne cedex 2
Rapport de période industrielle du 6 février au 22 juin 2012

How to Modelize Los Angeles in a Comprehensive System

University of Southern California

Los Angeles

Supervisor: Dr. Roger Ghanem

Professor Referent: Dr. Hanene Souli & Mr. Jacques Lipp

Summary

I) Introduction

A) Megacities Research Center

The Center on Megacities develops innovative solutions for megacities through its interdisciplinary expertise in both science and engineering, in area such civil and environmental engineering, information technology, architecture, economics, social science, policy and planning, and public health. Building on the partnerships between universities, the government, industries, and nongovernmental organizations, the Center on Megacities brings together engineering and other disciplines to innovate for a better future for megacities. Progress made in solving megacity issues will benefit cities of lesser size, thereby contributing to the broader world's welfare.

I have been integrated into this lab to lead a pre-study with a precise task: To apply existing theories in order to create a model of Los Angeles. I will first begin with a presentation on the basis of the subject: the city of Los Angeles.

B) Los Angeles, California

Los Angeles is the second biggest city in the United States, after the New York area. 18.5 million inhabitants are living in the big urban space, with 4 million in the proper city. It is the emblem of the American Dream and thus known all around the world. Its urban system, founded on a completely different basis than our European cities, brings about a lot of imperfections that an ambitious policy of the city had tried to correct.

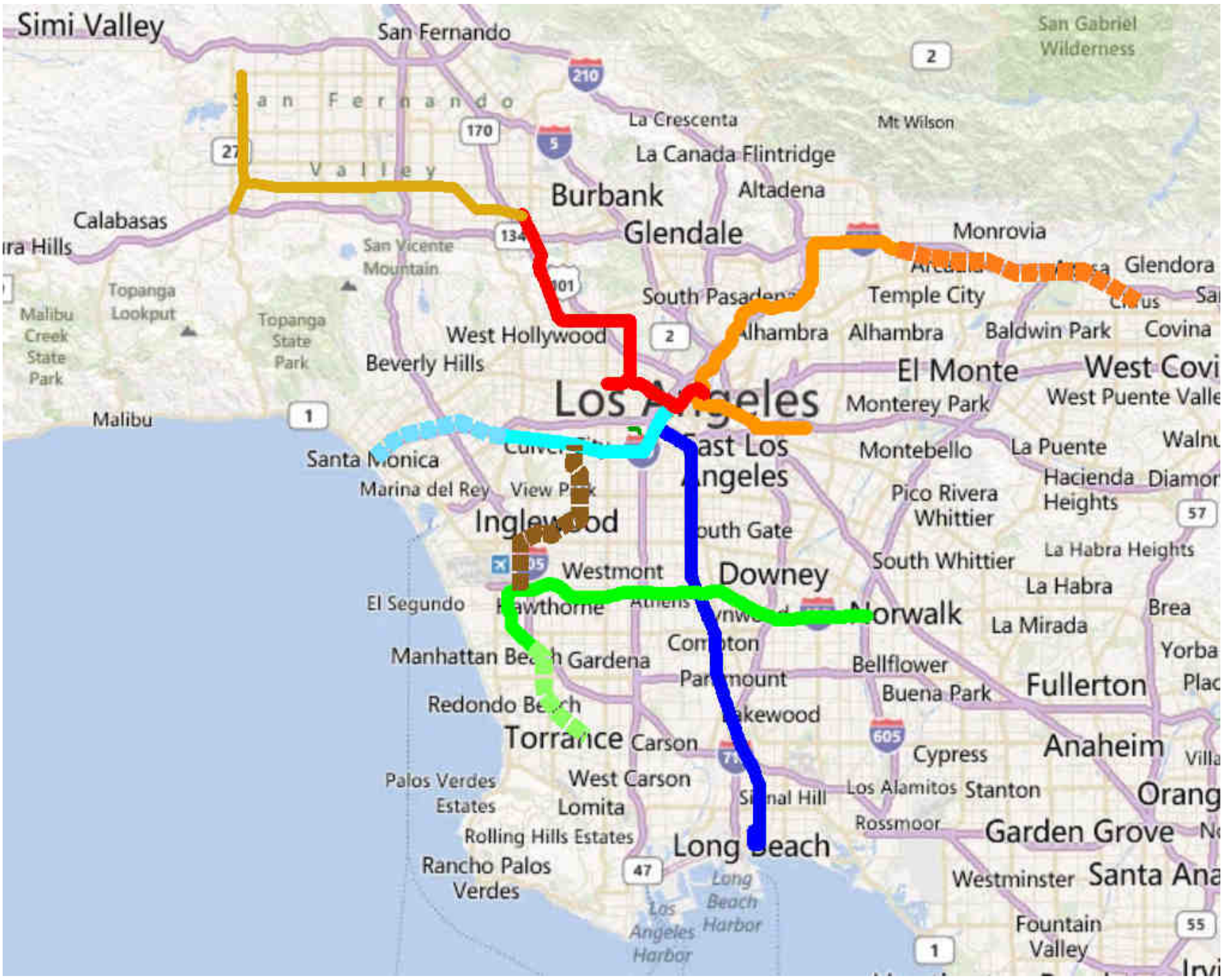
The Golden State was owned by Mexico before becoming an American state in 1848. The perfect weather of the gulf of Los Angeles made the region a fantastic orange producer; make this little 2500-people city in 1860 a famous place in the United States. People in the states who were searching for a milder climate and wild areas moved to Los Angeles, provoking a sudden demographic explosion. Cinema has also taken advantage of the Los Angeles climate. Hollywood has become the world capital of film, which has made Los Angeles the world’s headquarter of contemporary pop culture. Hollywood’s projection of the image of sparkling success continues to attract immigrants, therefore sustaining the crazy expansion of the city for many years.

Demographic history

Year	1887	1900	1920	1930	1960	1990	2010
Los Angeles population	11 500	102 000	577 000	1 238 000	2 479 000	3 485 000	3 792 000

The housing model in Los Angeles was, since the beginning, based on single houses with a garden, or another private personal space. The aim in coming here was to be close to nature. Nowadays, the expansion of the city has transformed the whole area. City expansion has reached all of the cities within the Los Angeles Basin, giving rise to one of the biggest megacities in the world. These spaces of very low density made private transportation the only way to travel in a giant freeway complex. Driving a long time into traffic rush became the daily routine of the population, which ironically brought about the opposite of the nature that people had sought for. With the promise that a brand new city can bring in a liberal American system where everything seems to be possible, poverty arose and began to be a real problem. People became confined to their communities, the wealthy moved into suburbs like Orange County, on the beach or on the bottom of the Hollywood hills, and the poor concentrated themselves within the southern neighborhood. In the 90's, gangs were a real problem. In 1992, because of the racial tension and the ignorance of the politics surrounding it, a riot began in the neighborhood of South Central, the same one within which USC is located. 60 people were killed and the damage done to the city amounted to a billion dollars. The city knew that the urban problem was real, and started to feel the immediate need to fix it to avoid such problems. Another obvious example of these tribulations is the Downtown of the city. Skyscrapers were built in the 70's and the neighborhood stayed for a long time an office place by the day and a no man's land by night, full of violence and homelessness.

Today, the on-working gentrification of this area shows the political will of the city to make this unique urban model more humane. Big recreational centers have been built downtown, bringing back the rich population, which usually chooses to live in a Downtown area, and creating the first real public center in the city where everybody seems to only drive. For a long time, public transportation was only made by bus, but in the 90's appears a Metro system appeared and continues to expand today. Even if the place is very big and serving all the superficies of the city seems to be impossible, it shows the real will to avoid this model that creates an inhuman city that is full of individualism and pollution. The Metro Gold Line and Silver Line were just created when I arrived, and I've been to the opening of the Exposition line, which is outlined in cyan on the following scheme. It links Downtown LA to Culver City, in passing by the USC campus. In few years, this metro line will reach all the way to Santa Monica (extension materialized by the dotted line). There are a lot of other metro projects in the whole city, like the extension of the Green airport line to Torrance, the extension of the gold line to Azusa, and the new Crenshaw line, in brown, which will link the airport to Culver City. All these projects show the real will of policy to get the city better, since better transportation can improve the enjoyability of living in an expanding place.



Los Angeles as a whole is divided into four different counties, and hosts 88 different cities, with different social and race localities. In Los Angeles, almost 50% of the population consists of Hispanic people, a sign of the huge diversity of this fascinating city.

