New matching for Housing, Health and Happiness

Maximilian Ben Ali maximilian.benali@epfl.ch
Quentin Bernet quentin.bernet@epfl.ch
Anne Donnet anne.donnet@epfl.ch

Abstract

In this paper we implemented new matchings, further confirming the results of the paper "Housing, Health, and Happiness" (Cattaneo, Galiani, Gertler, Martinez, and Titiunik, 2009). We explored three different feature sets with three different distance functions for the matching, and observed the differences on two different tables of the original paper.

1 Introduction

We will be extending the paper "Housing, Health, and Happiness" (Cattaneo, Galiani, Gertler, Martinez, and Titiunik, 2009), which investigates the effects of Piso Firme (a Mexican program to replace dirt floors with cement floors in households) on the health of the children and the happiness of their mothers. To draw their conclusions they used two similar Mexican cities which are geographically close, one of them having benefited from the Piso Firme program and the other not. They used a census from 2000 that had been done in both cities to match census blocks from the treated city with ones from the control. Their matching was made using different households features averages. Our addition will be to add matching on the household or individual level to answer the following:

- **Q1:** Would the paper's results have changed if they had added matching on the households?
- **Q2:** How do the different distance functions impact the results?
- **Q3:** How does the choice of matching features impact the results?

2 Data Collection

We used the datasets provided with the paper, and replicated the cleaning steps. More specifically, we used their provided information of the 2000 Mexican census and the 2005 survey.

3 Dataset Description

We worked on two distinct datasets, one containing the information at the household level and the other with information at the individual level. In these datasets, every row describes a household (resp. a person), with columns describing different characteristics, such as the number of household members, whether animals are allowed to enter the house, the number of times respondents washed hands the day before the survey, total transfers per capita from government programs (resp. Parasite count, McArthur Communication Development Test score, age). Table 1a shows the size of the different datasets and Table 1b shows the sample sizes after cleaning.

(a) Datasets statistics

	Household dataset	Individual dataset
number of element	2'783	6'693
number of features	77	88

(b) Sample Sizes in 2005 Survey

	Control	Treatment
Share of rooms with cement floors	1393	1362
Cement floor in kitchen	1393	1362
Cement floor in dining room	1393	1362
Cement floor in bathroom	1393	1362
Cement floor in bedroom	1393	1362
Parasite count	1566	1528
Diarrhea	2105	1930
Anemia	1951	1768
MacArthur Communicative Development Test score	302	291
Picture Peabody Vocabulary Test percentile score	817	757
Height-for-age z-score	2053	1865
Weight-for-height z-score	2058	1881

4 Methods

We started by replicating two of the paper's tables, trying to follow the same steps they did. The goal of this replication was to validate that any dispar-

Table 2: Sample Sizes after T5_match

	Control	Treatment
Parasite count	1419	1489
Diarrhea	1887	1884
Anemia	1750	1724
MacArthur Communicative Development Test score	267	283
Picture Peabody Vocabulary Test percentile score	739	742
Height-for-age z-score	1839	1825
Weight-for-height z-score	1843	1837

ities between their results and our post-matching results would be due to the matching and not to any misunderstanding of their processing or to any difference between the used data.

We chose two different tables, one describing the regression of the cement floor covering measures and one describing the regression of the children's health measures, respectively labeled *Table 4* and *Table 5*[1] in the original paper. We chose these tables because together they cover most of the data provided in the paper.

To replicate the tables, we first identified the dependant and independent variables, the treatment variable and the clustering variable. The two tables have three variation of the independent variables forming three different models, the first one without any control variable, the second with age, demographic, and health-habits control variables, and the third adding public social programs to the control variables. All the variables for the different tables are more thoroughly described in Annex A.

