Informing Policymakers on State Level Supplemental Security Income

2022 CUSP Capstone Project

Team:

Samuel Zierler
Jingxuan Xiao
Chih-Yun Lu

Sponsors:

Dani Hochfellner
Mary Hamman

1. Abstract:

Supplemental security income (SSI) benefits people with limited income and resources who are disabled, blind, or 65 or older. Blind or disabled children may also get SSI. And SSI is under federal and state governments, meaning eligible people can receive monthly payments from them.
The federal government basically provides the same amount to most states, even though the cost of living in those states is different. In addition, the federal government does not manage the amount of state the governments provide to recipients. Moreover, the couple penalty in SSI is also an issue that should be modified. Specifically, an individual in a couple gets a lower amount than an individual with no partner. Furthermore, with the technology being well advanced, more people can survive at an older age, and that is also an issue that should be solved because both the federal and state governments have a limited amount to support SSI. Therefore, the motivation of this project is to address the amount of the federal and state government, the unfairness of the couple penalty, and the aging issue.

2. Problem Definition:

We are facing several problems, and the biggest problem is that the previous data of state assistance programs for SSI recipients is a PDF form. In other words, we cannot access the data efficiently. We need to find out what techniques can help us get the data quickly. Moreover, not just the challenge of accessing the data, we also face issues analyzing those data. We expected we could explore methods to access those data to study them and create a dashboard to provide people to understand this program because so far, there is no dashboard to review.

3. Literature Review:

To deeply understand the SSI program, we focused on reports that introduced the couple penalty, living arrangements, as well as the amount the federal and state government provide. We found that the federal government supplies the same amount to most states. And the amount from state governments is different, which might be because of the cost of living in states. Moreover, we also study the aging issue since the recipients' population has changed every year. To secure the budget for the SSI program, this is also a topic that we focused on.

4. Data:

The primary data for this project was a collection of historic reports published by the Social Security Administration and provided to our team in the form of scanned PDF images. These reports contained annual information about the Federal mandatory recipients as well as State-level Optional Supplementation program implementations. Each state defines its own income and resource limitations, if any beyond the Federal thresholds, as well as describing what living arrangements are covered under SSI for that state and the benefit amounts for both individual and couple. These reports were provided for a nearly twenty year time span from 1990 to 2011, with a few years missing in that range. Our group chose to focus on extracting and rebuilding the payment tables for each state so that we could examine this time series in full and run a predictive model for generating future benefits outside of our provided range. This decision was made because the payment table data was the most significant data for addressing our project questions of the “Couple Penalty” and state-to-state comparisons. We chose to include several additional comparative values in our

modeling processes and visualization component to support the primary benefit data and provide more context to the adequacy of the numbers. As a means of simulating hypothetical past adjustments to the state level benefit amounts the modeling process employed a consumer price index value as a counterfactual to the provided values from the report. In the visualization tool itself we included the average median household income in each state and urban region as a comparative value to weigh the adequacy of benefit amounts against the expected cost of living in a given state.

5. Data Extraction:

The process for extracting the payment table data from the PDF reports was not a simple task at first due to the format of the documents. The underlying data was never created with the intention of being machinable in the way this project necessitated. Using pdfminer, a Python library for working with pdf files, our group designed an extraction pipeline to isolate the payment table data from each file and extract it, ultimately to a Pandas data frame structure. The extraction pipeline worked as follows:

1. The pdf (as a series of scanned images) is read into the notebook and converted to a raw text form. (pdfminer.high_level, extract_text)

From this raw text block, each "Payment Levels" page could be identified by a consistent table heading which provided a convenient capture for a regular expression (Figure 01). Regex was used throughout the extraction pipeline to isolate and select specific phrases or sections of data as well as for correcting mistakes in character recognition from the OCR process. Through some exploration of the raw text blocks we found several consistent mistypings that were adjusted or replaced with Regex.

2. The raw text block is split into separate payment pages for each state and extra characters are dropped using Regex.

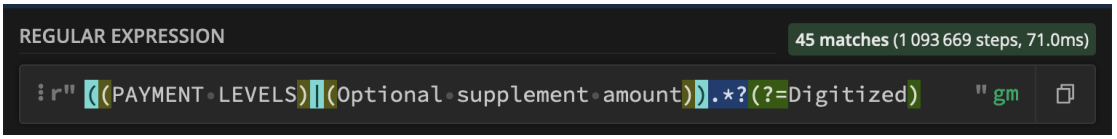


Figure 01: Regular expression for isolating payment table pages from raw pdf text. The Regex uses a "Digitized by Google" stamp as its lookahead cutoff. The 45 matches represent the number of states with an optional program that year.

Many of the PDF pages that contained the desired Payment Tables had additional, irrelevant information for our focus. Again a consistent heading following the table could be used as a cutoff anchor for the regular expression capture (Figure 02). Many tables also contained footnotes with possibly significant information to the state benefits. This data was captured alongside the table data to potentially be used later in the visualization.