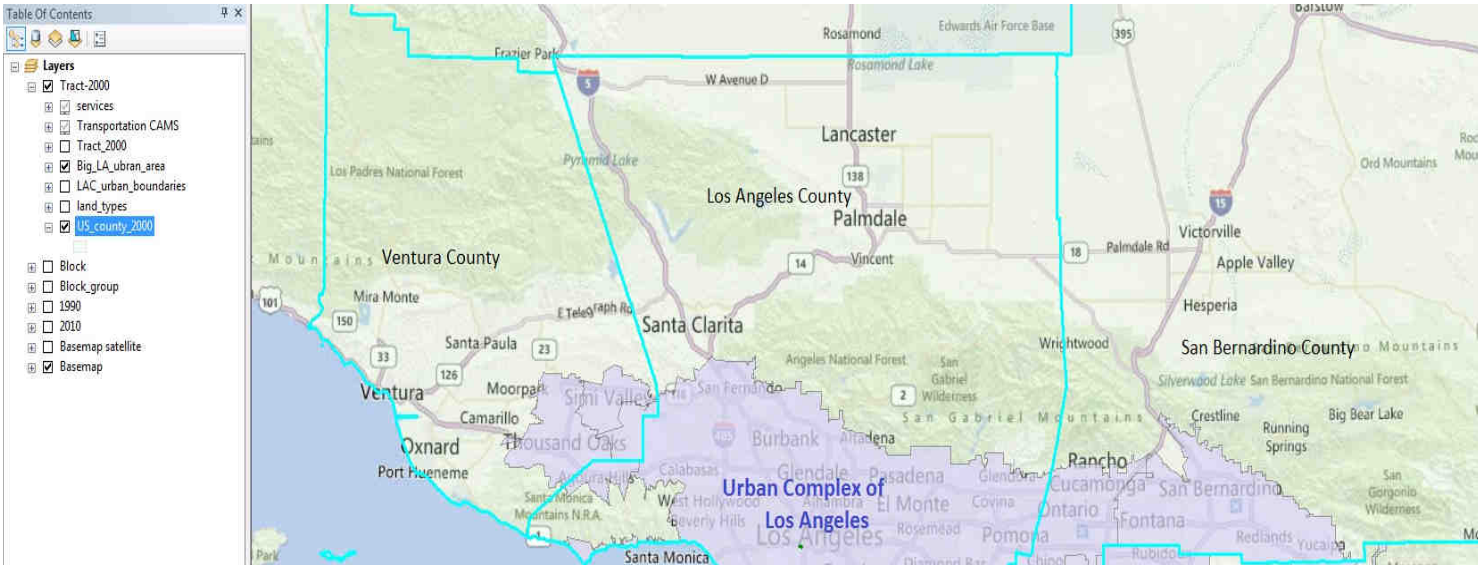


Table Of Contents

- Layers
  - ☑ Tract-2000
    - ☑ services
    - ☑ Transportation CAMS
    - □ Tract_2000
    - ☑ Big_LA_ubran_area
    - □ LAC_urban_boundaries
    - □ land_types
    - ☑ US_county_2000
  - □ Block
  - □ Block_group
  - □ 1990
  - □ 2010
  - □ Basemap satellite
  - ☑ Basemap

What is it like to live in LA	Los Angeles Area	California	United States
Median household income	$46 393	$61 283	$51 660
Median home price	$400 360	$342 040	$183 450
Cost of living (% of national avg)	141%	135%	100%
Unemployment	14%	12%	10%
Violent crime index (1 is lowest)	7	6	4
Days per year with some sun: 284			
Days per year with some precipitation: 26			

C) Modeling the city

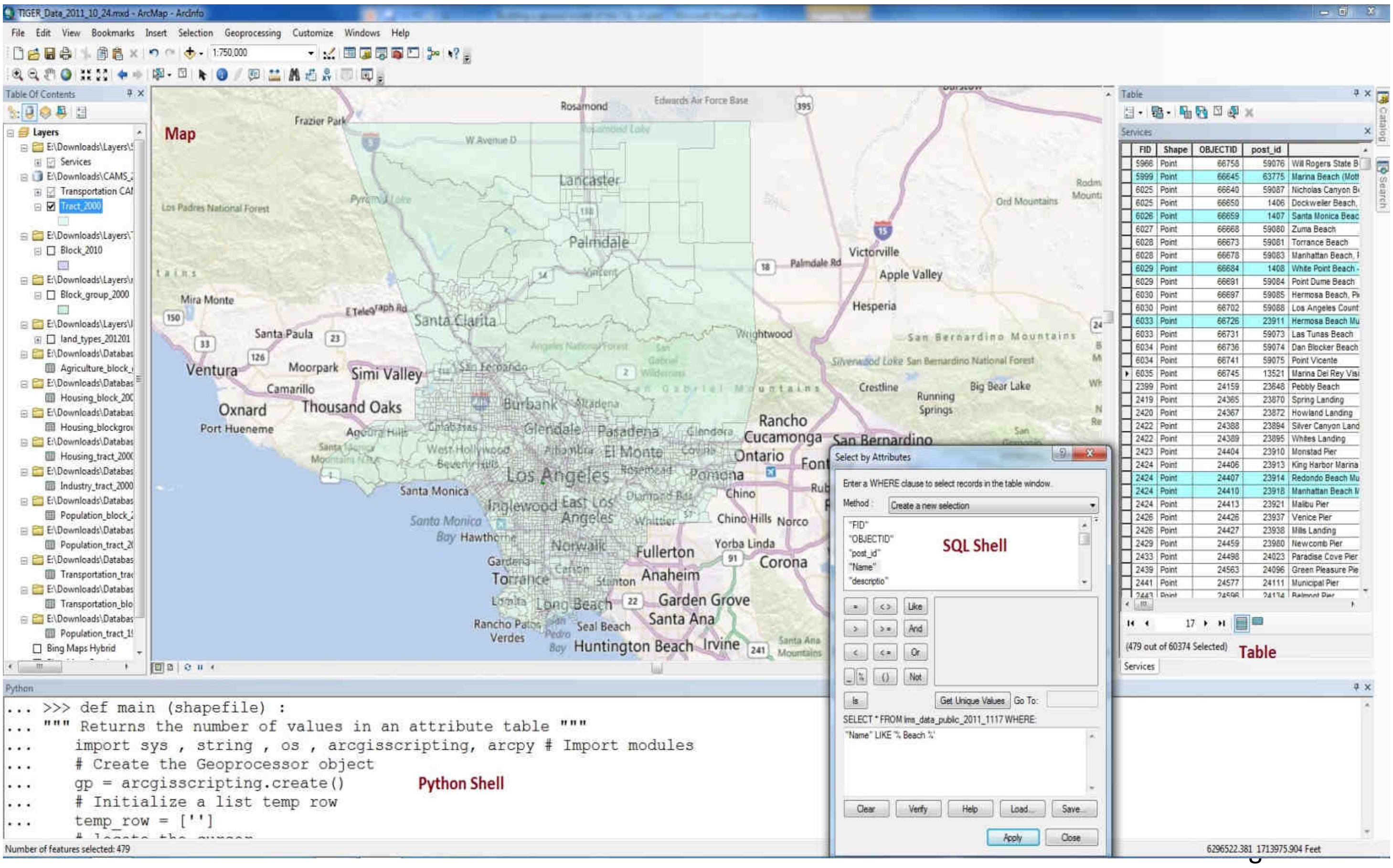
The need to model everything increased with the power of computers. Research became more advanced and efficient decisions are harder and harder to make for the policy makers who have to deal with a system that gets more complex everyday. Having a model of the city which allows us to test the repercussions of political decisions would be a great power. Theories and models have existed for years, but the increase of computer power and informatics tools are what make the application of these theories possible. If we could have a system dealing with the entire city's data and which represents the reality of how everything is interacting, we would be able to test how changing a general parameter can transform an area in a better way. Such as, creating a new business center, facilitating immigration in a place, what will be the repercussions on the transportation system, retail centers and economic backcloth… Because such projects necessitate a strong knowledge in urban planning, mathematics, geographical sciences and programming, we call it the "complexity sciences".

With the internet as a powerful tool, available data have become more and more complete and accurate. The US Census Bureau collects vast amounts of data about the whole United States, with a clear classification and allows us to access to some of it. Using Geographical Information Systems (GIS) to model the city and deal with all types of information, my goal during this research period has been to find a way to apply these models and to begin to build a comprehensive system for the city of Los Angeles. Working by myself and without a real geomantic formation, this huge mission has been vague at times. But I learned a lot of things that I am now passionate about, and I now know that I want to pursue a professional career in that field. I changed my major for the next year for a double diploma of Enise and a master II Humanities and Social Sciences, specializing in Geographic Information Systems at the University of Saint Etienne. I want to come back to USC to continue this project during a PhD period. I am going to present to you the models I've learned from, the work that I did, and the continuing research work needed.

II) Model used

A) Definition of a GIS

A Geographical Information System (GIS – SIG in French) is a system designed to capture, store, manipulate, analyze, manage, and present all types of geographical data like maps, layers of geometrics objects and databases. The most popular software, used by all geometricians is ArcGIS, from the private firm ESRI. It has three levels of licenses - ArcView, ArcEdito and ArcInfo - that allowed more and more accurate tasks. However, there exist many different kinds of GIS, with Grass being the most popular freeware. I had to learn how to use it to create my own core system of Los Angeles. Many data libraries are available on the Internet, offering different formats. Some are dispensing layers, geometrical objects which are synchronized with the base map to create areas representing all kind of information. For example, dots could represent services location, lines could represent the transportation network or polygons could display a segmentation of the urban area. The other big kind of information is database, non spatial tables that we can join or relate to spatial layers. The United State Census Bureau is the official organization charged to collect the geographical information about the nation's people and economy. They have an accurate system of splitting the cities into different scales of discrete areas. This is for example the representation on the Los Angeles County on ArcMap. The green transparent polygons shape is linked to the table in the right part of the screen.



