Description about the CLS in the tax99 data

Kyoung Hee Kim, Rao Fu, Adityanand Guntuboyina

Dec 2

Contents

1. Introduction

We are analysing data about properties in New Haven for the year 1999. The raw data for this topic came in the form of several spreadsheets from the Mayor's office. For each of 29925 properties, we have information about the location (Street number, name and Unit number), ownership (Owner names and their addresses), construction detail (Number of units and "CLS"), recent tax assessments (Land, Building, Miscellaneous and Gross Assessments for the year 1998) and recent sales information (Sale date and sale price for the most recent transaction involving which the property).

Our goal for this study is to learn something about tax assessments. For example, we are interested in finding meaningful answers to questions like:

1. Is there any truth to the urban legend that assessed values are grossly smaller than real values?

2. Are there any patterns that might suggest that some parts of the city get better treatment?

3. Where are the more valuable properties in New Haven? Does this pattern agree with what residents tell the folks who take the Decennial Census?

Unfortunately, upon careful investigation of the data, we found out that someone has worked on the Excel spreadsheets without exercising enough care. For example, each property in the dataset is supposed to have a unique identification code called the MBP. However we found out there are multiple records having the same MBP. It turned out these multiple records with the same MBP actually refer to the same property.

Also, there are some variables in the dataset whose meaning is not clear. For instance, the variable "CLS" assigns a code (A, C, E, I, O, R or U) to each of the properties. It is not clear what this code stands for. We have a feeling that this is related to the construction type of the property.

We decided that before analysing the data, we need to spend some time trying to understand the data better and for validity, compare the information in the dataset to some alternative sources of the data, if we can find some. Fortunately, The Assessors Online Database for New Haven also has similar and more detailed information about the various properties of the city.

To start with, we first tried to figure out the meaning of the CLS codes by comparing the tax99 database to the Assessors Online Database. In the next few sections, we discuss the basic methods that we employed in doing this and also the results that we obtained.

2. Data description

2.1 Tax database (tax99)

We are analysing data about properties in New Haven for the year 1999. We have been provided the raw data in the form of a few Excel spreadsheets which we had to combine first. Unfortunately, the data came with no documentation. In this short write-up, we describe basic features of the data: for instance, how big the dataset is, what do each of the variables mean, how does a typical record looks like etc.

Let us start with the size of the dataset. It has 29925 records. Each record represents a property in the city of New Haven. Apparently, the dataset is supposed to contain all the properties in New Haven (atleast those built before the year 1999). But we have no evidence to confirm this. For each property, we have information on 23 variables. The first of these variables (MBP) acts as a unique identifier for the properties. MBP stands for "Map Block Parcel". The city is divided into maps, which are then divided into blocks. The blocks are further divided into parcels. The map number (3-digit), block number (3-digit) and parcel number (5-digit) of a property uniquely identifies it. The combination of these three numbers has been termed MBP.

The remaining 22 variables can be classified in to the following groups:

1. *Location variables*:
The location of a property in the dataset is specified by four variables: Street Number ("ST.NO."), Street Name ("ST.NAME"), Unit Number ("UNIT.NO.") and Zip Code("ZIP"). There are 699 street names in all.

2. *Ownership Variables*:
There are six variables giving details on the ownership of the property: names of the first ("OWNER.NAME1") and second ("OWNER.NAME2") owners, and the mailing address of the first owner("MAIL.ADDRESS", "MAIL.CITY", "MAIL.ST" and "MAIL.ZIP").

3. *Construction Variables*:
There are a couple of variables giving details about the construction type of the property: The number of units ("NO..OF.UNITS") and the CLS code ("CLS"). The CLS variable assigns a code (A, C, E, I, O, R and U) to each property. There is no explanation given as to what these codes stand for. We are guessing that they give information on the construction detail. For example, "R" could mean residential, "C" commercial and so on. We shall come back and say something more about this.

4. *Assessment Variables*:
There are four variables ("GROSS98", "LAND98", "BLDG98", "MISC") which give information on the assessed values of the property for the year 1998. "LAND98" and "BLDG98" give land and building assessments respectively. "MISC" gives some additional assessment and finally, "GROSS98" is the sum of all these assessments and is the net assessed value for the property.