As described in the original paper, we wanted to drop all the data for which complete geographical data was unavailable. Looking at the sample size they had to compute their regressions, we noticed that we had more rows for which geographical data was missing than they announced, maybe due to some data corruption. Since the dataset ended with a cluster of as many rows missing geographical information as they had dropped, we assumed this was it and only dropped these final rows. Moreover, we confirmed that we had the expected number of non-`NaNs` values for all dependent variables, which further supports our decision.

Missing values in columns containing the independent variables were imputed with 0 and a dummy variable indicating whether the value was missing was added resulting in updated models. Those models were used to compute the regressions.

For the matching, we decided to use three different distance functions: L_1 (taxicab), L_2 (euclidean), and L_∞ (Chebyshev). We set our choice on these functions because they are widely used, and because Chebyshev distance was the one used in the original paper's matching.

To construct the feature sets, we looked at all the features that were not used for one of the regressions, then we studied what meaningful groups we could do with them. For the first table (*Table 4*), we created two different feature sets for the matching. The first one (T4_eco_match) contains eight economy-related variables[2], and the second one (T4_house_match) five house improvement-related variables[3]. For the second table (*Table 5*) we created only one economy-related feature set (T5_match)[4]. This is also why we use only the L_∞ distance in T5_match: in one dimension, all L distances are the same. The sizes of the datasets after matching are in annex C.

5 Results

Regarding the 6 different combinations of distance functions and feature sets for the matching for *Table 4*, none of them resulted in any major difference. As such, the results of our regressions can be found in Annex B. Table 3 shows a comparison of the paper's results with the results of our replication of *Table 5* and the regression with matching using L_∞ distance.

We answered our research questions thusly:

- **R1:** The results did not drastically change by adding matching on the household / individual level.
- **R2:** The different distance functions lead to really similar results.
- **R3:** No set of matching features lead to drastically different results.

[1] A reference to a table from the original paper will always be in italics

[2] total household income per capita, total value of household assets per capita, proportion of household members who work, household operates a microenterprise, hours worked by household members per capita, total household consumption per capita, log of self-reported rental value of house, log of self-reported sale value of house

[3] construction/expansion of sanitation facilities, restoration of sanitation facilities, construction of ceiling, restoration of walls, other house expansion

[4] total value of household assets per capita; this initially should have contained the log total income of parents of children 0 - 5 yrs, but unfortunately the dataset was missing all values for these columns

Table 3: Regressions of Children's Health Measures on Program Dummy

	Control		Model 1		Model 2		Model 3	
Dependent variable	Paper values	T5_match	Paper values	T5_match	Paper values	T5_match	Paper values	T5_match
			-0.065	-0.061	-0.064	-0.065	-0.064	-0.064
Parasite count	0.333 (0.673)	0.328 (0.661)	[0.032]**	[0.029]**	[0.031]**	[0.029]**	[0.032]**	[0.030]**
			-19.545	-18.637	-19.345	-19.983	-19.198	-19.578
			-0.018	-0.025	-0.020	-0.026	-0.018	-0.024
Diarrhea	0.142 (0.349)	0.150 (0.357)	[0.009]*	[0.010]**	[0.009]**	[0.010]***	[0.009]*	[0.010]**
			-12.819	-16.829	-13.834	-17.572	-12.803	-16.073
			-0.085	-0.088	-0.081	-0.085	-0.083	-0.086
Anemia	0.426 (0.495)	0.431 (0.495)	[0.028]***	[0.030]***	[0.027]***	[0.029]***	[0.027]***	[0.028]***
			-20.059	-20.436	-18.908	-19.625	-19.388	-19.962
			4.031	4.153	5.652	5.454	5.557	5.279
MacArthur Communicative Development Test score	13.354 (18.952)	13.352 (18.956)	[1.650]**	[1.643]**	[1.642]**	[1.645]***	[1.641]**	[1.646]***
			30.182	31.106	42.235	40.851	41.609	39.534
			2.668	2.525	3.206	3.146	3.083	3.076
Picture Peabody Vocabulary Test percentile score	30.656 (24.864)	30.729 (24.729)	[1.689]*	[1.650]	[1.430]**	[1.472]**	[1.410]**	[1.467]**
			8.702	8.218	10.460	10.238	10.058	10.009
			0.007	-0.008	0.002	-0.011	-0.002	-0.007
Height-for-age z-score	-0.605 (1.104)	-0.595 (1.110)	[0.043]	[0.045]	[0.038]	[0.041]	[0.039]	[0.041]
			-1.161	1.312	0.279	1.773	-0.323	1.228
			0.002	-0.004	-0.005	-0.009	-0.011	-0.015
Weight-for-height z-score	0.125 (1.133)	0.138 (1.134)	[0.034]	[0.036]	[0.036]	[0.039]	[0.037]	[0.039]
			1.790	-2.64	-4.119	-6.774	-8.727	-10.892

