Factor Analysis - Socio-Economic determinants (retail)

Joseph O'Malley

11 June, 2019

Contents

Data Overview

This dataset includes Socio-Economic determinants (SEDs) for 252 cities in the United States. Of these SEDs, 93 are independent variables such as education levels, demographic prevalence, and population density. I then removed several features to avoid perfect multi-collinearity to and ended with a final set of 79 variables. The 14 dependent variables pertain to food availability (i.e. - "Average grocery store sales by household", "Number of resturants available per household"). The first part of my analysis focuses on using factor analysis, which is a method that explores a reduced correlation matrix to find variables with common variance and weighs the variables based on this commonality to create "factors".

```
retail_df <- read.csv(fliepath,sep=",",header = T)

##check dimensions before transformation
dim(retail_df)
```

```
## [1] 252 108
```

Transform Data

```
retail <- retail_df
#We set the row names as the cities
rownames(retail) <- retail[, 1]
retail1 <- retail[, -1]

#head(retail1)

# replace missing observation with variable means
```

```r
retail1[] <- lapply(retail1, function(x) {
  x[is.na(x)] <- mean(x, na.rm = TRUE)
  x
})

## remove dependent variables
retail1_sub <- retail1[ , -which(names(retail1) %in% c("Groc_non_food","Groc_food", "S100_H", "S120_H"
                                                   ,"pq_g", "pq_r", "pqr_nonfood" ,"pqr_food", "Groc_non_
                                                   , "Groc_food1","Sr4451_100", "Sr722_120", "Nr4451_100"
                                                   ,"Nr722_120","Share100_4445_72","Share4451_722"
                                                   ,"L722", "L4451"))]

## remove one column from each subcategory to avoid perfect multicolinearity
retail1_sub <- retail1_sub[ , -which(names(retail1_sub) %in% c("t_other_race","t_male"
                                  , "t_age35_44","t_hhder35_44" ,"t_hhd_2p", "t_owner_hous"
                                  ,"t_work_inresidence","t_trans_priv", "t_traveltime15_29"
                                  ,"t_edu_hs","t_employed","t_hhdincomek60_75"
                                  ,"t_Vehic_1","t_nevermarried"))]

##check dimensions of transformed dataset
dim(retail1_sub)
```

```
## [1] 252  75
```

Pt.1: Factor Analysis

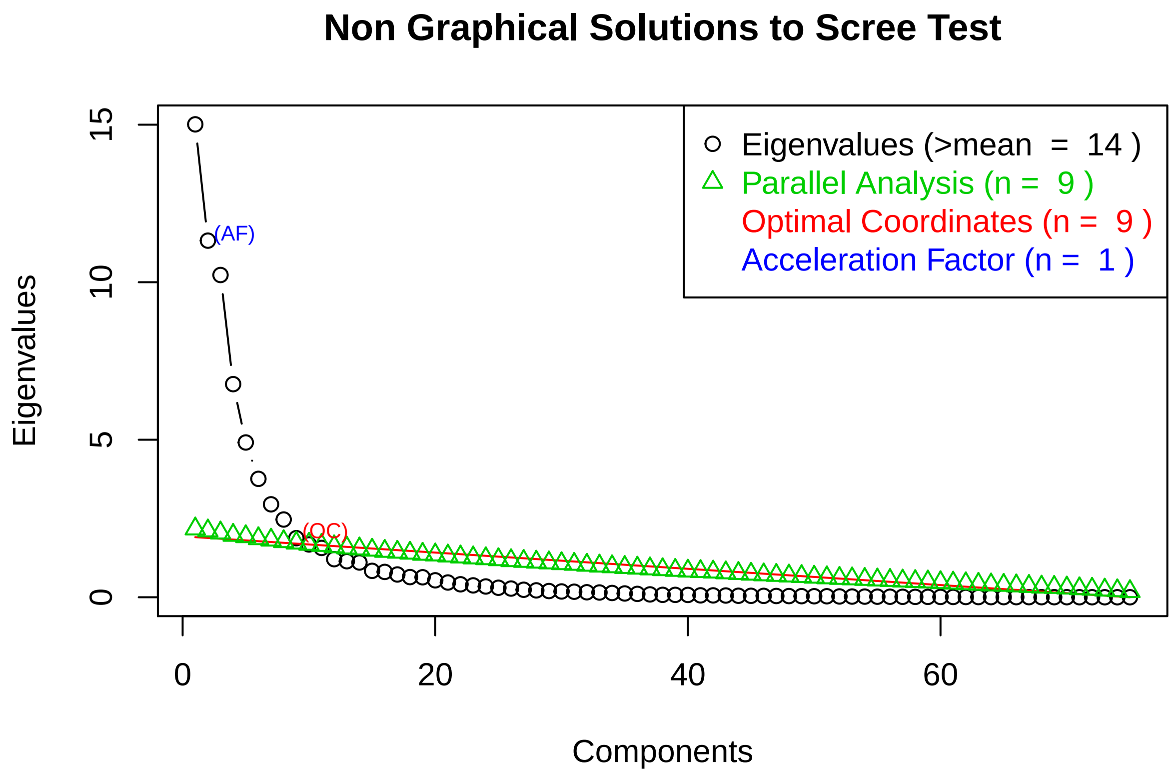
```r
# install.packages("nFactors")
library(nFactors)

#nScree give some indicators about the suggested number of factors, in this case 1 or 14
nScree(retail1_sub)
```

```
##   noc naf nparallel nkaiser
## 1  14   1        14      14
```