STATE ASSISTANCE FOR SPECIAL NEEDS

Figure 02: "State Assistance For Special Needs" heading consistently followed the end of the Payment Tables and operated as the lookahead cutoff.

3. Each page text is split into two list structures, *table_data* and *footnotes*.

With the table data isolated the next step was to separate the text-based living arrangements from the numeric benefit amounts. There was significant complexity in outlining an expression to adequately extract the living arrangement data simply because of the wide variety of possible categories and the potential for slight difference in wording between states. Furthermore again the issues of the underlying structure of this data and the lack of intention for machining in its design became relevant. While the documents appeared to have a meaningful table structure to them visually, the way in which the text was converted and read in meant that it was not reliable that their order in the text block would match the appearance of that of the table.

4. Each *table_data* structure is further split into two list structures, *living_arrangements* and *benefits*.

Many states subdivide their living arrangements along another category. This may commonly be a separation between aged, blind, and disabled recipients or for those living in different defined geographical areas within that state. These subcategories do not change significantly year to year within a state, as was found through exploratory analysis of the reports. For that fact, a data structure was designed containing these subcategories and joined with the living_arrangements data structure.

5. Each *living_arrangements* list is merged with *adjustments* to make *adjusted_living_arrangements* (final form).

With the living arrangements complete, the process focuses next on rebuilding the benefits data to match the tables. Each table contains four benefit columns, Combined (Federal + State) Individual, Combined Couple, State Individual, and State Couple. These may have an amount (if covered) or may have a zero, dash, or empty spot if not. Again the underlying structural issues were the most significant with converting the benefit areas. The benefit amounts were read into a list with a single dimension but the table structure is a list of lists, rows with 4 columns. It would seem simple enough to split a single list into groups however because of the unreliability of the ordering of the benefits list it was not possible to trust that this operation would provide the correct arrangement as an output. What was discovered was that the inconsistencies in the benefit order was itself somewhat consistent, meaning a given state's benefits list was most likely in the same incorrect order year and year when extracted from the reports. To address the issue of incorrect ordering we identified the expected order for each state benefits list and stored that as a pattern in a list structure. A small compiler was designed to read in the pattern and reorder the benefits list to match. Based on the pattern structure the compiler would assemble a list of swap or shuffle instructions for the specific benefit list. For more information on these operations please refer to the "Example Parser" notebook file.

6. Each benefits list is reordered and pivoted to be a list of lists where outside length is the number of living arrangements, inside length of each list is 4.
7. Each *sorted_benefits* list is appended to a single structure of *state_payments* (final form).

At this point the two structures, *adjusted_living_arrangements* and *state_payments*

should be parallel to each other. A check is performed to ensure that they are. If not the data must be checked by hand at the state(s) with issue. Most likely the benefits shuffle was not correct and needs to be manually adjusted for that year. If they match, the final data frame can be generated for that year. Each year's output file was a CSV structure of the joined payment tables for all states for the report year.

6. Modeling:

In order to study the specific distribution of the different situations in each state for funding SSI program, our group needed a large amount of data to analyze by building the model; notwithstanding, we had discussed different models such as Decision Tree, Random Forest, Support Vector Machine, CNN, RNN and Time Series Models, etc. Finally, we made decision to use ARIMA Model of Time Series Model to analyze, due to the fact that compared to other models, ARIMA Model has some significant superiorities are that by calculating the difference between values in the time series instead of the actual values to predict the future 'value' and trends, which are via 'signature' with significant pattern of autocorrelations and autoregressive signals. Besides that, the ARIMA Model is not only capable of visualizing the data we have extracted from the historical literature for each state, but it also can make non-stationary data be transformed into stationary data. First of all, our step is to filter the specified type of data in accordance with our goal, which is to help policymakers to fund people who are in the SSI program at different state levels, so we ought to identify some 'similar' patterns among states at the beginning. Secondly, the next step is to analyze the residuals in the time-series data, because we need to detect whether the past data in the time-series is consecutive and to correct the inference of future values. Thirdly, after completing the linear regression for the analysis, we should find the relevant parameters which are MA(q), d and AR(p) before building the ARIMA Model; thus, by applying the ACF and PACF methods we easily gain them. Fourthly, we use the 75% dataset as the training set and the chosen parameters p, d and q to build the model and train it, and we also obtained the calculated values in the two-dimensional axis. More than that, our group also used the additional statistics method 'Ljung Box' to test the residuals at each data point to ensure that the autocorrelation and kernel density estimations are consistent and sustainable. Finally, we can plot the ARIMA Model line based on the previous modeling steps, and our group focused on predicting the future trends of SSI program funding, so we also plotted the five out-of-sample trends from 2011 to 2016 to help policymakers to make a more comprehensive and informed decision. Figures below present several elements of this modeling process.

```
In [8]:
#Anlyaze the residuals by using the linear model,and it helps us to find the ACF and PACF, which both of them belonged to ARIMA terms
lm=sm.OLS(df_AK.state_indv_sum[:16],sm.add_constant(df_AK.year.index)).fit()
print (lm.summary())
```