Notes: As in the original paper, regressions computed using survey information (sample sizes reported in Table 1a). Missing values in covariates were imputed with zero, and a corresponding dummy variable was then added to the regressions. Model 1: no controls; Model 2: age, demographic, and health-habits controls; Model 3: age, demographic, health-habits, and public social programs controls. T4_eco_match: results when matching on economic. T4_house_match: results when matching on house improvement. T5_match: results when matching on economic. Reported results: estimated coefficient, clustered standard error at census-block level in brackets (136 clusters), and 100 × coefficient/control mean.
***/**/* Significantly different from 0 at the 1/5/10 percent level

6 Conclusion

No additional matching, whatever the distance function used, changed drastically the original paper's results. Since they matched blocks of houses together based on average values of a block's houses' features, and that those features are likely correlated with the individual / household level features, it was not surprising to us that no real difference appeared.

It would have been really interesting to see the result of **replacing** the block-level matching with individual / household-level matching. However this would require using the original datasets, which the authors did not provide.

References

Matias D. Cattaneo, Sebastian Galiani, Paul J. Gertler, Sebastian Martinez, and Rocio Titiunik. 2009. *Housing, Health, and Happiness*, 1(1):75-105..

A Variables description

In both *Table 4* and *Table 5* of the original paper, all models use `idcluster` for clustering, control and treatment groups are identified by `dpisofirme` and model 1 has no control variables.

For *Table 4*, the data related to the dependent variables can be fount in the following columns:

- Share of rooms with cement floors
- Cement floor in kitchen
- Cement floor in dining room
- Cement floor in bathroom
- Cement floor in bedroom

And for *Table 5*:

- Parasite count
- Diarrhea
- Anemia
- McArthur Communication Development Test score
- Picture Peabody Vocabulary Test percentile score
- Height-for-age z-score
- Weight-for-height z-score

Model 2 is split in two groups, *demographic* and *health*. The *health* variables are identical for both *Tables 4* and *5*, but the *demographic* ones differ:

- health:
  - Household has animals on land
  - Animals allowed to enter the house
  - Water connection outside
  - Water connection inside the house
  - Electricity
  - Number of times respondent washed hands the day before
  - Uses garbage collection service
- demographic (*Table 4*):
  - Number of household members
  - (Number of rooms; this one is mentioned in the paper, but not used for the regression so we dropped it)
  - Head of household's years of schooling
  - Spouse's years of schooling
  - Head of household's age
  - Spouse's age
  - Proportion of Males 0-5yrs in household
  - Proportion of Males 6-17yrs in household
  - Proportion of Males 18-49yrs in household
  - Proportion of Males 50+yrs in household
  - Proportion of Females 0-5yrs in household
  - Proportion of Females 6-17yrs in household
  - Proportion of Females 18-49yrs in household
  - Proportion of Females 50+yrs in household
- demographic (*Table 5*):
  - Number of household members
  - Number of rooms
  - Age
  - Male
  - Mother of at least one child in household present
  - Mother's age (if present)
  - Mother's years of schooling (if present)
  - Father of at least one child in household present
  - Father's age (if present)
  - Father's years of schooling (if present)
  - (Trimester * Gender) Dummy for children 0-5yrs