	FID	Shape	OBJECTID	post_id	
	5966	Point	66758	59076	Will Rogers State B
	5999	Point	66645	63775	Marina Beach (Mott
	6025	Point	66640	59087	Nicholas Canyon B
	6025	Point	66650	1406	Dockweiler Beach,
	6026	Point	66659	1407	Santa Monica Beac
	6027	Point	66668	59080	Zuma Beach
	6028	Point	66673	59081	Torrance Beach
	6028	Point	66678	59083	Manhattan Beach, I
	6029	Point	66684	1408	White Point Beach -
	6029	Point	66691	59084	Point Dume Beach
	6030	Point	66697	59085	Hermosa Beach, Pi
	6030	Point	66702	59088	Los Angeles Count
	6033	Point	66726	23911	Hermosa Beach Mu
	6033	Point	66731	59073	Las Tunas Beach
	6034	Point	66736	59074	Dan Blocker Beach
	6034	Point	66741	59075	Point Vicente
▸	6035	Point	66745	13521	Marina Del Rey Visi
	2399	Point	24159	23848	Pebbly Beach
	2419	Point	24365	23870	Spring Landing
	2420	Point	24367	23872	Howland Landing
	2422	Point	24388	23894	Silver Canyon Land
	2422	Point	24389	23895	Whites Landing
	2423	Point	24404	23910	Monstad Pier
	2424	Point	24406	23913	King Harbor Marina
	2424	Point	24407	23914	Redondo Beach Mu
	2424	Point	24410	23918	Manhattan Beach M
	2424	Point	24413	23921	Malibu Pier
	2426	Point	24426	23937	Venice Pier
	2426	Point	24427	23938	Mills Landing
	2429	Point	24459	23980	Newcomb Pier
	2433	Point	24498	24023	Paradise Cove Pier
	2439	Point	24563	24096	Green Pleasure Pie
	2441	Point	24577	24111	Municipal Pier
	2443	Point	24596	24134	Belmont Pier

17

(479 out of 60374 Selected)

Select by Attributes

Enter a WHERE clause to select records in the table window.

Method : Create a new selection

"FID"
"OBJECTID"
"post_id"
"Name"
"descriptio"

SELECT * FROM lms_data_public_2011_1117 WHERE:

"Name" LIKE '% Beach %'

Python

```
... >>> def main (shapefile) :
... """ Returns the number of values in an attribute table """
...     import sys , string , os , arcgisscripting, arcpy # Import modules
...     # Create the Geoprocessor object
...     gp = arcgisscripting.create()
...     # Initialize a list temp row
...     temp_row = ['']
```

Number of features selected: 479

6296522.381 1713975.904 Feet

In order to automate tasks, ArcGIS allows us to use two informatics languages. We can see the windows shell on the principal screen. The Python language is now fully integrated with the 2010 version of ArcGIS, and the modules arcpy and arcgisscripting give a large set of geo-processing tools and functions. I had to establish a strong basis of this powerful language in order to write the main script of the program. The syntax of this interpreted programming language is clear, and even if there are an infinite amount of different complicate libraries (arcpy is one of those), I enjoyed learning this useful language.

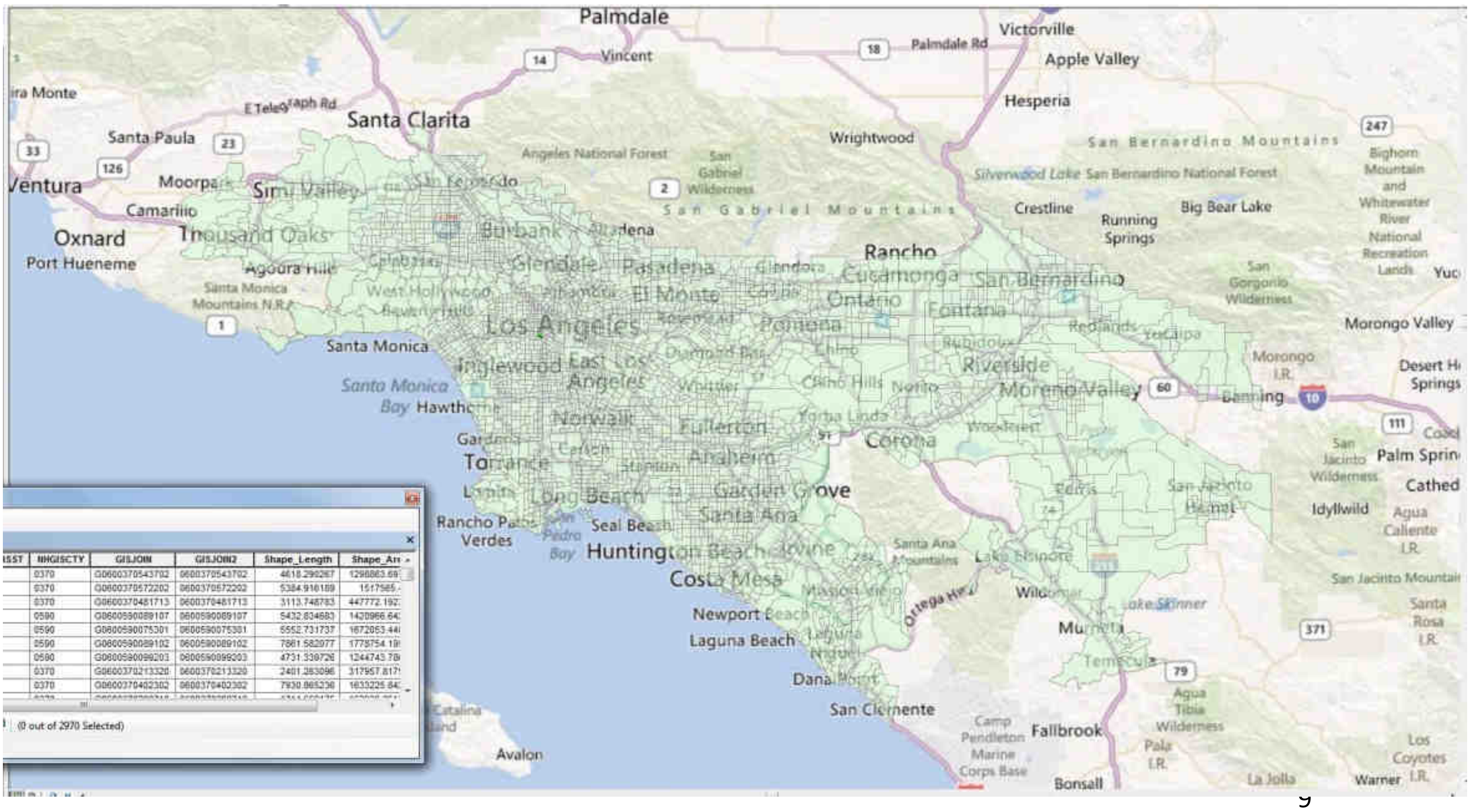
Also to operate on tables, ArcGIS is using a SQL shell to select data. The request is already on the form SELECT * FROM WHERE: I have already taken an initiation to this language, I knew the basis about the DataBase Management Systems (DBMS).

B) Universe

Many different settings are possible in using GIS systems. I will briefly define the basis universe I am going to work with.

Base: other systems exist, but the Census system is the most common in the GIS library websites, and depends on the federal Bureau, so all information are official.

Universe: I'll start in modeling only the Los Angeles County, but as seen on the map, the system is only complete in the greater Los Angeles area. The work has to be extended in the future to these counties – Orange, Riverside, San Bernardino and Ventura. To avoid a lot of data manipulation, I kept working with only the Los Angeles County.



ISST	NHGISCTY	GISJOIN	GISJOIN2	Shape_Length	Shape_Ar
	0370	G0600370543702	0600370543702	4618.290267	1298863.69
	0370	G0600370572202	0600370572202	5384.916189	1517585.
	0370	G0600370481713	0600370481713	3113.748783	447772.192
	0590	G0600590089107	0600590089107	5432.834683	1420966.64
	0590	G0600590075301	0600590075301	5552.731737	1672063.44
	0590	G0600590089102	0600590089102	7861.582077	1778754.19
	0590	G0600590099203	0600590099203	4731.339726	1244743.78
	0370	G0600370213320	0600370213320	2401.283096	317957.817
	0370	G0600370402302	0600370402302	7930.865236	1633225.84

(0 out of 2970 Selected)

Year : I will use the census tract of 2000. I chose this year because it was a complete census year and the most information was available; the results of 2010 are not yet totally complete. We can collect other years like 1990 and the beginning of the 2010 set because we will calibrate the model but the fact that new census set comes only every ten years is a big limit of the system. Data won't be really accurate and few sets are really workable. We can imagine that in a couple of years, tasks will become more automated and a complete set will be available each year.