5. *Sales Variables*:
There are two variables ("SLE.DATE" and "SLE.PRICE") which give information on sales involving that property. There is atmost one sale date and price for a property and so we are guessing that these correspond to the most recent sale involving the property.

6. *Others*:
There are four other variables in the dataset ("VOL", "PAGE", "INST" and "CODE"). Volume ("VOL") and page ("PAGE") most probably refer to the physical location in the City Hall where the property details are recorded. We are unclear as to what exactly the variables "INST" and "CODE" stand for.

We shall see some more features about the data in this section. Let us start by looking at a few typical records ([Table 1]). We do not show all the variables for these properties but only a few important ones. A few immediate observations can be made from [table 1]. For the second property in the table, there is no Sale Date but there is a Sale price of zero. The property is owned by the Parks Department and presumably, they got it for free. But we do not know when they aquired it. There are a few other properties as well that have a Sale Price but do not have a Sale Date.

Also, there are a lot of properties in the dataset which have a sale price of zero. In fact, 62.34% of all the properties have a zero most recent price. This is just too unsual to be true: we cannot believe that most properties around are being sold for a zero price. One of our tasks in the analysis of this dataset is to find out the meaning of at least some of these zero prices.

For the numeric variables in the dataset: Assessment variables and Sale Price, [table 2] gives their summary statistics. This table shows that the maximum sale price is about 892 million dollars but the maximum assessed valued is only about 86 million. Perhaps, the entry for that sale price is incorrect by a factor of 10.

MBP	CLS	OWNER.NAME1	OWNER.NAME2
303 0065 02700	R	BANKS ROBERT	
208 0550 00100	E	WOOSTER SQUARE	PARKS DEPT
208 0549 01502	O	BADER JOEL S	
161 0785 00900	R	WEBB	SAMUEL L & SUSAN R

MAIL.ADDRESS	MAIL.CITY	MAIL.ST	MAIL.ZIP
22 ADELINE ST	NEW HAVEN	CT	06519
720 EDGEWOOD AV	NEW HAVEN	CT	06515
12 ACADEMY ST UNIT 2A	NEW HAVEN	CT	06511
109 ATWATER ST	NEW HAVEN	CT	06513

ST.NO.	ST.NAME	UNIT.NO.	ZIP	NO..OF.UNITS	GROSS98
22	ADELINE ST		06519	1	39382
	ACADEMY ST		06511	0	497490
12	ACADEMY ST	2A	06511	1	108500
109	ATWATER ST		06513	1	78141

LAND98	BLDG98	MISC	VOL	PAGE	SLE.	SLE.PRICE	INST	CODE

					DATE			
15092	24290	0	4932	152	11/15/95	11000	WD	B
497490	0	0				0		
0	108500	0	5116	306	3/13/97	115000	WD	A
29701	48440	0	4838	206	3/1/95	85000	WD	A

[Table 1] A few typical properties

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SLE.PRICE	0	0	0	138,800	59,000	892,500,000
GROSS98	0	57,080	80,520	196,600	109,100	86,190,000
LAND98	0	12,730	21,490	45,900	32,560	15,790,000
BLDG98	0	38,500	56,910	142,100	77,980	83,180,000
MISC	0	0	0	8,578	2,450	35,360,000

[Table 2] Summaries for numeric variables

2.2 Assessor Online Database (DVA)

The DVA dataset comes from the website: http://data.visionappraisal.com/NewHavenCT/search.asp. From that website, we can learn detail information of properties. The MBLU numbers(3 digits for Map, 4 digits for Block, 5 digits for Lot and Unit), locations, and current owner names are given on the top of webpage. The assessment values in year 2007 and 2005 are shown in part of Parcel Value. Also detail information is provided by tables of Owner Record, Ownership History, Land Use, Land Line Valuation, Construction Detail, Building Valuation, Outbuildings, Extra Features, Building Sketch, and Subarea Summary.

There are several search options available online, such as, query by locations, by owner names, by MBLU numbers, or by version internal IDs. Apparently, searching by version internal IDs (also called Parcel ID) is a good way to get complete records of all properties because PID are unique across all properties and are coded by a sequence of numbers starting from 1.