OLS Regression Results

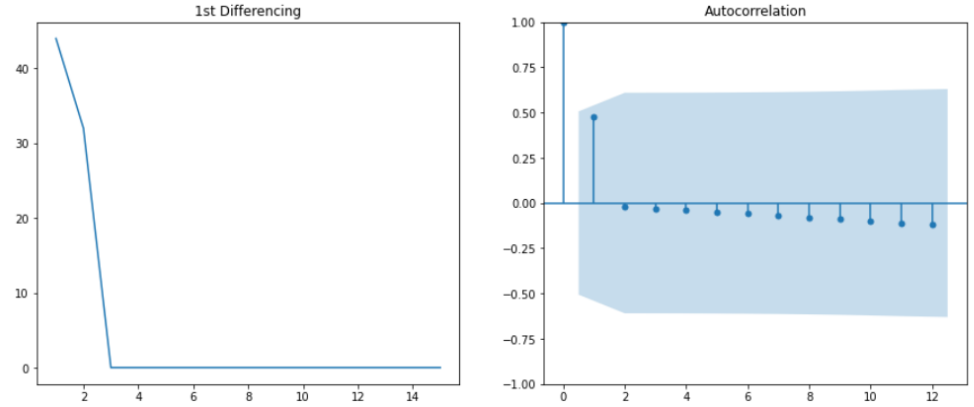
Dep. Variable:	state_indv_sum	R-squared:	0.293
Model:	OLS	Adj. R-squared:	0.243
Method:	Least Squares	F-statistic:	5.809
Date:	Mon, 25 Jul 2022	Prob (F-statistic):	0.0303
Time:	16:22:50	Log-Likelihood:	-67.436
No. Observations:	16	AIC:	138.9
Df Residuals:	14	BIC:	140.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	859.0882	8.358	102.784	0.000	841.162	877.015
x1	2.2882	0.949	2.410	0.030	0.252	4.325

Omnibus:	17.990	Durbin-Watson:	0.627
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17.568
Skew:	-1.783	Prob(JB):	0.000153
Kurtosis:	6.694	Cond. No.	17.0

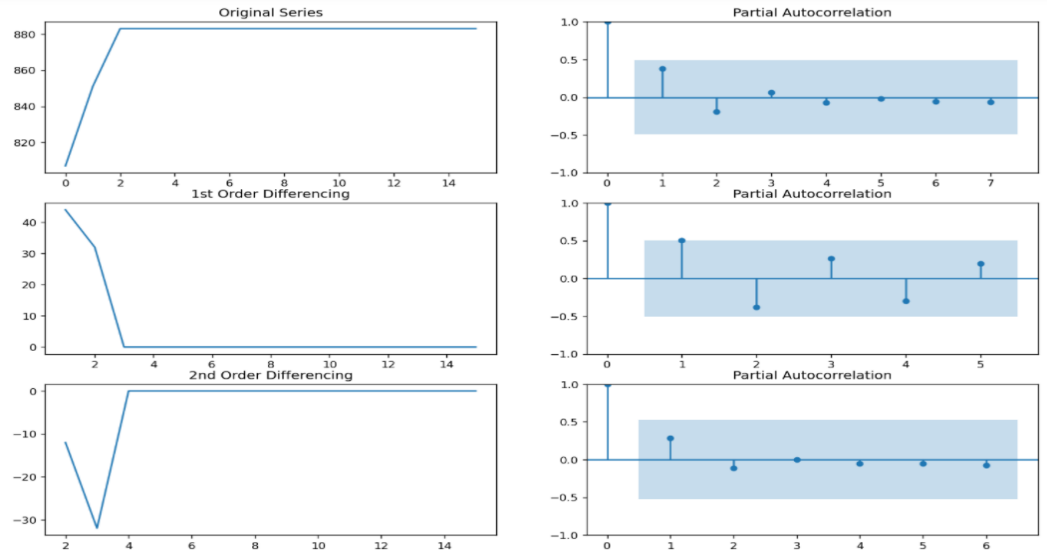
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 03: Results of primary Linear Regression model



Obviously, we should choose MA(q)=1 from the above graph.

Figure 04: 1st Order Differencing and ACF plot from above regression. ACF plot would suggest choosing an MA (q) value of 1 for ARIMA.



From the above results, we might choose d=0 or 1, but d=1 will be appropriate in this situation.

Figure 05: Series, 1st, and 2nd Order Differencing and PACF plots from above regression. PACF would

suggest choosing an AR (d) value of 1 for ARIMA.