Scale: Here is the splitting of all the American territories:

- State: 50 federal states with a proper federal government.
- County: There are 3,033 organized counties or county-equivalents. The average number of counties per state is 62 and the average county population is about 100,000. The Los Angeles County is the most populated. With almost 10 million inhabitants, it is bigger than 42 US states.
- Census tract: All Counties are divided in areas from 1500 to 8000 inhabitants. The average is 4000. There are 2055 different ones in the 2000's Los Angeles county census set.
- Blockgroup: census tract divided in 3 parts: 6352 areas in L.A County, 211,267 block groups in the US, each containing an average of 39 blocks.
- Census Block: on the smallest scale, there are 89,615 in the L.A County, about 8,200,000 in the US

All those scales are changing from a dataset year to another. We are working with the census level because it's accurate enough and the total number is not too big. But to get more accurate results for future applications, we could extend it to the Block group or block. The calculation would become heavier though.

C) Comprehensive Model of the city

The first work on the task of inventing a model of the city I read was "The dynamics of Urban Spatial Structure: The progress of a Research Programme", written by M. Clarke and A.G. Wilson in 1985. It introduces a way of modeling a complete city in building subsystems which are run by two mathematical patterns: The use of dynamical equations and the maximization of entropy. First, it dispenses a list of different principal headings and subheadings that we need to use in constructing our model:

9 substantive systems of interest: agriculture; industry; private services such as retailing; public services such as health and education; housing; transport flow and infrastructure; population backcloth; economic backcloth; and the sum of all those sub systems into a comprehensive model.

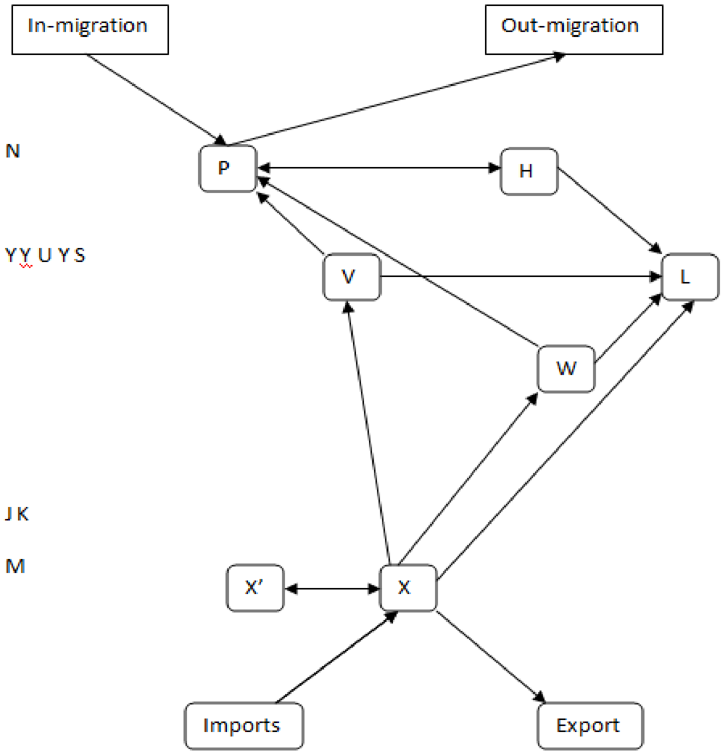
7 methods: spatial interaction, accounting, mathematical programming, network analysis, micro simulation, dynamical analysis and planning theory.

4 utilities : problem analysis, policy formulation, design and interactive planning

These are the different structural topics we need to coordinate in order to build the system of the city. We have to build the models of each of the subsystems, choose the right scales and method of resolution, and make them work together. In a more recent work: “Entropy in Urban and Regional Modeling: Retrospect and Prospect“ (2009), Alan Wilson lists what would be the main steps to construct the system:

First, assemble the core data to build a Geographical Information System (GIS), a set of geographical data that we can easily manipulate. There are certain tasks that we can do as routine, and the more the system will be automates, the better it will be. We have to seek for as much data as possible, from different years and scales. If some of them are missing, once the system is built, we would be able to operate estimations. In this phase, we decide too which settings will be used to build the base of the system.

Then, we have to decide for the sets of subsystem which will represent this general urban model and formally how do they will interact between each others in a realistic way. The following scheme is an example of the Alan Wilson city model. Each of the subsystems has a known way to be modeled, but decisions on the way to adopt details and link them are still not well defined. There are tasks that we can represent as a routine. There are GIS system like TransCad especially made to model the transportation network.



Key :
Stock and structure Flows
P : population
H: housing
X: the economy
X’: another branch of the economy
W: retail
L: land
V : Public services

N: Demographic moves
S: to retail
U: to public services
Y: Journey to work
J: economy to public services
K:economy to retail
M: business to business

When all of the sub systems are built, we have to calibrate each of them to make them represent the real situations. We would move the different coefficients in the formulas used. The different types of data would be useful on that part.

When all systems are correctly built, we need to articulate the links between the submodels and assemble a general comprehensive model in a unique program. In my opinion, this would clearly be the hardest task. Once done, we will add a graphic interface to the program to create what would be called a CityIS. The first goal of this will be to send « What if . . . ? » requests and stock outputs in data warehouses. After manipulations and system testing, we would be able to build interactive planning.

D) Mathematical model

There are two big mathematical notions used to build the subsystems. The first one is the application of the entropy maximizing, which offers what might be called an optimal blurring of the social or economic system. This concept was first used in thermodynamics but was later applied to many fields. The second concept is the use of non linear dynamics equations, which brings some interesting proprieties:

- There can be multiple equilibrium solutions
- These solutions are very dependent on the initial solutions. We call it path dependence.
- There may be critical values at which there are sudden changes in the structure. We will have to study the phase which leads to those critical values in order to predict it, to avoid it and to change it into favorable phases.
- There are different types of dynamics. If there are any modifications to the core of the city system, some submodels will interact faster than other. Transportation for example will find a solution very quickly, while housing will instead need more time to return to the equilibrium. We have to consider the fast dynamics and slow dynamics independently.

I will explain the meaning of the basic demonstrations developed by the geomathematicians to define the common core of each of the sub systems.

Imagine T_{ij} is a measure of the interactions between zones i and j
We can build a model as $T_{ij} = T_{ij}(O_i,Z_j,c_{ij})$ with
O_i as the demand of the service

Z_j as the attractiveness of the facility
c_{ij} as a measure of cost travel

Then, let D_j be the total use made of the facility like $D_j = \sum_i T\mathrm{ij}$

And $C_j(Z_j)$ is the cost of running the facilities

We can hypothesis that : Zj will increase if $D_j>C_j$
Zj will decrease if $D_j<C_j$

Or, written as a differential equation, $\frac{\partial Zj}{\partial x} = \varepsilon(Dj - Cj).Zj$

The parameter ε determine the speed of the response to the movement signal

The most probable state of the system can be found by maximizing the entropy function

Max T_{ij} = - $\sum_{i\ j} Tij \log(Tij)$

We finally obtain an equation of the type : T_{ij} = Ai . B_j . O_i . D_j . exp (β c_{ij})

A_i and B_j are expressions that look like partition functions in statistical mechanics. This led to the Boltzmann statistical mechanics analogy. They will be in the forms :

B_j =1 / ($\sum_k \mathrm{Ak} . \mathrm{Ok} . \mathrm{c\ kj}$ ^($-\beta$)) and A_j =1 / ($\sum_k \mathrm{Bk} . \mathrm{Dk} . \mathrm{c\ ik}$ ^($-\beta$))

We can see them as the competitive effect from the zone k.

Because this model combines Boltzmann's statistical mechanics and Lokta's and Voltera's dynamics (LV), this has been characterized as the BLV model.

A lot of system bases are built on this BLV model.

Once all of the models are built, we would be able to search and work on the proprieties of the BLV equations. A dynamical system is really dependent on the initial conditions, the initial state of the city. It determines the region of state space which is accessible in the future. We will be able to create a vector of many dimensions; each would represent a deep path characteristic of the city. This will represent the DNA of the city during a span of time.

{e, P, H, X, L, W, V, c} is, for example, the DNA of the city that A. Wilson imagined in his model displayed earlier.

Urban planners will thus have a very accurate tool, the representation of their city in simple terms and they will see immediately see the feasibility and the rapidity of their actions. If they would like to change a deep characteristic of the city, they will have to act to modify this structural vector.

There are research questions at each level. Even if every kind of system has a known representation, they are not always compatible to be articulated between each other. Universes are different, and hypotheses too; but the common scientific base, BLV equations, will well-represent cases, and much progress has been achieved this way in very complicate fields like genetics. We should be able to build it, and an efficient model could be one of the greatest evolutions of the twenty-first century in the way of thinking and building the city.