Model 3 adds 4 economic control variables, the same for both Tables *4* and *5*:

- Transfers per capita from government programs
- Household beneficiary of government milk supplement program
- Household beneficiary of government food program
- Household beneficiary of seguro popular

B Detailed regression results

Table 4: Regressions of Cement Floor Coverage Measures

		Control group					
Dependant variable	Paper values	T4_eco_match (L_1)	T4_eco_match (L_2)	T4_eco_match (L_∞)	T4_house_match (L_1)	T4_house_match (L_2)	T4_house_match (L_∞)
Share of rooms with cement floors	0.728 (0.363)	0.748 (0.345)	0.763 (0.333)	0.715 (0.383)	0.719 (0.368)	0.719 (0.368)	0.719 (0.368)
Cement floor in kitchen	0.671 (0.470)	0.690 (0.463)	0.715 (0.452)	0.673 (0.470)	0.662 (0.473)	0.662 (0.473)	0.662 (0.473)
Cement floor in dining room	0.709 (0.455)	0.720 (0.449)	0.740 (0.439)	0.700 (0.459)	0.697 (0.460)	0.697 (0.460)	0.697 (0.460)
Cement floor in bathroom	0.803 (0.398)	0.821 (0.384)	0.816 (0.388)	0.760 (0.428)	0.796 (0.403)	0.796 (0.403)	0.796 (0.403)
Cement floor in bedroom	0.668 (0.471)	0.717 (0.451)	0.732 (0.444)	0.692 (0.463)	0.656 (0.475)	0.656 (0.475)	0.657 (0.475)
		Model 1					
Dependant variable	Paper values	T4_eco_match (L_1)	T4_eco_match (L_2)	T4_eco_match (L_∞)	T4_house_match (L_1)	T4_house_match (L_2)	T4_house_match (L_∞)
	0.202	0.173	0.16	0.207	0.209	0.209	0.209
Share of rooms with cement floors	[0.021]***	[0.026]***	[0.024]***	[0.03]***	[0.021]***	[0.021]***	[0.021]***
	27.746	23.119	20.951	28.968	29.132	29.132	29.1
	0.255	0.228	0.209	0.24	0.263	0.263	0.263
Cement floor in kitchen	[0.025]***	[0.032]***	[0.026]***	[0.033]***	[0.024]***	[0.024]***	[0.024]***
	37.936	33.038	29.249	35.593	39.716	39.716	39.669
	0.21	0.191	0.169	0.224	0.218	0.218	0.217
Cement floor in dining room	[0.026]***	[0.034]***	[0.035]***	[0.035]***	[0.025]***	[0.025]***	[0.025]***
	29.633	26.539	22.901	32.065	31.201	31.201	31.166
	0.105	0.067	0.065	0.129	0.11	0.11	0.11
Cement floor in bathroom	[0.022]***	[0.029]**	[0.029]**	[0.04]***	[0.022]***	[0.022]***	[0.022]***
	13.071	8.194	7.958	17.0	13.864	13.864	13.851
	0.238	0.196	0.195	0.228	0.25	0.25	0.25
Cement floor in bedroom	[0.020]***	[0.025]***	[0.03]***	[0.034]***	[0.02]***	[0.02]***	[0.02]***
	35.598	27.292	26.641	32.967	38.141	38.141	38.095
		Model 2					
Dependant variable	Paper values	T4_eco_match (L_1)	T4_eco_match (L_2)	T4_eco_match (L_∞)	T4_house_match (L_1)	T4_house_match (L_2)	T4_house_match (L_∞)
	0.208	0.178	0.163	0.222	0.215	0.215	0.215
Share of rooms with cement floors	[0.019]***	[0.024]***	[0.023]***	[0.032]***	[0.02]***	[0.02]***	[0.02]***
	28.512	23.837	21.398	31.082	29.902	29.902	29.872
	0.260	0.234	0.211	0.249	0.268	0.268	0.268
Cement floor in kitchen	[0.023]***	[0.03]***	[0.026]***	[0.035]***	[0.023]***	[0.023]***	[0.023]***
