BBVA Research

Unsupervised Learning for Industry Classification, State Economies, and MSA Sustainability

Marcial Nava

May 2021

Creating Opportunities

Key messages

- GMM and K-means allow us to produce classification systems to evaluate industry performance, analyze state economic resiliency, and assess MSA sustainability.
- Our industry classification system (ICS) helps identify common patterns and the impact of shocks for 1075 industries.
- States less exposed to severely impacted industries show greater economic resilience despite poor pandemic control.
- Cluster analysis suggests that the biggest upside for sustainable banking products is in Western markets.
- The analysis of extreme observations helps detect risks and opportunities.
- Unsupervised learning is a powerful tool for regional and industry economic analysis, lending strategy, customer segmentation, fraud detection, talent acquisition, etc.

An in-house Industry Classification System

OBJECTIVES:

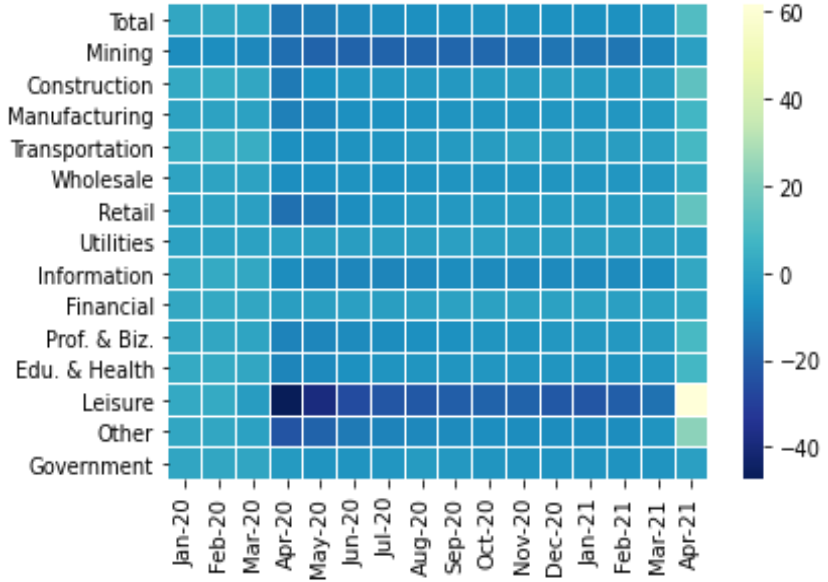
- Develop an industry classification system (ICS) to identify risks and opportunities for the bank.
- Apply unsupervised machine learning techniques that can learn patterns from untagged data.
- Assess the potential and limitations of K-means and Gaussian Mixture Models (GMM) to construct an effective ICS.

USERS:

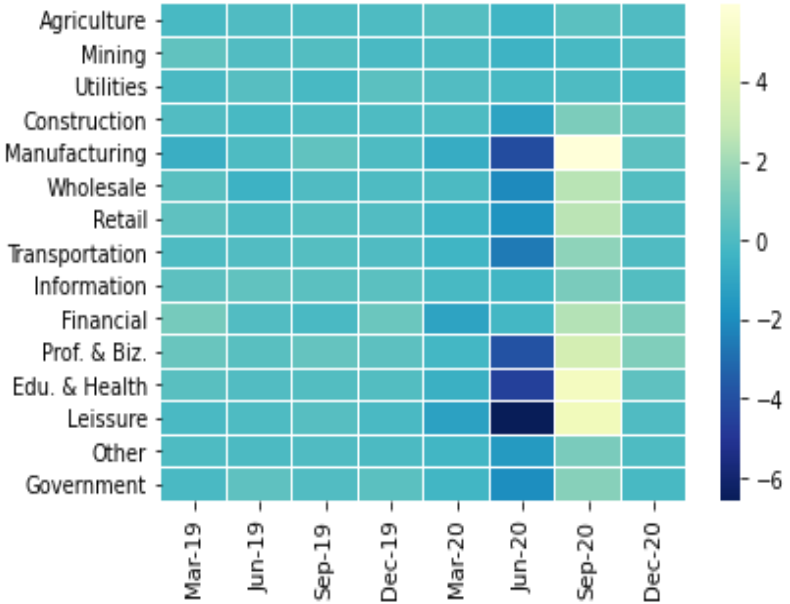
- Commercial and Industrial Lending.
- Corporate and Investment Banking.
- Risk Department.
- Auto Lending.
- Strategy.