III) Work undertaken

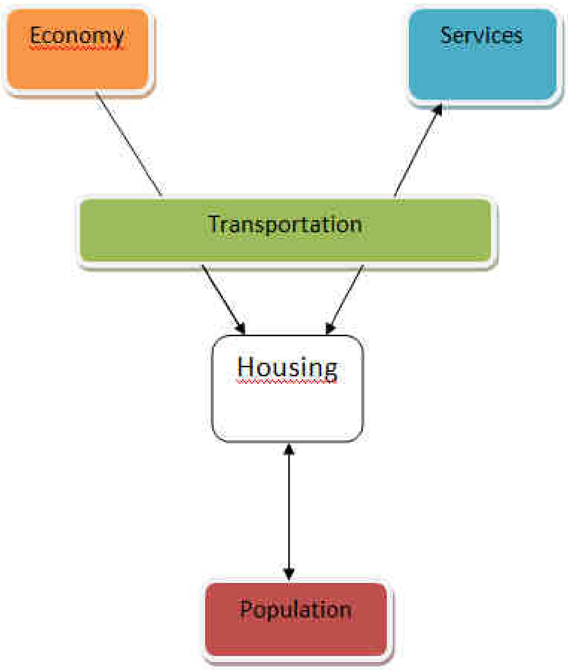
A) Isolating a system : Housing

Of course, in a mere 20 weeks and without a team and deep knowledge about all the subjects presented, I didn't have the time to consider the whole system. I chose to begin in isolating a single subsystem, fixing exogenous variables from other systems that will interact with it, and then applying the theories I've explained before to build this system. As a lot has been done on the retail center, choosing this system again wouldn't be very interesting. I could have tried to work on it further in order to improve it, but instead I prefer trying to build my own. I chose the housing subsystem. This is a central subsystem as it overlaps with a lot of other systems, and it is based on very slow dynamics. The choice of moving out to a new place is a decision that takes time and efforts from a population. I was interested in modeling that, and to see how this special and hard characteristic could be represented. I decided to carry out these steps:

- Isolate my unique subsystem : Housing
- Modelize a good submodel in fixing exogenous variables from the other models.
- Assemble a GIS
- Write the script
- Calibrate it with all my data
- Build other submodels, such as retail or population, and articulate the links

Unfortunately, due to facing the multiple problems and my missing skills in program writing, I didn't have the time to finish it. I will however present the Housing subsystem I built:

Housing is supposed to follow the employment location, but in Los Angeles, even if with the traffic rush hours that paralyzes the city, people would still choose to live in a place far from their workplace, for many reasons. The two persons of a couple could work very far from each other. The neighborhood is, for certain people, more important. This leads us to a point about Los Angeles: Driving is a part of the Angeleno life. People may spend long hours driving just to go to a restaurant. Here is the scheme representing the isolation of housing system:



Population is strongly linked to the housing system; it defines where people lives.

Services will be adapted to the housing supply, but we'll see that the housing will deal with the services available too.

We are examining the hypotheses that Housing will only be affected by economy and that changing housing parameters won't affect the economy

The transportation system is only a way for people to move in Los Angeles, it won't be a decisional subject, but it will affect journeys between housing and economy/services

I didn't put the Land subsystem in the scheme because it can be assumed as a simple idea. The sum of each of these facilities must not exceed the available surface of the total city

$A_{house} + A_{facilities} + A_{unusable} < A_{zones}$

To continue with the mathematical subsystem, we need to introduce some indices:

w as a range of household income
k as the type of house
i as the living zone (2055)
j as the working zone (2055 too)

Let $W^{res}{}_{i}{}^{k,w}$ ^α^{w} be a representation of the residential attractiveness with

Accessibility to services
Affinity to social group
Bid rent term/affordability
Available housing stock
And α^{w} a parameter associated with how each class will value each part of the W representation.

$E^{w}{}_{j}$ a known distribution of employment,

μ^{w} a parameter relating to work travel distance. It will be lower for the higher class population

$c_{i\,j}$ a travel cost constant. It represents the travel time from I to j, it would be facilitated by a freeway, for example. Common transportations are still not present enough in the urban area to be modeled.

$T_{i\,j}^{k\,w}$ is the number of w income people working in zone j, who live in type k house in a zone i

$T_{i\,j}^{k\,w} = B^w_i\, W^{res}{}_i^{k,w} \,.\, E^w_j \exp(-\mu^w \,.\, c_{i\,j})$

Where $B^w_i = 1/\, (\sum_m W_m^{res\;k\;w} .\, \exp(-\mu^w \,.\, c_{i\,m}))$

We can see the m zone as a measure of the competition of the other zones.

The equilibrium equation is:

$\sum_w T_i^{k\,w} \,.\, q^{\,kw} = H^k{}_i . p^k{}_i$

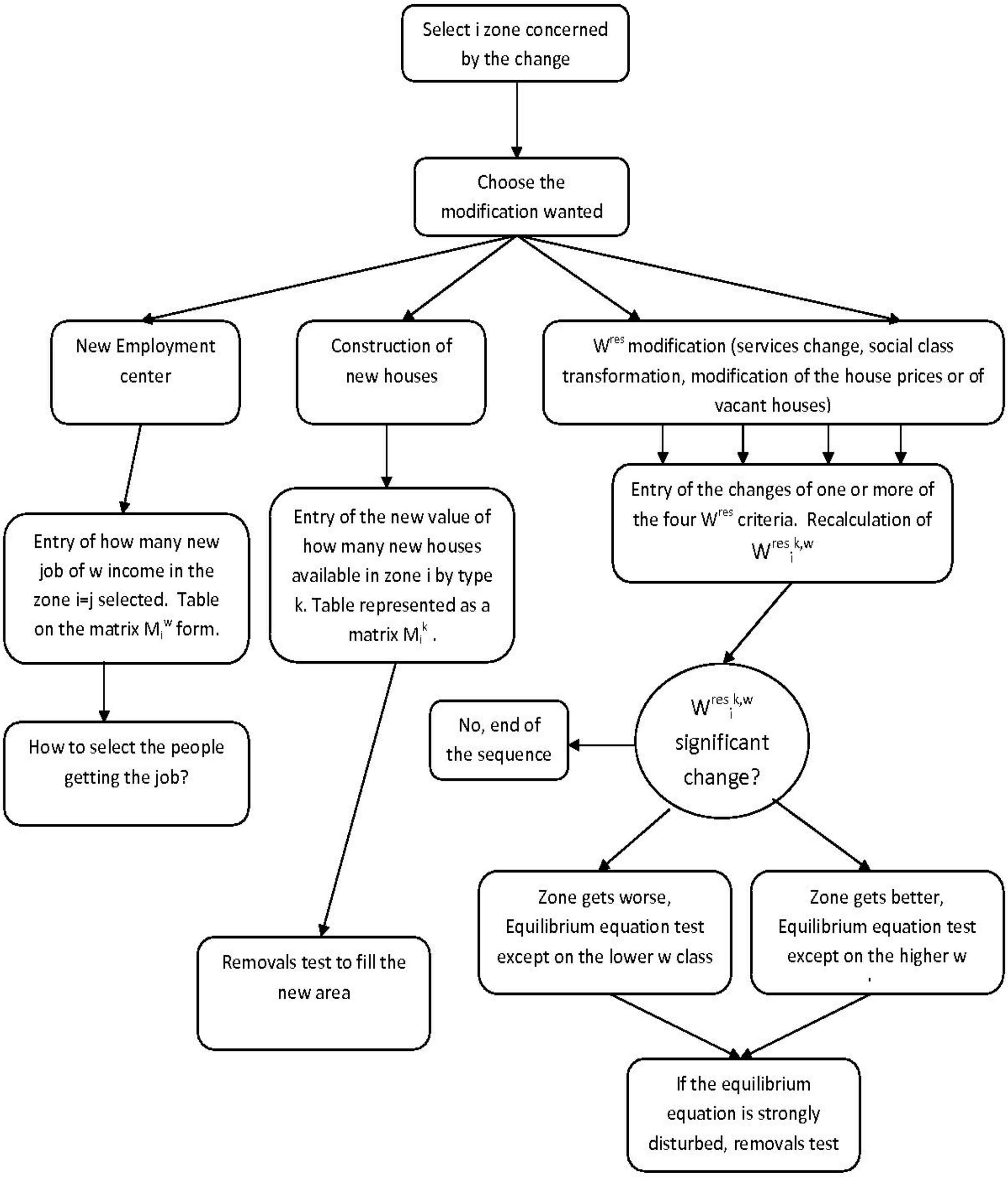
$q^{\,kw}$ is the average price that w household income are willing to pay for a k house
This value can be found in building a map which recency prices in the house market by it class k

$H^k{}_i$ is the amount on type k housing in i

$P^k{}_i$ is the price of type k housing in i

We need to create programs that will execute these formulas that Python can totally operate on. Once it is done, we have to calibrate all the different coefficients to make it reflect the real situation in Los Angeles. Then we must run it for a global test in the whole city system. All the $W^{\,res}{}_i{}^{k,\,w}$ values will be stocked in a database folder, as it will be easier for the program to use the data from this folder whenever it needs to. Changes in .dbf table files don't depend on the number of raw, it will be fast to calculate and replace a new value. We can consider the $T_{i\,j}^{k\,w}$ term as a tensor of order 4, as other terms with more than two indices. Python can deal with this representation as it creates links to databases. For now it would be easier to use simplifications. In another folder, we will stock if the tract is increasing, decreasing or equal concerning the equilibrium equation. These heavy data would take a long time to calculate at first, but once done, it will be very easy to request for it upon the system.