Mike checked that there are no parcels with PID beyond 50000. Using Mike's Perl script GetRealEstateData.pl, we get 27307 properties, among which 508 PIDs show "no parcel found", and the information for PID 8162 has been suppressed at the owner's request. So finally we get 26798 properties with useful information. And there are some gaps among their PID numbers. Here I have to mention that we are not sure whether or not this is a complete record for all properties of New Haven.

Moreover, we also miss something, because there are some properties with multiple buildings such as pid4911 and pid14090. The information about Construction Detail, Building Valuation, Outbuildings, Extra Features, Building Sketch, and Subarea Summary for each building are shown on separate webpages. In our script, we will miss those multiple parts; only get information for one building.

David combines all pid files into a file allPID.txt using the script getNHR5.pl. Each line in the text file contains information of one property. It begins with "pid#\PARCEL SUMMARY (MBLU: ###/ ####/ #####/ //; Location:……, … …)", then several tables follow. For each table begins with <table> tag and ends with </table> tag. The rows within tables are separated by <nl> tags. The fields within rows are separated by characters (\t).

Here is an example after splitting the whole line by “</table>”.

```
> strsplit(x[1],"</table>")
[[1]]
 [1] "pid1\tPARCEL SUMMARY ( MBLU : 013/   0853/   00400/    / /; Location:51 SOUTH END
RD)<table> 51 SOUTH END RD <nl>   \t MBLU :\t013/  0853/  00400/   / /  <nl> Location:\t51
SOUTH END RD  <nl> Owner Name:\tCITY OF NEW HAVEN AIRPORT  <nl> Account Number:\t   <nl>"
 [2] "Parcel Value<table> Item\t Current Assessed Value\t\t\t\tFY 2005 Assessed Value  <nl>
Buildings\t0\t0    <nl>  Extra  Building  Features\t0\t0    <nl>  Outbuildings\t0\t0    <nl>
Land\t866,250\t625,730  <nl> Total:\t866,250\t625,730"
 [3] "Owner of Record<table> CITY OF NEW HAVEN AIRPORT  <nl> CITY OF NEW HAVEN"
 [4] "Ownership History<table> Owner Name\tBook/Page\tSale Date\tSale Price  <nl> CITY OF
NEW HAVEN AIRPORT\t\t7/23/1975\t0"
 [5] "Land Use<table> Land Use Code\tLand Use Description <nl>  <nl> 902V\tCITY  MDL-00"
 [6]  "Land  Line  Valuation<table>  Size\tZone\tNeighborhood\tAssessed  Value    <nl>  3.80
AC\tAIRP\tFLY\t866,250"
 [7] "Construction Detail<table> Item\tValue  <nl> STYLE\tVacant Land  <nl> MODEL\tVacant"
 [8]  "Building  Valuation<table>  Item\tValue    <nl>  Living  Area\t0  square  feet    <nl>
Replacement Cost\t0  <nl> Year Built\t  <nl> Depreciation\t100%  <nl> Replacement Cost Less
Depreciation\t0"
 [9] "Outbuildings<table> Code\tDescription\tUnits  No Outbuildings"
[10] "Extra Features<table> Code\tDescription\tUnits  No Extra Building Features"
[11] "Building Sketch  <strong>Vacant Land, No Sketch</big>"
```

[Figure 1] Example

If we need any information, we can extract them according to this format. Here we give an overview of some tables we are interested in.

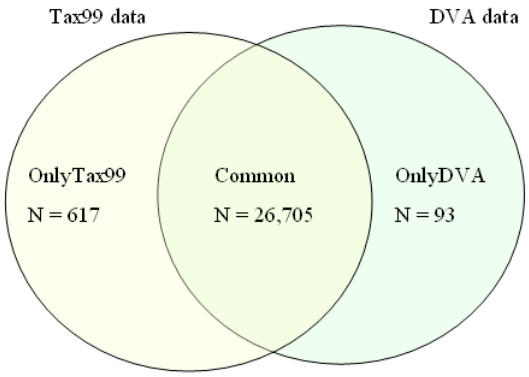
Name	Overview
MBLU	12 digits to represent map, block, lot and unit numbers. It roughly