	38.708	33.869	29.521	37.072	40.445	40.445	40.402
	0.217	0.194	0.168	0.236	0.224	0.224	0.224
Cement floor in dining room	[0.025]***	[0.034]***	[0.037]***	[0.038]***	[0.025]***	[0.025]***	[0.025]***
	30.588	26.936	22.679	33.765	32.191	32.191	32.162
	0.113	0.075	0.072	0.156	0.118	0.118	0.118
Cement floor in bathroom	[0.018]***	[0.025]***	[0.027]***	[0.04]***	[0.019]***	[0.019]***	[0.018]***
	14.043	9.143	8.851	20.502	14.804	14.804	14.788
	0.245	0.204	0.204	0.234	0.258	0.258	0.258
Cement floor in bedroom	[0.021]***	[0.023]***	[0.029]***	[0.039]***	[0.021]***	[0.021]***	[0.021]***
	36.735	28.4	27.899	33.87	39.303	39.303	39.256
		Model 3					
Dependant variable	Paper values	T4_eco_match (L_1)	T4_eco_match (L_2)	T4_eco_match (L_∞)	T4_house_match (L_1)	T4_house_match (L_2)	T4_house_match (L_∞)
	0.210	0.181	0.164	0.227	0.218	0.218	0.218
Share of rooms with cement floors	[0.019]***	[0.024]***	[0.023]***	[0.033]***	[0.02]***	[0.02]***	[0.02]***
	28.876	24.221	21.547	31.753	30.302	30.302	30.273
	0.265	0.237	0.213	0.258	0.272	0.272	0.272
Cement floor in kitchen	[0.023]***	[0.029]***	[0.026]***	[0.037]***	[0.023]***	[0.023]***	[0.023]***
	39.440	34.346	29.757	38.321	41.132	41.132	41.089
	0.221	0.197	0.168	0.243	0.229	0.229	0.229
Cement floor in dining room	[0.025]***	[0.033]***	[0.037]***	[0.039]***	[0.025]***	[0.025]***	[0.025]***
	31.189	27.367	22.636	34.763	32.84	32.84	32.812
	0.117	0.081	0.082	0.16	0.122	0.122	0.122
Cement floor in bathroom	[0.018]***	[0.024]***	[0.027]***	[0.041]***	[0.019]***	[0.019]***	[0.018]***
	14.536	9.888	10.063	21.004	15.381	15.381	15.366
	0.245	0.204	0.201	0.233	0.258	0.258	0.258
Cement floor in bedroom	[0.020]***	[0.022]***	[0.029]***	[0.04]***	[0.02]***	[0.02]***	[0.021]***
	36.695	28.489	27.481	33.706	39.307	39.307	39.261

Notes: Regressions computed using survey information (sample sizes reported in Table 1a). Missing values in covariates were imputed with zero, and a corresponding dummy variable was then added to the regressions. Model 1: no controls; Model 2: age, demographic, and health-habits controls; Model 3: age, demographic, health-habits, and public social programs controls. T4_eco_match: results when matching on economic. T4_house_match: results when matching on house improvement. T5_match: results when matching on economic. Reported results: estimated coefficient, clustered standard error at census-block level in brackets (136 clusters), and 100 × coefficient/control mean.
***/**/* Significantly different from 0 at the 1/5/10 percent level

C Number of datapoints before and after matching

Table 5: Number of datapoints

		Before matching	L_1	L_2	L_∞
Households dataset	T4_eco_match	2755	1308	708	526
	T4_house_match		2556	2556	2558
Individuals dataset	T5_match	4052	-	-	1984