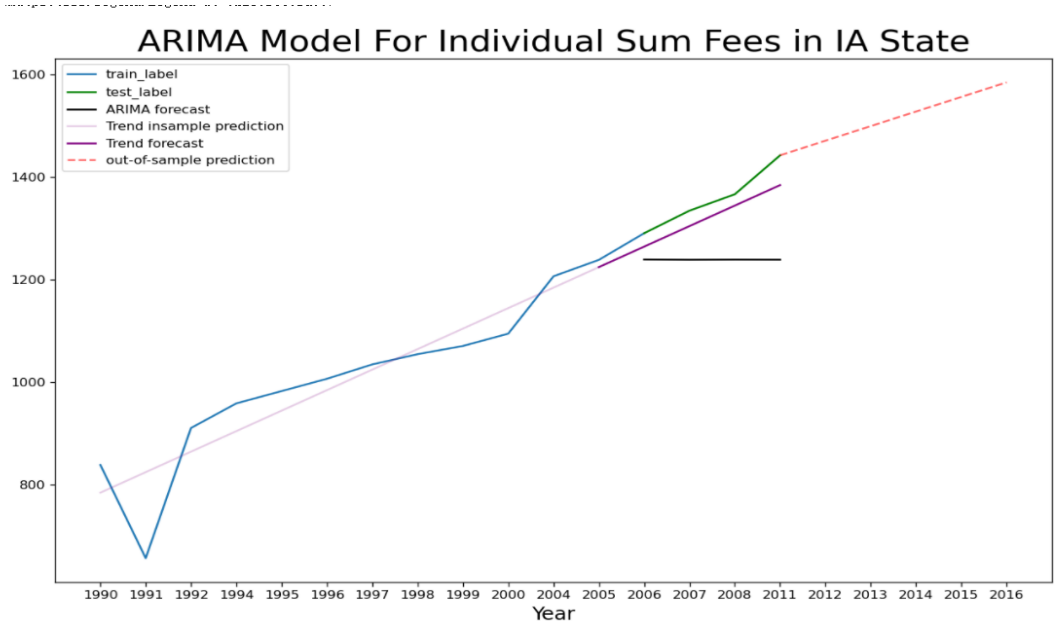


Figure 06: ARIMA model output for Iowa

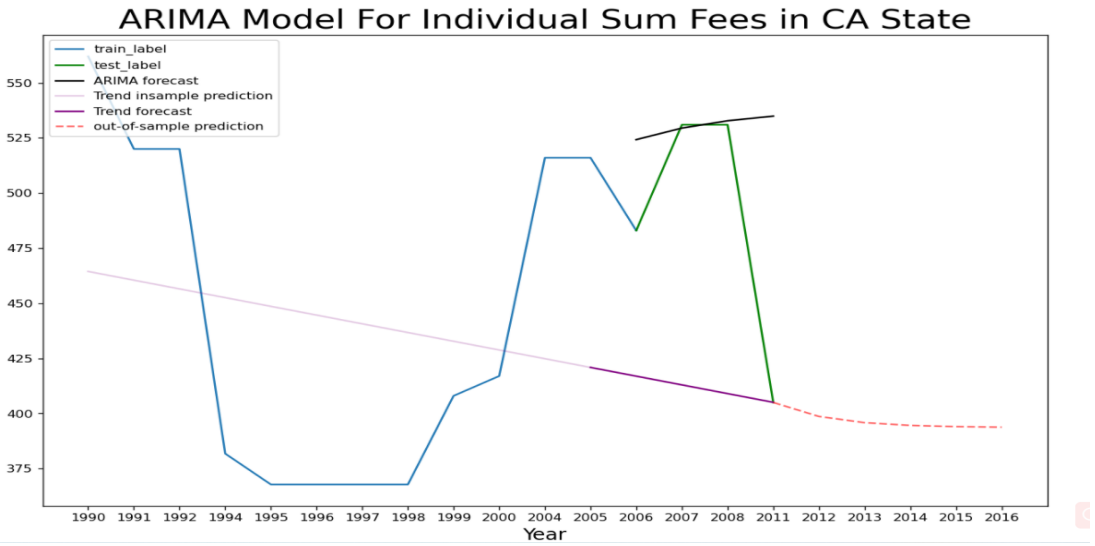


Figure 07: ARIMA model output for California

7. Data Visualization

One of our ultimate goals with this project was to create a space where other people could continue exploring and interacting with the data we extracted. To achieve this we built a Tableau dashboard that is hosted publicly through their services. When a user first arrives on the dashboard they will arrive on an informational page that describes the SSI program and this Capstone project. From there they can navigate to several dedicated pages.

- The “Program Coverage” page presents a map of states that participate in the SSI program and can be filtered by year or by specific living arrangements.
- The “State Overview” page shows all the living arrangements defined by a specific state and can be viewed in multiple arrangements, selecting for

Individual or Couple benefits.

- The “Compare States” page is built to show how different states provide benefits for the same living arrangement types. One chart shows all the states that match so comparisons can be made within the same visual space.
- The “Couple Penalty” page shows the Individual vs. Couple benefits for all the living arrangements in a selected state. Each chart also includes a hypothetical gray line that is the Individual benefit x2. It is expected that with no couple penalty this line should match the Couple amounts.

More information on how to use these pages can be found on the dashboard itself.