<table>
<tr><td></td><td>corresponds to MBP in TAX99 data.</td></tr>
<tr><td>Parcel Value</td><td>Assessment values for buildings, extra building features, outbuildings, land, and total values in year 2007 and 2005 are given in Parcel Value table. The format of money is ###,###. If we want to do some calculations, we need convert them into numeric form.</td></tr>
<tr><td>Ownership History</td><td>Owner Names, Book/Page, Sale Date, Sale Price of repeated sale for each property are given in this table. This may be not the complete history.
The sale price is represented by ###,###.
The sale month, date, and year are split by “/”.
Looking at sale year, there are some strange values: 247, 704, 705, 829, 1502, 2014, 2020, 2106, 2205. The range of others is 1916 – 2006. There are1963 missing values for it.
We are not sure the meaning of Book/Page.</td></tr>
<tr><td>Land Use</td><td>Land use code and land use description are given in this table. There is a list of codes and descriptions online.</td></tr>
<tr><td>Construction Detail</td><td>STYLE and MODEL are two variables we are interested in, because they are helpful to explain the meaning of CLS in TAX99. There are 96 styles and 8 models in our data.

<u>Summary table of MODEL</u>
<table><tr><td>Com Condo</td><td>Commercial</td><td>Ind/Lg Com</td><td>Multi -Family</td><td>Res Condo</td><td>Residential</td><td>Svc Sta/Gar</td><td>Vacant</td></tr><tr><td>172</td><td>2180</td><td>360</td><td>246</td><td>3123</td><td>17400</td><td>211</td><td>3106</td></tr></table></td></tr>
<tr><td>Building Valuation</td><td>In this table, we can learn living area, replacement cost, year built, depreciation, and replacement cost less depreciation of each property.
The unit of living area is square feet, the form of it is ###,###.
The range of built year is 1763 to 2006. And there are 3106 missing values for it</td></tr>
</table>

3. Merging the two data sources

In order to understand what the CLS code in the tax database(tax99) means, we needed to find related information in Assessor Online Database(DVA). While tax99 has a unique ID called "MBP", DVA data contains MBLU code (which will be used later as an alternatives for "MBP"). After deleting some rows not having any information, the number of rows of the DVA was 26,799. Among these, one row whose MBLU is "166 0714 00800" has interesting comment such as "The information for this parcel has been suppressed at the owner's request". Without this data, we compared MBP and MBLU. Though the number of rows in MBP (27,323) is more than that in MBLU, MBP does not contain the whole MBLU. Thus, common part of MBP and MBLU could be used for comparison if matched MBP and MBLU data indicates the same property.



[Figure 2] Two data sources

Using the two data sources, the first thing we need to check is whether the intersection of two IDs(MBP, MBLU) have the same information. For this purpose, we extracted street name and owner name from Common part in Tax99 data and DVA data and compared them.

Unfortunately, we found that 14 cases among 26,706 properties contain different information even though their MBP and MBLU are the same. From detailed information in DVA and google maps, we noticed that these are the cases where they have either different street name because the property is adjacent to each street, or different owner name because the house is sold after 99. Thus, we regarded two IDs (MBP, MBLU) in two sources are the same.

Now, we can combine two datasets in terms of MBP (or MBLU). We extracted "Land Use Code", "Land Use Description", "Style" and "Model" from DVA datasets which presumaly

help to describe the CLS code. Then, we combined tax99 data and extracted variables from DVA data in case both data have the same MBP and MBLU.

4. Meaning of the “CLS” codes

In the merged dataset, we have more construction detail information about each of 26705 properties (Construction style and model) as well as information on land use. Our hope is that we can use this information to figure out the meaning of the "CLS" codes. [Table 5] classifies all the properties (in the merged dataset) according to their CLS code and the construction model:

Model/CLS	A	C	E	I	O	R	U
Com Condo	0	2	12	0	154	2	0
Commercial	434	1,179	553	12	0	16	2
Ind/Lg Com	0	144	69	126	0	1	18
Multi-Family	233	9	2	0	0	2	0
Res Condo	0	0	15	0	3,074	2	0
Residential	466	177	179	0	7	16,652	1
Svc Sta/Gar	0	190	12	8	0	0	1
Vacant	22	528	870	163	46	1,424	18

[Table 5] Classification according to Construction Model and CLS

From the 2229 properties with CLS code C, 528 are vacant land. Among the other, 77.78% of the properties have either a commercial or an Industrial construction model. This makes it plausible to believe that the CLS code "C" stands for "commercial". Likewise, all the properties with CLS code "I" that are not vacant land have an Industrial Construction model. Hence, there is a good possibility that "I" stands for "Industrial". We can also check this using another construction detail: style. [Table 6] gives the construction styles of all the properties with CLS code "I".