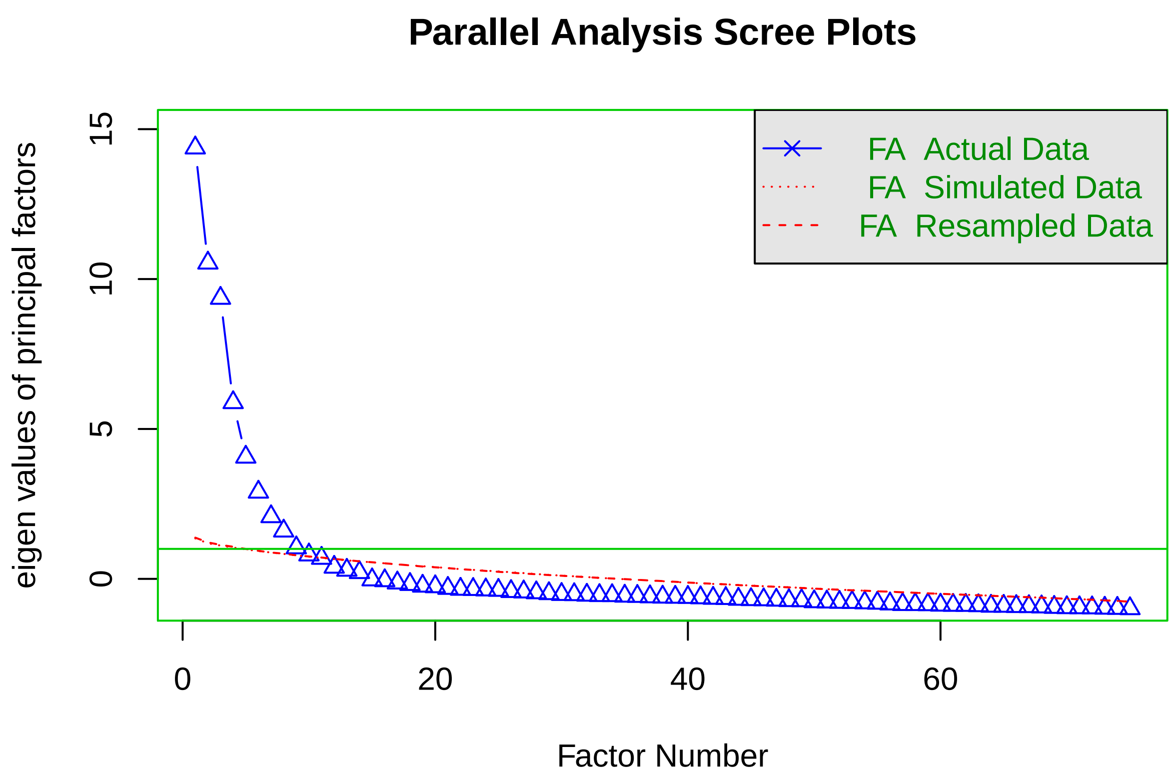
```r
# get eigenvalues. need to be >1 in order to be considered.
eig<- eigen(cor(retail1_sub))
print(eig$values[1:20])
```

```
##  [1] 15.0103182 11.3188806 10.2278603  6.7643970  4.9125342  3.7577786
##  [7]  2.9498920  2.4674841  1.8754002  1.6755659  1.5649237  1.2082521
## [13]  1.1431682  1.1023571  0.8373688  0.8043155  0.7224117  0.6398100
## [19]  0.6341227  0.5371376
```

```r
#graph showing suggested number of factors
ap <- parallel(subject=nrow(retail1_sub),var=ncol(retail1_sub))
nS <- nScree(x=eig$values, aparallel=ap$eigen$qevpea)
plotnScree(nS)
```




```
# install.packages("psych")
library(psych)
library(GPArotation)

## check eigenvalues of components (14>1.0 required threshold)
parallel <- fa.parallel(retail1_sub, fm = 'minres', fa = 'fa')
```




```
## Parallel analysis suggests that the number of factors =  10  and the number of components =  NA
```

```
#parallel
```