8. Data Analysis and Results

With the final output file we created several calculated fields for the Federal Individual and Couple benefits as well as the State level Individual to Couple difference (Figure 08). From the Federal amounts we were able to map each State level living arrangement to an expected Federal benefit category by matching the values. We found that states were generally consistent in their definitions of living arrangements and that the benefits appropriately mapped to the Federal categories. There is some complexity with “Essential Person” which was an antiquated classification from older State level programs that were replaced with SSI. These differences can be accounted for by including the EP adjustment that is defined in the Federal payment table for each year. The calculated field for *state_diff_dbl* is a boolean for if the State level Couple benefit is twice that of the Individual. This was relevant to our exploration of the “Couple Penalty” issue in the benefit level setting of these programs.

	A	B	C	D	E	F	G	H	I	J	K	L
1	year	stat	liv_arr	sub_cat	combn_i	combn_c	state_ind	state_cpl	fed_indv	fed_cpl	state_diff	state_diff
2	1990	AK	Living independently		717	1063	331	484	386	579	153	FALSE
3	1990	AK	Living independently with ineligible spouse		862	0	476	0	386	0	-476	FALSE
4	1990	AK	Living in household of another		594.34	883	337	497	257.34	386	160	FALSE
5	1990	AK	Living in household of another with ineligible spouse		681.34	0	424	0	257.34	0	-424	FALSE
6	1990	AK	Medicaid facility		75	150	45	90	30	60	45	TRUE
7	1990	AL	Receiving IHC in a private home or a personal care home	Level A	446	699	60	120	386	579	60	TRUE
8	1990	AL	Receiving IHC in a private home or a personal care home	Level B	442	691	56	112	386	579	56	TRUE
9	1990	AL	Receiving IHC and support and maintenance in a private home or personal care home	Level A	317.34	506	60	120	257.34	386	60	TRUE
10	1990	AL	Receiving IHC and support and maintenance in a private home or personal care home	Level B	313.34	498	56	112	257.34	386	56	TRUE
11	1990	AL	Receiving specialized IHC in a private home or personal care home		446	699	60	120	386	579	60	TRUE
12	1990	AL	Receiving specialized IHC and support and maintenance in a private home or personal care home		317.34	506	60	120	257.34	386	60	TRUE
13	1990	AL	Living in foster home with IHC or specialized IHC		496	799	110	220	386	579	110	TRUE
14	1990	AL	Living in cerebral palsy treatment center		582	971	196	392	386	579	196	TRUE
15	1990	AZ	Requires housekeeping services		456	0	70	0	386	0	-70	FALSE
16	1990	AZ	Licensed private nursing homes		466	739	80	160	386	579	80	TRUE
17	1990	AZ	Licensed county operated nursing homes		174	348	174	348	0	0	174	TRUE
18	1990	AZ	Licensed supervisory care homes		436	679	50	100	386	579	50	TRUE
19	1990	AZ	Adult foster care homes		436	679	50	100	386	579	50	TRUE
20	1990	AZ	24-hour treatment facilities		436	679	50	100	386	579	50	TRUE
21	1990	CA	Independent living with cooking facilities	Aged and disabled	630	1167	244	588	386	923	344	FALSE
22	1990	CA	Independent living with cooking facilities	Blind	704	1372	318	793	386	1054	475	FALSE
23	1990	CA	Nonmedical out-of-home care		709	1418	323	839	386	1095	516	FALSE
24	1990	CA	Independent living without cooking facilities	Aged and disabled	698	1303	312	724	386	991	412	FALSE
25	1990	CA	Living in household of another	Aged and disabled	501.34	974	244	588	257.34	730	344	FALSE
26	1990	CA	Living in household of another	Blind	575.34	1179	318	793	257.34	861	475	FALSE
27	1990	CA	Disabled minor in home of parent/guardian/relative		499	0	113	0	386	-113	-113	FALSE
28	1990	CA	Nonmedical out-of-home care living in household of another		580.34	1225	323	839	257.34	902	516	FALSE
29	1990	CA	Disabled minor in the household of another		370.34	0	113	0	257.34	-113	-113	FALSE
30	1990	CA	Medicaid facility		42	84	12	24	30	72	12	TRUE
31	1990	CO	Living independently or in the home of another	Aged	444	888	54	309	390	579	255	FALSE
32	1990	CO	Living independently	Blind/Disabled	390	770	4	191	386	579	187	FALSE
33	1990	CO	Adult foster care	Blind/Disabled	581	0	195	0	386	0	-195	FALSE
34	1990	CO	Home care	Aged	763	0	377	0	386	0	-377	FALSE
35	1990	CO	Home care	Blind/Disabled	709	0	323	0	386	0	-323	FALSE
36	1990	CO	Individual with essential spouse	Blind/Disabled	514	0	128	0	386	0	-128	FALSE
37	1990	CT	Independent community living		752	1104	366	525	386	579	159	FALSE
38	1990	DC	Adult foster care home	(50 beds or less)	533.2	1066.4	147.2	487.4	386	579	340.2	FALSE
39	1990	DC	Adult foster care home	(over 50 beds)	643.2	1286.4	257.2	707.4	386	579	450.2	FALSE
40	1990	DC	Living independently		401	609	15	30	386	579	15	TRUE
41	1990	DC	Living in household of another		272.34	416	15	30	257.34	386	15	TRUE
42	1990	DC	Living independently with an essential person		594	802	15	30	579	772	15	TRUE
43	1990	DC	Living in household of another with an essential person		401	544.67	15	30	386	514.67	15	TRUE
44	1990	DC	Medicaid facility		60	120	30	60	30	60	30	TRUE
45	1990	DE	Living in adult residential care facility		526	1027	140	448	386	579	308	FALSE
46	1990	FL	Community care program	Adult foster care	583	0	197	0	386	0	-197	FALSE
47	1990	FL	Community care program	Adult congregate living f	583	0	197	0	386	0	-197	FALSE