Style	Properties	Style	Properties
Apt 1-7 St	2	Office	7
Cold Storage	2	Research/Devel	4

Food Process	2	Self Storage	1
Garage	5	Service Garage	2
Garage/Office	4	Vacant Land	163
Ind/Office	10	Ware House	37
Industrial	55	Whse/Office	6
Mill Bldg	9		

[Table 6] Construction Styles for "I" Properties

We can see that apart from the ones which are just vacant land, the majority of the rest have the style: Industrial, Warehouse or some kind of Office. Thus, it seems that "I" most likely means "Industrial".

Let us now get to the CLS code "O". From [table 5], we see that all but 7 non-vacant land properties with code "O" are condos (commercial or residential). We can infer that "O" means "Condominium". We can also check this by looking at the Construction Style of these properties ([table 4]). Again, we see that all these styles are related to condominium. Hence "O" stands for condominium. Since, the code "C" has been used for commercial, the folks at the Mayor's Office had to find a different code for condominium and then they thought of "O".

Style	Properties	Style	Properties
Condominium	1,613	Raised Ranch	1
Duplex	2	Retail Condo	21
Indust Condo	3	Row House	4
Med Offc Condo	83	Townhouse	1,461
Office Condo	47	Vacant Land	46

[Table 7] Construction Styles for "O" Properties

We have taken care of three codes (C, I and O). Now let us turn our attention to "R". We would obviously guess that the meaning of this code must be Residential. Let us check this from [table 5]. We find that 99.86% of properties with code "R" that are not just vacant land have the residential construction model. Clearly, this is enough evidence to take "R" to

mean "Residential".

APT 4-Unit	APT 5-12 R	APT over 12	APT5-12	Others
406	229	213	215	92

[Table 8] Land Description for "A" Properties

For understanding the CLS code "A", we looked at the land description of the properties with code "A". The resulting table is [table 8]. The evidence is convincing enough to conclude that "A" stands for Apartment.

Only two codes remain: "E" and "U". We have summarized the land desciption details for the properties with code "E" in [table 9]. Looking at the descriptions that mostly include the CITY, STATE, CHURCH and other EDUCATIONAL properties, we find it highly persuasive to believe that "E" stands for "Exempt".

Desc	Properties	Desc	Properties
CEMETERY	33	MUNICIPAL	465
CHARITABLE	9	NON PROFIT	90
CHURCH	62	PARISH HSE	2
CITY	94	PVT (Hosp & Educ)	227
EDUC BLDG	4	REC FACILITY	76
EX CONDO	12	RELIGIOUS	202
EXEMPT	31	STATE	75
EXEMPT COM	56	US GOVT	6
FRATNL ORG	2	VACANT	38
HOUSING AUTH	100	VET ORG	2

[Table 9] Land Descriptions for "E" Properties

For the code "U", we have [table 10] that classifies the properties with CLS code "U" according to their Owner Names. We infer that "U" stands for "Utilities".

Owner Name	Number of Properties
Connecticut Properties LLC	1

Jenmar LLC	1
Southern Connecticut Gas Company	2
Southern New England Telephone	8
United Illuminating Company	28

[Table 10] Owner Names for "U" Properties

Finally, Table 11 summarizes our understanding of the CLS codes:

CLS code	Meaning
A	Apartment
C	Commercial
E	Exempt
I	Industrial
O	Condo
R	Residential
U	Utilities

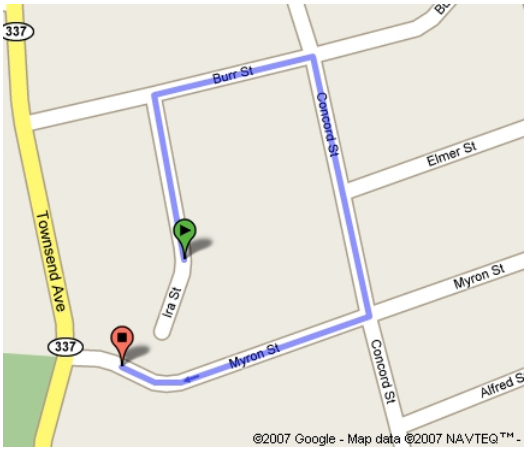
[Table 11] Meaning of the CLS codes

5. Anomalies

From now on, we call the data in the tax99 which is not included in DVA as "OnlyTax99" data. Similarly, we call the data in the DVA which is not included in tax99 as "OnlyDVA" data.