initial findings

The first part of my analysis focuses on using factor analysis, which is a method that explores a reduced correlation matrix to find variables with common variance and weighs the variables based on this commonality to create “factors”. I first looked at the eigenvalues when deciding the number of factors I wanted to keep. These eigenvalues represent the sum of squared loadings and should be >1.0 to be valid. There were 14 that met this criteria in this set. We can to look at the chart (“Non Graphical Soulutions…”) to see that as we add more factors we are able to explain away more of the variance, but this comes at the sake of interpretability and simplicity of the model. I then used parallel analysis and confirmed with another look at the eigenvalues to choose a cut-point of 9 factors. These eigenvalues for the selected number of factors ranged from 1.9-15.0, well above the needed threshold (1.0).

```
## output factor loadings for interpretation
#We run FA using "Varimax" rotation with "minres" (minimum residual)
retail.fa <- fa(retail1_sub,nfactors = 9,rotate = "Varimax",fm="minres",scores="regression")
retail.fa

#trim loadings to ease intepretation
print(retail.fa$loadings,cutoff = 0.35, sort = TRUE)
```

```r
#This function visually shows the loading
retail.score<-as.data.frame(retail.fa$scores)
```

```r
#rename the new factors
colnames(retail.score) <- c("rich_retail", "young_educated","retail_no_vehicle", "poor_large_household"
                            ,"black_single_mom", "young_uneducated_unemployed" ,"divorced_single_dad"
                            ,"rural_multivehicle", "high_growth")
# retail.score

#Combine the new factors to the original dataset (if needed).
retail.final<-cbind(retail, retail.score)

## write to csv for part 2
#write.csv(retail.final, file="retail_w_clusters.csv")
```

factor analysis findings

Now that I had chosen my number of factors, I then started my analysis of the factor loadings (correlation between the original features and the newly formed factors). The first step in this process was to eliminate values in the variable-feature table with correlation coefficients below the absolute value of 0.35, which made the process of focusing in on the top weighted variables easier. I then analyzed each of the factors starting in descending order of their eigenvalues. The top factor had high correlations to no vehicle, high population density, renters, single– so I gave it the nickname "city_no_vehicle". The next feature had positive correlations with younger age groupings (18-34), negative with older groups, and high education levels (bach/grad degrees) – so I gave it the nickname "young_educated". The next group had high commute times (30+ minutes), high income, and population density metrics – so I called this factor "rich_city". The 4th highest factor had high correlation with large families (3+), Hispanic, low education levels (high school or less), and renters – this group is "poor_large_household". Factor 5 had high correlations with black race, single parent females – "black_single_moms". The next group had high correlation with low education, low income or unemployed, young people (18-34) – so I called them "young, uneducated, unemployed". Grouping 7 was high on divorced, single dads, working in city – "divorced_single_dads". The 8th factor had correlation with multivehicle (2+), workers in different area, low population density metrics – so I called them "rural_multivehicle". My final group captured two of three growth metrics (household and population) – "high_growth". After naming each of my factors, I tied it to the original set of cities (MSAs) to analyze the results (see Appendix A-1). I started through each of my factors looking at the top and bottom 10 city weightings. Based on my knowledge, "higher income" met my expectations with several east coast (including Suffolk county, NY – where the Hamptons are) and Bay area cities in the top 10. The top "young_educated" cities were spot on with every one of the top 10 being in relatively smaller cities with large state schools (with College Station, TX – Texas A&M as the largest). The bottom 10 also made sense, where all but three were in Florida where, presumably, older retirees live. "City no-vehicle" was highly weighted by the top few and fell off significantly after, but met expectations with NYC area and San Francisco as the top two. "Poor large household" was next and also was not surprising considering many were in more rural locations in Texas and California where, presumably, relatively more people of Hispanic roots live. "Black single moms" had Kansas City as the 4th highest, which is not surprising given my knowledge that there is a long wait list at BBBS KC for young boys, predominantly black, who do not have a male role model in their household. Nothing stood out to me in the "young, uneducated, unemployed" category except that I may have expected more than two cities in the south to make the top list. "Divorced single dads" was interesting in that many of the top cities (i.e. Las Vegas, Reno, Anchorage) would seem to be difficult on a marriage lifestyle with conditions/vices. "Rural multicar" top cities all made sense, but several in the bottom of this group were surprising (i.e. – Wichita, Kansas City). "High growth" met expectations, seeing Ann Arbor among the top and Detroit among the bottom are two that I would've expected.