Figure 08: Final output data of all SSI benefits for all years

With the data in a visual space it became much easier to quickly recognize interesting elements to the SSI programs. Figure 09 shows the benefits for “Living Independently” as covered by all states over the time period. It can be seen that most benefits are clustered below $50 with a handful between 50 and 150, however a few lie well above that, notably Alaska. It is possible Alaska’s unique conditions explain

their decision to provide significantly higher benefits to recipients on their own than in almost every other state; more detailed research into Alaska's history of SSI would be needed to be sure.

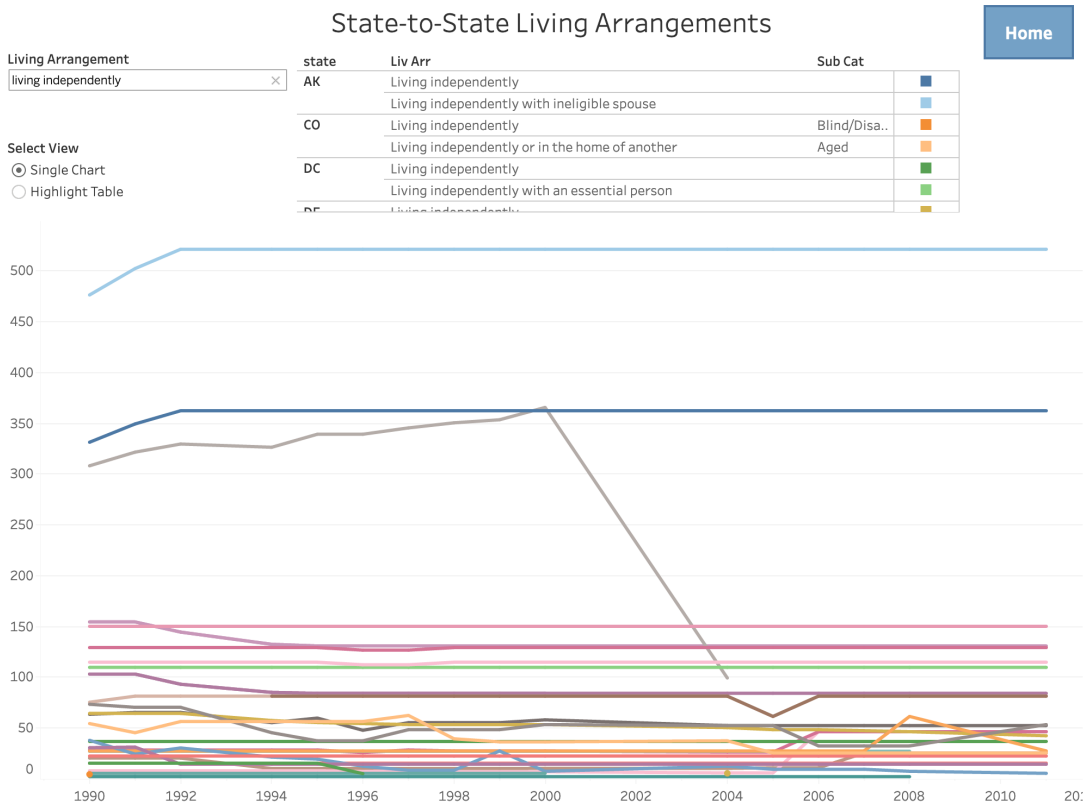


Figure 09: "Living independently" benefits for all states with SSI

Through analysis we found out that there are 133 living arrangements in different states, and most of the states cover these three arrangements: Living in the household of another, Living independently, and Medicaid facility. Moreover, these three living arrangements also are in the federal government category. To easily understand the trends in the past, we applied Tableau as our data visualization tool. Medicaid facilities for individuals and couples in most states didn't change a lot in the past decades (Figure 10 and 11).