To get to know how the system moves, I built an algorithm to make modifications to the core city. It would be the most massive modification that we can do on a city's system, and would show us how it affects the housing situation.


Select i zone concerned by the change
Choose the modification wanted
New Employment center
Construction of new houses
W^{res} modification (services change, social class transformation, modification of the house prices or of vacant houses)
Entry of how many new job of w income in the zone i=j selected. Table on the matrix M_i^w form.
Entry of the new value of how many new houses available in zone i by type k. Table represented as a matrix M_i^k .
Entry of the changes of one or more of the four W^{res} criteria. Recalculation of $W^{res}{}_i^{k,w}$
How to select the people getting the job?
No, end of the sequence
$W^{res}{}_i^{k,w}$ significant change?
Zone gets worse, Equilibrium equation test except on the lower w class
Zone gets better, Equilibrium equation test except on the higher w
Removals test to fill the new area
If the equilibrium equation is strongly disturbed, removals test

We can distinguish between two different types of removal tests. One will be a research of a population living in a i zone and moving in a i' zone, and the other will be a research of the population finding a new place to live. In the case of the new employment center, it is an interesting criteria but I can't resolve it now because it's too linked with economic issue. It would have to submit the two tests, finding people to fit in and is it worth to move in, and where. The way we can model interaction with economy would be with a personal path, which is on too small a scale. As I have hypothesized on my subsystem isolation, housing can't interfere in this way with economy.

We have another problem. How could we choose between all the discrete areas of the system to operate a removal test on the whole population?
In the $T_{ij}^{k\,w} = B^{w}_{i}\, W^{res\,k,w}_{i}\ .\ E^{w}_{j}\, \exp(-\mu^{w}\ .\ c_{ij}) / (\sum_{m} W^{res\,k,w}_{m}\ .\ \exp(-\mu^{w}\ .\ c_{im}))$ equation :

- i is known
- There are 2000 census tract j zones
- For example there could exist 3 types of w income and 3 types of k housing
- There are 2000 census tract m zones

It is obvious that we shouldn't run 36 million possible population calculations. A large number of entries would be zero. A random test to know who would move in to the new place would not represent the reality well enough. There is the necessity to seek for a microsimulation representation.

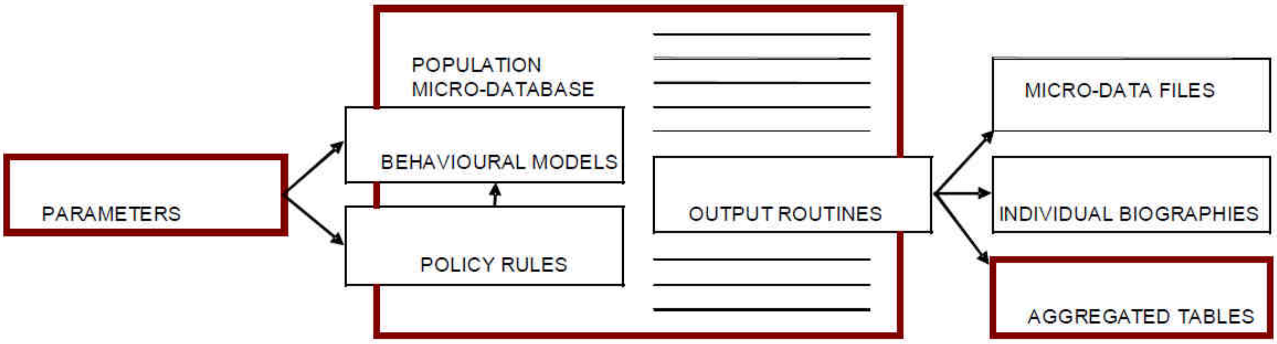
B) Building a microsimulation model

a) What is a Microsimulation?

In the social sciences, a dynamic microsimulation is the projection of the demographic and socioeconomic society settings in simulating action and interactions from a large sample of inhabitants. For that, we need accurate data, complicated models, and heavy tests. The system really implies macro scales, but the main idea of the model can be summarize in two points: First, to model and simulate acts and interactions of the micro smallest scales unit and aggregate results to get macro results. People are making parallel choices sometimes hard to balance, like having babies, doing studies, working, getting married, moving out, divorcing… All these decisions are modelized in the scale in which they are taken. The first point would be to define what the important settings for a move are. Secondly, we need to program the mass behavior effects at a macro scale, which will show us the moving of the complete system. Fortunately I found a powerful tool for helping to construct those complicate simulation. ModGen is a programming language, developed by Statistics Canada, which run on Visual Basic and simplifies the way of building our own model without having a strong knowledge in programming sciences.

For certain reasons, microsimulation is still contested by specialists "imbued with physics envy and ascribing the highest status to mathematical elegance rather than realism and usefulness" (Wolfson 2007). The most critical point is of the stochastic format of decisions made. Since microsimulation produces not expected values but instead random variables distributed around the expected values, it is subject to another type of randomness: Monte Carlo variability. Every simulation experiment will produce different aggregate results. Many repeated experiments on the large populations can eliminate this sort of randomness to find a more redundant solution and deliver valuable information on the distribution of results. Nevertheless it worth building such a model if all the individuals are very different and the number of each combination is too big to class the population in a useful class. In terms of our precise removals problem, it's perhaps not the easiest way to solve it but it's the more realistic and all conditions can be applied to. In addition, this model could be reused in the future to solve other issues in reusing the settings and the sample of the model. This would give us another accurate kind of representation of the people's behavior.

b) How does it work?



Main components of a typical data-driven microsimulation model

The microdata sample is a group in the central core part. The settings define how the simulation is going to work, and is delivered in an aggregate table. All data are stocked in a personal table row, which composes the central core of the simulation. We will have to think about the size of the sample and how to get it. The set of personal options could be increased, creating a lot of all indices slices, or even doing it continuously would make result more accurate. We have to define parameters to give the way the simulation is going to work. The outputs will be new tables of the sample, with the new place of living. The system is able to give an individual biography of a special agent too, but it won't be of use in our case.

c) Settings used

There are two aspects covered in the scope of a simulation – we first distinguish general models from specialized ones, then population models from cohort models. Finally, looking at the methods on how we simulate populations, we focus our discussion in three ways. The first is the population type we simulate; open versus closed population models, as well as cross-sectional versus synthetic starting populations. The second is the time framework used, either discrete or continuous. The third is the order in which lives are simulated, leading to either a case-based or time-based model. Even though the two different types of test don't ask the same questions, we will construct only one model with the same population and universe. Only the parameter will change in order to use a different test.

- The aim of the micro simulation model is to test which populations want to move to a new place or select a new place to live because the current zone is not satisfying. These are specialized questions but a lot of the attributes of the population will need to be modeled, and we don't know about the future of the system. So our simulation will be based on a general model. This is the usual way of building a microsystem model.

- The cohort model would only be used to simulate population segments, in order to study and compare it. So we will use a population model.

- The difference between an open and a closed population model is that the first would allow adding new resident to the sampling, especially for matching spouses or representing immigration. A closed model won't allow new entries. Ignoring immigration in Los Angeles would be a devastating simplification, but the model which will represent the in and out migration is not directly dealing with the housing subsystem, and we can imagine that only people form a part of the city would move to a new place. The Demographcs of LA is quite stable now, so even if it is a big part missing in the model of the city, for this problem we will create a closed system. We could add a "rest of the world" zone if we want to immigration to interfere.

- In population models, we can distinguish two main starting population types: cross-sectional and synthetic. In the first case, we read into a starting population from a cross-sectional dataset, and age all individuals from this moment until death (while of course also adding new individuals at birth events). In the second case all individuals are modeled from their moment of birth onwards. As this doesn't really matter in our model and we don't know a lot about the history of the population, we will opt for a cross-sectional population model.

- We would choose a discrete representation of time rather than a continuous one because it will only be a single test. It will be more useful to represent it in order to ordinate the actions happening. It will be a great tool to represent the slow dynamics.

- The distinction between case-based and time-based models lies in the order in which individual lives are simulated. In case-based models one case is simulated from birth to death Before the simulation of the next case begins in time-base models, all individuals are simulated simultaneously over a predefined time period. Therefore, we will choose this model.

d) The Sample

First, we'll have to define the size of the sample. All Census tracts have to be present on the test. Microsimulation must represent each person individually to deal with a lot of different personal paths. It can't be a cell-grouped information like we used until now.