Pt. 2: Linear Regression

The second half of my analysis focuses on using the our newly formed factors to analyze their relationship and predictive power on our set of dependent variables. This set of outcome variables includes 16 measures of grocery store and restaurant availability/usage.

```
## combine factors with dependent variables
retail1_sub <- retail1[ , which(names(retail1) %in% c("Groc_non_food","Groc_food", "S100_H", "S120_H"
                                                    ,"pq_g", "pq_r", "pqr_nonfood","pqr_food", "Groc_non_f
                                                    ,"Groc_food1","Sr4451_100", "Sr722_120", "Nr4451_100"
                                                    ,"Nr722_120","Share100_4445_72","Share4451_722"
                                                    ,"L722", "L4451"))]

colnames(retail.score) <- c("rich_retail", "young_educated","retail_no_vehicle"
                             ,"poor_large_household","black_single_mom"
                             , "young_uneducated_unemployed","divorced_single_dad"
                             ,"rural_multivehicle", "high_growth")

#Combine the new factors to the dependent variables
retail.final2<-cbind(retail1_sub, retail.score)
#head(retail.final2)

## look at prior distribution of target - Grocery food
summary(retail.final2$Groc_food)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.931   2.568   2.829   2.858   3.018   5.380
```

```
## create linear regression(s) to check feature importance
Sample.model1<-lm(data = retail.final2, Groc_food ~ 0 +rich_retail+young_educated+retail_no_vehicle
                  + poor_large_household + black_single_mom + young_uneducated_unemployed
                  + divorced_single_dad +rural_multivehicle + high_growth)

summary(Sample.model1)
```

```
##
## Call:
## lm(formula = Groc_food ~ 0 + rich_retail + young_educated + retail_no_vehicle +
##     poor_large_household + black_single_mom + young_uneducated_unemployed +
##     divorced_single_dad + rural_multivehicle + high_growth, data = retail.final2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  1.929  2.606  2.811  3.048  4.635
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## rich_retail                  0.190552   0.182300   1.045    0.297
## young_educated              -0.029189   0.174612  -0.167    0.867
## retail_no_vehicle           -0.002679   0.170316  -0.016    0.987
## poor_large_household         0.070410   0.184458   0.382    0.703
## black_single_mom            -0.077649   0.182474  -0.426    0.671
```

```
## young_uneducated_unemployed -0.050715   0.174316  -0.291    0.771
## divorced_single_dad         -0.082617   0.186515  -0.443    0.658
## rural_multivehicle          -0.013682   0.115159  -0.119    0.906
## high_growth                 -0.018201   0.178592  -0.102    0.919
##
## Residual standard error: 2.936 on 243 degrees of freedom
## Multiple R-squared:  0.007051,   Adjusted R-squared:  -0.02972
## F-statistic: 0.1917 on 9 and 243 DF,  p-value: 0.9949
```

```
## get odds ratio
exp(Sample.model1$coefficients)
```

```
##                  rich_retail               young_educated
##                    1.2099168                    0.9712331
##            retail_no_vehicle        poor_large_household
##                    0.9973243                    1.0729482
##             black_single_mom young_uneducated_unemployed
##                    0.9252888                    0.9505493
##          divorced_single_dad          rural_multivehicle
##                    0.9207039                    0.9864107
##                  high_growth
##                    0.9819632
```

```
## repeat process for:
## Share100_4445_72 - Number of grocery stores
## Nr4451_100 - ratio of grocery sales to total food sales
## S120_H - total resturant sales
## Sr722_120 - sales per resturant
## Nr4451_100 - sales per resturant
```