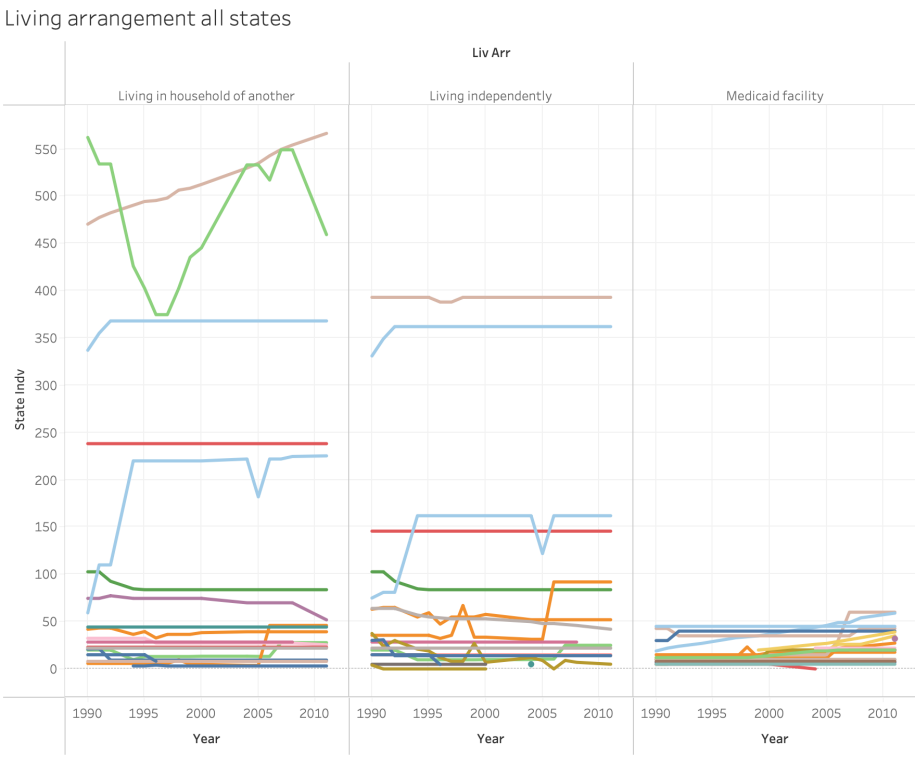


Figure 10: Individual benefits for all states based on Federal living arrangements

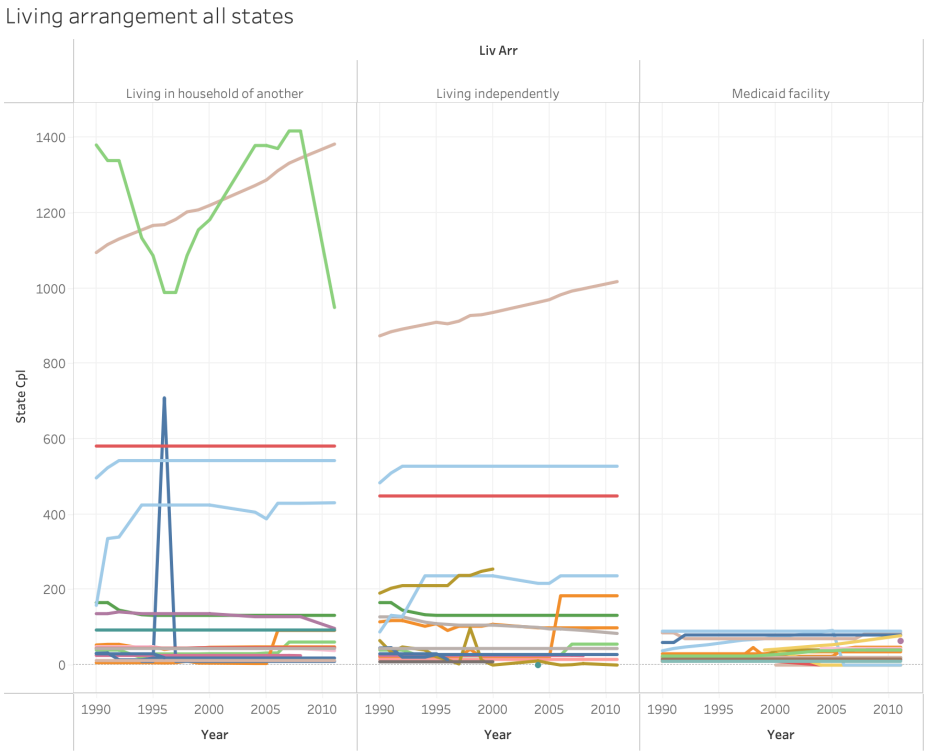


Figure 11: Couple benefits for all states based on Federal living arrangements

To deeply understand the couple penalty, we decided to look into some states that have the complete information and also cover living arrangements, which are living in the household of another, living independently, and Medicaid facilities. So we picked up Washington, D.C. and Washington. The interesting factor we got is that in Washington, D.C, the amount from the state for the individual in a couple is higher

than that for an individual (Figure 12). We believe the reasons can be multiple, such as the cost of living or the aging issue.