First, we'll define attributes to the people that matter in a moving choice. Some of them are already used in the previous work.

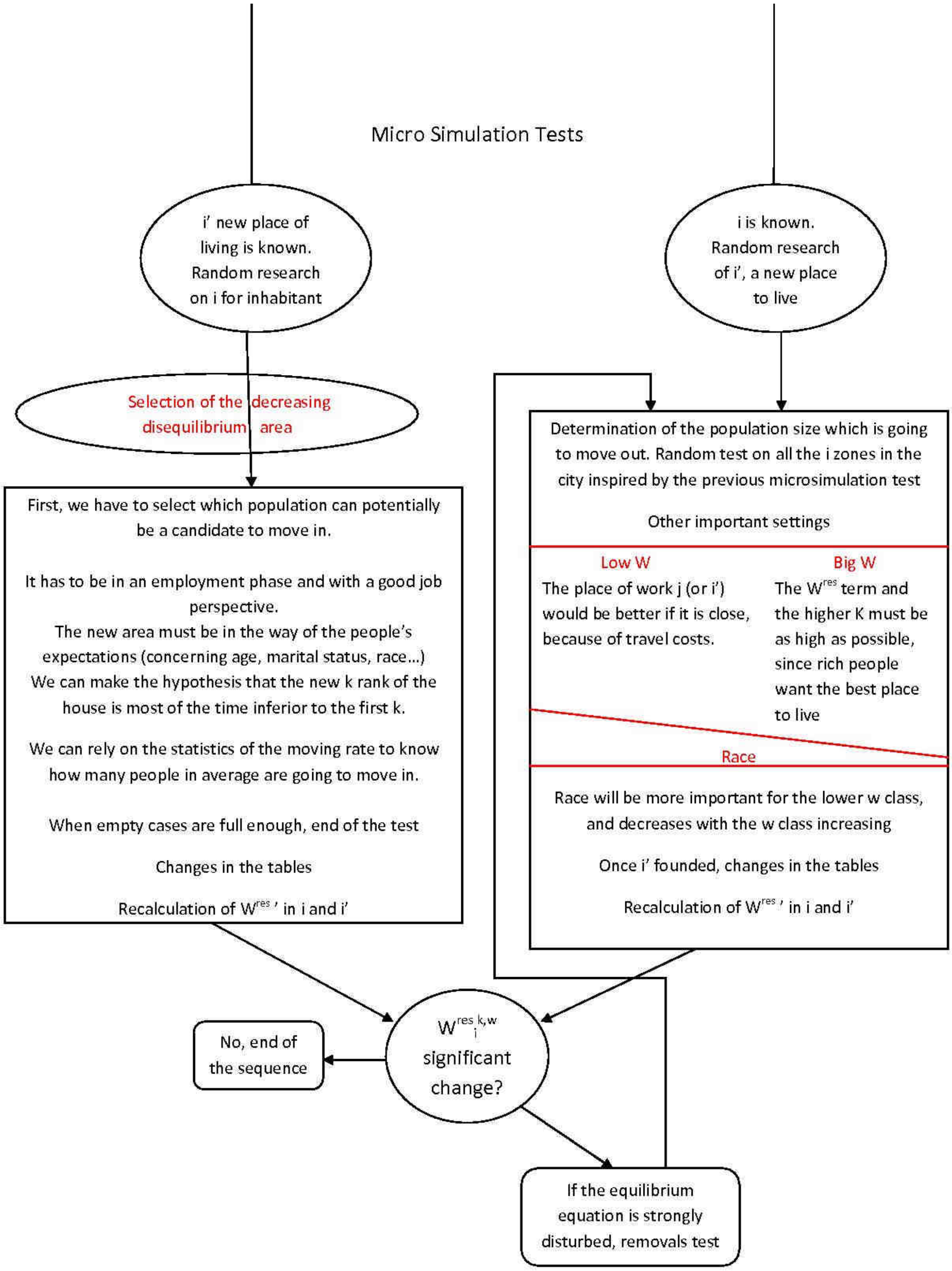
- i : the place of living which will be the only parameter to change. It will be the most important. It will be linked to the rate of the zone W^{res} and its equilibrium situation. People will move out their area only if the equilibrium equation is decreasing.
- j: place of work. It will always be an important information
- k: current kind of house. In most of the case, the new house is not worse than the old one.
- w: class of the household income.
- Race (a lot of more race can be represented in order to represent communities)
- Age
- Sex
- Marital status
- Employment/unemployment status
- Educational level and job perspective
- Immigrant status

Then we define the higher scale effects that can occur to give a general comportment to the population. We can notice these kinds of phenomena:

- Racial segregation in neighborhoods (more for low w income)

- The search for a bigger place which leads the suburbanization of the urban space (more for high w income)

Even with ModGen which really simplifies really the construction of such models, it's a very time consuming task. The housing stock cannot change very quickly and this component has to be strongly represented. As I wasn't able to build the core of the system by myself without the adequate geomantic and probabilistic formation, I haven't tried to build an accurate model of microsimulation. However I explained the mains ideas and how I think the test is going to proceed.


Micro Simulation Tests
i' new place of living is known. Random research on i for inhabitant
i is known. Random research of i', a new place to live
Selection of the decreasing disequilibrium area
First, we have to select which population can potentially be a candidate to move in.
It has to be in an employment phase and with a good job perspective.
The new area must be in the way of the people's expectations (concerning age, marital status, race…)
We can make the hypothesis that the new k rank of the house is most of the time inferior to the first k.
We can rely on the statistics of the moving rate to know how many people in average are going to move in.
When empty cases are full enough, end of the test
Changes in the tables
Recalculation of W^{res}' in i and i'
Determination of the population size which is going to move out. Random test on all the i zones in the city inspired by the previous microsimulation test
Other important settings
Low W
The place of work j (or i') would be better if it is close, because of travel costs.
Big W
The W^{res} term and the higher K must be as high as possible, since rich people want the best place to live
Race
Race will be more important for the lower w class, and decreases with the w class increasing
Once i' founded, changes in the tables
Recalculation of W^{res}' in i and i'
$W^{res\,k,w}_{i}$ significant change?
No, end of the sequence
If the equilibrium equation is strongly disturbed, removals test

When a significant change appears in a zone, it will always modify the W^{res} value. The toleration of the spectrum of a W change can be changed in settings. The system will maybe make multiple phases of changes before it will return to an equilibrium phase. I order to respect the number limitation made by the slow dynamics, we have to control the number of the moves.

Because the data can be saved and can replace the new tables on the core system, users of the software can choose more than one type of change to operate. We could have a political simulation of planners, modifying an area in different ways.

Depending on how heavy the micro test is, we can run multiple numbers of simulations in order to avoid the Monte Carlo effect.

C) Assembling the core data

Once the model is created, I began to conduct the first step of my program: searching for the right data required. In training myself to manipulate ArcGIS, I have been collecting a lot of general data about Los Angeles. The Minnesota Population Center has a website with a powerful Database Management System. We are allowed to select scales, years and topics on a very useful finder requesting their personal DBMS website. An extract of the Housing database and its codebook are provided in annex 1.

I had to deal with structural problems with the types of table files. .cvs and .xls format were incompatible with the Python requests sent on ArcGIS, which needs .dbf table files to have the full set of operations available. I found the evaluation software of DBase, a DBMS software which run the .dbf format. The software allows extremely fast manipulating of the data, but the interface is not convenient at all. I created my folders with Microsoft Access, which can extract tables in the .dbf format. This part has taken more time that I expected, as I didn’t know about all these different types of tables and software used to manipulate them. If the system becomes stronger, we would be able to automate all of these tasks and upgrade the system with more recent data.

Indices building:

w

w will represent the social status of the member, and the best way to model it is with the household income. I began with the hypothesis that the income class is divided into three parts. It’s not a very accurate scale, but in a first program it will be tolerable, and it shows maybe more the effects between three real classes: the poor, the middle class and the rich. Of course, we will easily be able to extend the number of slices in the future.

W Household Income (k$/year)	In Los Angeles County	In US
0-30	36%	35%
30-75	39%	42%
75+	25%	23%

k

k represents the type of house. The census information concerning housing characteristics are obsolete (for example, telephone availability and fuel heating system are not current indices). I tried to class different types of houses, but the dominant kind of house units are the individual house, and this is almost the unique type in most of the zones. Half of the occupied houses in Los Angeles are rented and the other half is owned. I decided to build my different k class on this criterion. As it can't be the only way to characterize a type of house, I added a representation of the value of the house. To not represent the unique value of the house, let's say that three cases of k values will be represented by the population the house is built for: low cost, middle class, and luxury houses. However my only way to model it for now is with the value\bid rent of the house.