regression findings

The first dependent variable I looked at was the ratio of grocery sales to total food sales (grocery + restaurants). The range for this category had a minimum value of 0.418 (41.8% grocery sales) to 0.779 (77.9% grocery) with a median ratio of grocery sales at 61.5 percent. After analyzing the covariates, only one (city_no_vehicle) did not pass the p-test for statistical significance. "Poor large household" had the largest positive weight (more grocery store sales) and "young educated" had largest negative weight (more likely to eat out). These results seemed intuitive, given the cities of primarily large state universities where people are more likely to have irregular schedules and eat out more, though a majority of factors had a negative weights. I then shifted my analysis to look at total usage and looked at restaurant sales per household. The range is between 1.4-5.9, which I will assume is visits per week. After running the model to look at the coefficients - "rich city" is the highest weighting and "poor large household" is the lowest, which is not surprising after seeing that the large families had the highest relative expenditures on groceries. Additionally "rural multi-vehicle" was also negatively weighted, affirming that this group may not have as easy access to restaurants.

The next area of focus was the sales per restaurant. The cities ranged from 246-730, with a median of 490 (I will assume this number to be $/month by household). "Black single moms" had the highest weight for this grouping, which makes sense if you consider all of how busy it would be to have a job and raise kids all by yourself (leaving no time for groceries/cooking). The most negative weight for this group was the "rural multivehicle" and confirms my theory about their situation. I finished my analysis of relationships between the factors and dependent variable by looking at actual availability of grocery stores hoping to identify "food deserts". The range for number of grocery stores ranged from 0.35-1.71, with a median of 0.85 (assuming this is number is average number of grocery stores within a mile). “City no vehicle” and “rich city” had the highest positive coefficients with “rural multivehicle” having a negative weight, which leads me to believe that this is just a proxy for population density – so naturally there will be more grocery stores in your area. The concept of “food deserts” talks about the availability of quality/healthy grocery stores with fresh produce, which would require more granular (somewhat subjective) data.

Conclusion

Overall, factor analysis proved to be very powerful in making a large set of features interpretable. These insights could help city planners and government officials better understand the people they serve. If acted upon, these insights could make a material difference for certain populations living in these cities.