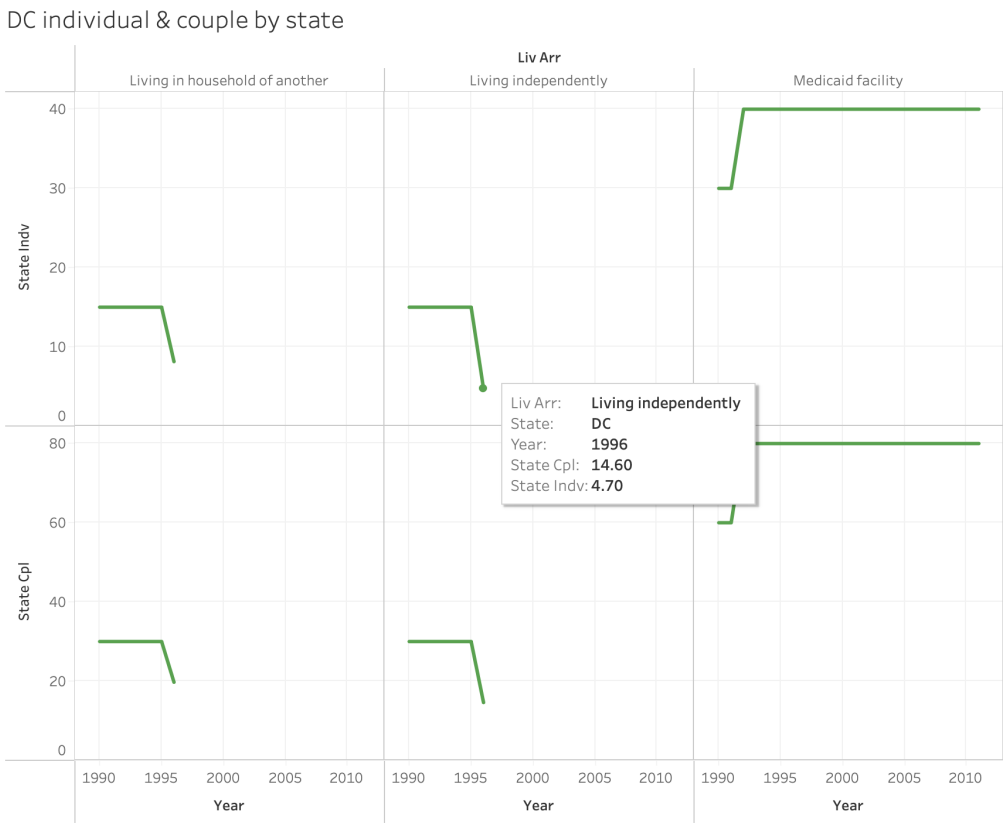


Figure 12: Individual and Couple benefits for recipients in Washington, D.C.

9. Conclusions

We approached this project by initially identifying a series of issues regarding the SSI program. The records were not converted to a format that was appropriate for today's capabilities and therefore the data was unavailable for proper study. This meant it was impossible for both policymakers and outside researchers to truly understand the adequacy of the SSI benefits historically and to apply this knowledge to considering adjustments in the present and future. Through our processes we succeeded in engineering this historical data and extracting it to a machinable format. Furthermore this process could be applied to similar reports for the missing years if these were to be produced, either by the Social Security Administration or another party.

The dashboard tool allows for a richer opportunity of exploration of this data and facilitates the ability to compare states with similar living arrangements that was previously very tedious by sifting through the reports. We hope that this will allow policymakers on the state level to examine their own state's SSI implementation and compare the benefits to other states to better evaluate the adequacy of their own program.

Through our process of researching and learning about the SSI program we discovered other issues, outside the scope of our project, that are relevant to the policies of the program. An example of this would be the policy for recipients to update SSA of any overpayments of benefits and return the mistaken funds within a

short period of time. Ultimately we felt there was too much of a burden on the recipients stemming from the complexity and bureaucracy connected with SSI.

We feel there is significant public good in putting attention to the SSI program. Although the total number of recipients nationally may be relatively small the people who qualify for this program are among the most vulnerable and reliant on their assistance programs. We hope this project serves as a resource to others who feel a similar responsibility to these individuals.

10. Team Collaboration Elements:

Team Member	Areas of Contribution
Sam Zierler	Data extraction Visualization development Co-authored project documents
Chih-Yun Lu	Background research Literature review Visualization case study Co-authored project documents
Jingxuan Xiao	Data modeling Co-authored project documents

References

Mary C. Daly, Mark Duggan.2019. When One Size Does Not Fit All: Modernizing the Supplemental Security Income Program.

Berkowitz, Edward D., DeWitt, Larry. 2013. The other welfare: Supplemental Security Income and U.S. social policy. Ithaca, NY: Cornell University Press.

Duggan, Mark, Kearney, Melissa, Rennane, Stephanie. 2016. The Supplemental Security Income program. In Economics of means-tested transfer programs in the United States, vol. 2, ed. Moffitt, Robert A. , 1–58. Chicago, IL: University of Chicago Press.

Mark Duggan, Melissa S. Kearney & Stephanie Rennane. 2016. The Supplemental Security Income Program.

Mary K. Hamman. 2020. The Demographics Behind Aging in Place: Implications for Supplemental Security Income Eligibility and Receipt.

Jack Smalligan, Chantel Boyens. 2019. Improving the Supplemental Security Income Program for Adults with Disabilities.