5.1. Data in both tax99 and DVA with same MBP and MBLU but different street name and owner name

As I mentioned in section 3, there were 14 cases[1] of this kind. All cases have different street name because the property is at the border line of both streets. The following figure shows this property is adjacent to MYRON ST and 1A IRA.



[Figure 3] Google map of the streets of the property (MBP: 029 0892 00900)

Among 14 cases, we also found that 9 properties have different owner name. This is because tax99 data does not reflect the changes after 1999.

5.2. Data in tax99 but not in DVA

As we can see in the figure 2, there are 617 properties in OnlyTax99 data. From the analysis we have done in the previous section, we want to understand the characteristics of this data.

[1] MBPs of those kinds are as follows: 029 0892 00900, 066 0951 00200, 163 0725 00900, 167 0768 00400, 167 0769 00100, 209 0591 00200, 224 0578 00500, 246 0362 00100, 250 0493 00100, 257 0351 00100, 279 0177 00100, 290 0469 02201, 321 0323 00100, 339 0210 01200

Data/CLS	A	C	E	I	O	R	U	Total
OnlyTax99	20	106	154	14	52	270	1	617
Common	1,155	2,229	1,692	309	3,281	17,999	40	26,705
Total	1,175	2,335	1,846	323	3,333	18,269	41	27,322[2]

[Table 12] CLS in OnlyTax99 data

From [table 12], we notice that the ratio of E is much higher in OnlyTax99 data than that in Common data. Also, from the summary statistics of OnlyTax99 data, we found that 22.7 percent of owner name were either Yale University or City of New Haven. Compared to the 3.2 percent in Common data, 22.7 percent is considerably higher. Thus, we can reconfirm the meaning of "E" as exemption.

On the other hand, we can regard these properties as those whose status changed after 1999, so that these properties no longer have the same MBP number (or, MBLU). If there existed the same street name in the DVA data, we could utilize the ownership history information. If there were the same owner name in the history information, we can conclude that the MBP has changed after 1999 for some reasons. Unfortunately, after careful checking, there were no same streets in DVA data. However, interestingly, we were able to obtain supplemental information directly from online database. For example, the property whose address is "1 Admiral st" was found from the visionappraisal webpage[3].

Information	OnlyTax99	DVA	Information from direct URL
MBP(=MBLU)	282 0347 00900	No information	282 0347 00106
Location	1 Admiral St	No information	1 Admiral St
Owner	Elm Haven Homeownership	No information	Kinkade Alphonse
Previous owner	No information	No information	Elm Haven Homeownership

[Table 13] Case of 1 Admiral Street

[2] This number is one less than 27,323 (number of rows in Tax99 data) since we deleted one row (MBLU : 166 0714 00800, pid 8162) saying that "PARCEL SUMMARY (MBLU : 166/ 0714/ 00800/ / /; Location:85 CHAPEL ST)</title> The information for this parcel has been suppressed at the owner's request.")

[3] http://data.visionappraisal.com/NewHavenCT/findpid.asp?iTable=pid$\backslash$\&pid=100182

Table 15 implies that our DVA data does not contain the all updated data from the website.

5.3. Data in DVA but not in tax99

In this category, we have 93 properties. Among these, 14 properties are located in the "Whalley" street, 13 cases are in the "Prospect" and 9 cases are in the "Ashmun" street. Also, the owner of the 13 cases were WHALLEY AVENUE FUNDING LLC. [Table 14] shows the model of OnlyDVA. From this, we notice that 38 percentage of the properties in this category are vacant while 34 percentage are Residentail Condo.

	Com Condo	Commercial	Ind/Lg Com	Res Condo	Residential	Vacant	Total
Frequency	2	4	2	32	18	35	93

[Table 14] Model of properties in OnlyDVA data