Appendix - A1

				Top 10				
rich_city	young_educated	city_no_vehicle	poor_large_household	black_single_mom	young_uneducated_unemployed	divorced_single_dads	rural_multicar	Top 10 - high_growth
Stamford--Norwalk, CT PMSA	Bryan--College Station, TX MSA	New York, NY PMSA	McAllen--Edinburg--Mission, TX MSA	Lowell, MA--NH PMSA	McAllen--Edinburg--Mission, TX MSA	Reno, NV MSA	Bellingham, WA MSA	Ann Arbor, MI PMSA
San Jose, CA PMSA	Ann Arbor, MI PMSA	Jersey City, NJ PMSA	Brownsville--Harlingen--San Benito, TX MSA	Nashville, TN MSA	Bellingham, WA MSA	Anchorage, AK MSA	Kenosha, WI PMSA	Fayetteville--Springdale--Rogers, AR MSA
Nassau--Suffolk, NY PMSA	Bloomington, IN MSA	San Francisco, CA PMSA	Visalia--Tulare--Porterville, CA MSA	Atlanta, GA MSA	Bryan--College Station, TX MSA	Santa Cruz--Watsonville, CA PMSA	Santa Cruz--Watsonville, CA PMSA	Bellingham, WA MSA
Middlesex--Somerset--Hunterdon, NJ PMSA	Raleigh--Durham--Chapel Hill, NC MSA	Lowell, MA--NH PMSA	Merced, CA MSA	Kansas City, MO--KS MSA	Gainesville, FL MSA	Medford--Ashland, OR MSA	Bremerton, WA PMSA	Daytona Beach, FL MSA
Santa Cruz--Watsonville, CA PMSA	Columbia, MO MSA	Hagerstown, MD PMSA	Provo--Orem, UT MSA	Baton Rouge, LA MSA	Monroe, LA MSA	Las Vegas, NV--AZ MSA	Medford--Ashland, OR MSA	Charlottesville, VA MSA
Bergen--Passaic, NJ PMSA	Provo--Orem, UT MSA	Madison, WI MSA	El Paso, TX MSA	Flint, MI PMSA	Santa Cruz--Watsonville, CA PMSA	Olympia, WA PMSA	Ann Arbor, MI PMSA	Ocala, FL MSA
Washington, DC--MD--VA--WV PMSA	Killeen--Temple, TX MSA	Eugene--Springfield, OR MSA	Riverside--San Bernardino, CA PMSA	Fort Worth--Arlington, TX PMSA	Stamford--Norwalk, CT PMSA	Tucson, AZ MSA	Charlottesville, VA MSA	Naples, FL MSA
Danbury, CT PMSA	Colorado Springs, CO MSA	Roanoke, VA MSA	Stockton--Lodi, CA MSA	New Orleans, LA MSA	Brownsville--Harlingen--San Benito, TX MSA	Albuquerque, NM MSA	Elkhart--Goshen, IN MSA	Hickory--Morganton--Lenoir, NC MSA
Newark, NJ PMSA	Champaign--Urbana, IL MSA	Santa Rosa, CA PMSA	Bakersfield, CA MSA	Roanoke, VA MSA	Bloomington, IN MSA	Sacramento, CA PMSA	Akron, OH PMSA	Kenosha, WI PMSA
Boston, MA--NH PMSA	Bloomington--Normal, IL MSA	Kansas City, MO--KS MSA	Modesto, CA MSA	Houston, TX PMSA	Medford--Ashland, OR MSA	Spokane, WA MSA	Huntsville, AL MSA	Las Vegas, NV--AZ MSA
				Bottom 10				
Yakima, WA MSA	Sarasota--Bradenton, FL MSA	Bellingham, WA MSA	Columbia, MO MSA	Bellingham, WA MSA	Lancaster, PA MSA	State College, PA MSA	Lowell, MA--NH PMSA	Lowell, MA--NH PMSA
Merced, CA MSA	Fort Myers--Cape Coral, FL MSA	Kenosha, WI PMSA	Champaign--Urbana, IL MSA	Santa Cruz--Watsonville, CA PMSA	Lowell, MA--NH PMSA	Lowell, MA--NH PMSA	Kansas City, MO--KS MSA	Santa Rosa, CA PMSA
Columbus, GA--AL MSA	Naples, FL MSA	Medford--Ashland, OR MSA	Bloomington, IN MSA	Medford--Ashland, OR MSA	York, PA MSA	Danbury, CT PMSA	San Francisco, CA PMSA	Springfield, IL MSA
Fort Walton Beach, FL MSA	Fort Pierce--Port St. Lucie, FL MSA	Ann Arbor, MI PMSA	Tallahassee, FL MSA	Bloomington, IN MSA	Boise City, ID MSA	Monmouth--Ocean, NJ PMSA	Madison, WI MSA	Kansas City, MO--KS MSA
Killeen--Temple, TX MSA	Chico--Paradise, CA MSA	Akron, OH PMSA	Gainesville, FL MSA	Santa Barbara--Santa Maria--Lompoc, CA MSA	Fort Walton Beach, FL MSA	Provo--Orem, UT MSA	Chico--Paradise, CA MSA	Wichita, KS MSA
Lakeland--Winter Haven, FL MSA	Daytona Beach, FL MSA	Huntsville, AL MSA	Bangor, ME MSA	Kenosha, WI PMSA	Grand Rapids--Muskegon--Holland, MI MSA	Lancaster, PA MSA	Boise City, ID MSA	Chico--Paradise, CA MSA
Longview--Marshall, TX MSA	Steubenville--Weirton, OH--WV MSA	Bremerton, WA PMSA	Boulder--Longmont, CO PMSA	Bremerton, WA PMSA	Kansas City, MO--KS MSA	Nassau--Suffolk, NY PMSA	Wichita, KS MSA	Bangor, ME MSA
Sharon, PA MSA	West Palm Beach--Boca Raton, FL MSA	Santa Cruz--Watsonville, CA PMSA	San Francisco, CA PMSA	Elkhart--Goshen, IN MSA	Appleton--Oshkosh--Neenah, WI MSA	Bryan--College Station, TX MSA	Santa Rosa, CA PMSA	Detroit, MI PMSA
Parkersburg--Marietta, WV--OH MSA	Ocala, FL MSA	Elkhart--Goshen, IN MSA	Charleston, WV MSA	Charlottesville, VA MSA	Hagerstown, MD PMSA	Reading, PA MSA	Fort Walton Beach, FL MSA	Columbus, GA--AL MSA
Biloxi--Gulfport--Pascagoula, MS MSA	Sharon, PA MSA	Charlottesville, VA MSA	Madison, WI MSA	Ann Arbor, MI PMSA	Wichita, KS MSA	Lafayette, IN MSA	Roanoke, VA MSA	Brockton, MA PMSA

“‘