Price class	1	2	3
Range for buying(k$)	0-175	175-300	300+
%	36%	36%	28%
Range for Renting ($)	0-550	550-900	900+
%	42%	38%	20%

Race

As we can see on the following table, Los Angeles has very strong ethnic diversity. It may be shocking for the French I am to speak about race like this, and even to use it as a tool to class and operate population calculus, but as we have seen earlier, it is a fact and macro effects are admitted. The American Census Bureau has a different way to classify race, but I used the ones that people are most familiar with.

Group	*Los Angeles County*	*California*	*United States*
White	*31.1%*	*46.7%*	*63.7%*
Hispanics	*44.6%*	*32.4%*	*9%*
Black	*9.5%*	*6.4%*	*12.6%*
Asian	*11.3%*	*10.8%*	*4.9%*
Others	*3.5%*	*3.7%*	*9.8%*

D) The attractiveness term: $W^{res,}{}_{i}{}^{k, w}$

The first topic I wanted to script was the representation of the attractiveness of a type k house in a tract i for a w income household. This is the first step to build the basis of the housing subsystem, and it could give us a first representation of the city. We defined this term as $W^{res,}{}_{i}{}^{k,\ w} = \prod X\hat{}\alpha^{w}$, α is the representation of how much a population of the w class cares about each topic, and X are:

- Accessibility to services
- Affinity to social group
- Bid rent term/affordability
- Available housing stock

This term will be a coefficient between 0 and 1. I am writing the script to modelize this term. The rates I'm using are what I think people in Los Angeles are caring about, but it will certainly change when we move into the calibration phase.

Affinity to social group:

To feel good in one's home area is one of the most important choices in place selection. Two main topics can represent that term: the level of life and the preponderant race of the other neighbors. Like I mentioned earlier, the segregation of people in an area is very strong. Los Angeles is the third largest Hispanic city in America, and a neighborhood like Koreatown could be the third largest city in Korea.

Income (70%)

We will use the percentage of the same w class in the neighborhood. This is of course the most important point people cares about.

Race (30%)

We too will use the percentage of the same race in the neighborhood.

Affinity = 0.7*SameIncome + 0.3*SameRace

Affordability:

It represents the accessibility of the houses price for each kind of income. We have already created tables for both the k types, renting or owning, with 3 slices of prices and rent to figure out how classes can afford the price of living.

I took the hypothesis that buying a house of the price class which falls in the w income level is represented by 0.75. If an income class wants to buy or rent a lower price class house, its affordability will increase by 50%. If the income class wants to buy or rent a better class than his rank, its affordability will decrease by 50%.

W	K price class	Coefficient
1	1	0.75
	2	0.37
	3	0.18
2	1	0.87
	2	0.75
	3	0.37
3	1	0.93
	2	0.87
	3	0.75

Stock Houses

It represents the supply of houses in the tract. The basis I used to build the rate of how attractive the attractiveness of the zones are, if there are too much houses on sale it's a mark on desertion of the area, and if it is negatively regarded by those with higher incomes. By the supply and demand rules, it means prices are lower, so it's better viewed by the lower w class. I used the same thinking for places with a low house stock available.

Ratio	Population	W	Rate
0-1.99	338	1	0,5
	16,4%	2	0,7
		3	1
2.00-3.99	930	1	0,8
	45,3%	2	1
		3	0,9
4.00-6.99	503	1	1
	24,5%	2	0,8
		3	0,7
7.00-9.99	159	1	0,7
	7,7%	2	0,6
		3	0,5
10-100	117	1	0,5
	5,7%	2	0,4
		3	0,3

Accessibility to services:

I & w

This term has to represent how consumers value all the services, private or public, close to their home. The strongest hypothesis of this part is that nobody walks in the Los Angeles streets.

I found a Geodatabase Feature Class of points which represents a lot of points of interest in the Los Angeles County, cooperatively built by the County's Location Management System (LMS). It registers more than 66 000 different locations modeled by a point, in 270 unique types, that gives me all the information I was needed to accomplish this task.

This submodel will later be divided in two: the private and public sector. For now, they are just exogenous values that will all be combined.

Schools *temporary rates: 0.4*

In speaking with people, I heard that access to the highest rated public school seems to be one of the biggest influence in moving to a new place. The Layer Services represented all types of school but without the rates, and I haven't found a useful database of schools which did. Therefore, I had to build it. Ranks are edited on the Academic Performance Index (API) published by the California Department of Education. The scores given are out of 1000, and I found a website which reinterprets these data to make them accessible to families: school-ratings.com. It's a site people are looking at (the second most popular Google result) and it deals with official data. I asked for a raw database file, but they never answered, so I filled the ArcLayer services by hand. As the layer was protected, I created a new row in utilizing the .dbf table of the service's geobox, and filled it in ArcMap because Dbase failed to interpret the number I was writing.

For the higher w income class, I made the hypothesis that children are being put into private schools, which are always good. For this class, we will test only on the presence of a private establishment in the proximity of the zone.

Public services *0.2*

It ranks the presence of police centers and fire stations near the tract. People care a lot about their security and want to be protected the best way possible.

Rank: 70% for the police and 30% for the fire station in a five-miles square

Retails center *0.2*

The size and number of the retail centers

Attractions points *0.2*

The presence of important points of touristic interest, like beaches & marinas, parks, museums, and recreational centers

My Program :

I wrote the main script of this program. Once problems about the use of ArcGIS in Python orders were resolved, the first three topics were easy to solve. The public school system is harder. It's divided into zones, and people living in them have to go to the school assigned to that zone. As it is not the same as the Census Tract, I had a problem in constructing the program. The solution I thought of was to use the GPS proprieties of the GIS system and input the address of the sample, but it is too individual of a solution. As my period of research is almost done, I will try to find the solution before my presentation.

I attached in Annex 2 the beginning of the script I wrote. It is divided into three folders. One contains the data basis, like the names of tracts. On the other are the definitions of the functions and the class I created, and the last contains the main part of the script.

IV) Conclusion

A) The Continuing Task

This is really a massive subject that demands a lot in different fields or study. Once this first model is built, I'll have to choose another subsystem which is linked to housing and conduct the same process with other hypotheses and schemes to build it up. With the experiences, it shouldn't be that hard to build the other ones, and some topics like microsimulation can be reused. Once all of them are working, coordinating them to each other will be harder. But like I said earlier, in many fields a lot of progress has been made in much more complicate way. It can be a thought-provoking task for a team comprised of engineers, computer scientists and urban planners.

Like I said in my introduction, I will change my major next year to acquaint myself with Geographical Sciences, how to program on a GIS system and many more useful proficiencies. The goals of this formation are axed in three points: first, to master the concepts and tools of informatics, secondly, to analyze and process with geographical data and lastly, supervision and project management. I would be able to work efficiently on a stronger geographical basis. As I know the subject well, I will be able to program and think of other solutions during my next year's classes, and if the opportunity is still available, I would love to enter in the PhD program at the University of Southern California. It would be a very gratifying continuation of my professional plans, and I would love to carry on exploring and materializing this fascinating subject.

B) Bibliography

Articles

M. Clarke and A. G. Wilson. *The dynamics of urban spatial structure: the progress of a research programme*. The Royal Geographical Society : 1985

Alan Wilson. *Entropy in Urban and Regional Modelling : Retrospect and Prospect* (2010)

Morton E. O'Kelly. *Entropy-Based Spatial Interaction Models for Trip Distribution* (2010)

Alan Wilson - *The general urban model: Retrospect and Prospect* (2009)

Alan Wilson - *A general representation for a comprehensive urban and regional model* (2006)

California Department of Finance, Economic Research Unit. *California Statistical Abstract 2000*

Jørgen Lauridsen, Niels Nannerup and Morten Skak *Dynamic and Geographic Patterns of Home Ownership.* Discussion Papers on Business and Economics No. 9/2006

Websites

Census Bureau of United States - http://www.census.gov/

Minnesota Population Center. *National Historical Geographic Information System: Version 2.0*. Minneapolis, MN: University of Minnesota 2011 https://www.nhgis.org

Los Angeles County - *Location Management System (LMS)* – http://egis3.lacounty.gov/lms/

Environmental Systems Research Institute - http://www.esri.com/

C) ACKNOWLEDGEMENTS

Dr. Roger Ghanem

For giving me the opportunity of conducting this period of research, helping me with problems and introducing me to the wild world of research and the opportunities it has to offer.

Dr. Hanene Souli & Jacques Lipp

For their presence at the ENISE, and their presence during the video conference presentation.

Hélene Hennion

For her help in the function of international relations at ENISE

The PhD student team

For their kindness and their help in bringing a fresh perspective into my project.

Dr Laetitia Dablanc

Specialist of fret transportation, who has helped and she helped me with a French view of what it is to live in Los Angeles.

Finally, I would like to thank all the people who, one way or another, helped me during my internship, also on my project and in my life in America.

D) Annex

I a) : Extract from a NHGIS table codebook.

I b) : Extract from a database table

II a) : donnee draft script

II b) : function draft script

II c) W^{res} draft script