High levels of industry aggregation are not enough for decision making since they mask differences at the granular level

NONFARM PAYROLL
(YOY % CHANGE)



CONTRIBUTIONS TO PERCENTAGE CHANGE IN REAL GDP GROWTH BY INDUSTRY (%)

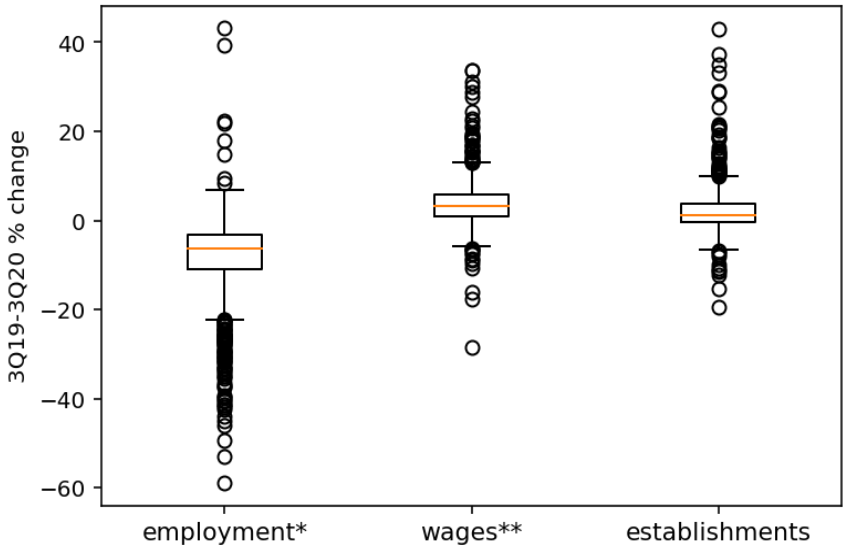


Source: BBVA Research, BLS and BEA.

QCEW is one of the most disaggregated sources of industry data

- Quarterly Census of Employment and Wages.
- Covers 95% of U.S. jobs.
- Available at county, MSA, state and national levels.
- Includes information of 1,075 industries under the 6-digit North American Industry Classification System (NAICS).
- Latest available data: 3Q20.

BOXPLOT: DISTRIBUTION OF INDUSTRIES
(6-DIGIT NAICS CODES, MEDIAN IN ORANGE)



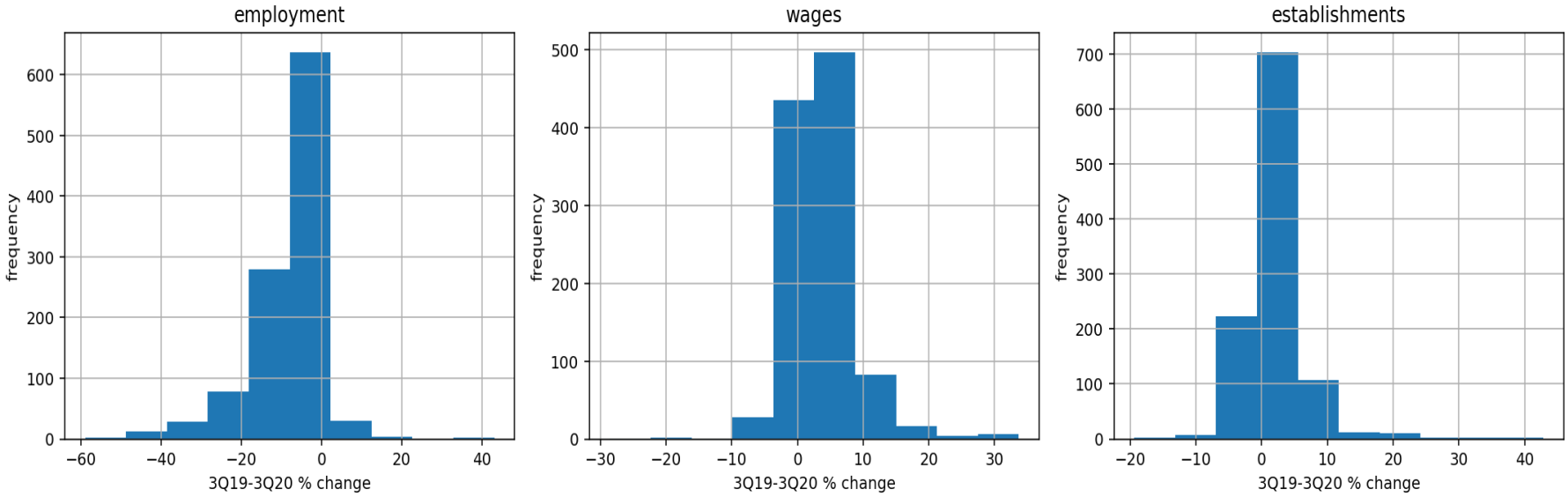
*Employment per establishment. **Wages per employee.
Source: BBVA Research and BLS.

Source: BBVA Research and BLS.

The distribution of employment growth is negatively skewed as the labor market absorbed the impact of the pandemic

DISTRIBUTION OF 6-DIGIT NAICS NDUSTRIES BY EMPLOYMENT*, WAGES** AND ESTABLISHMENTS

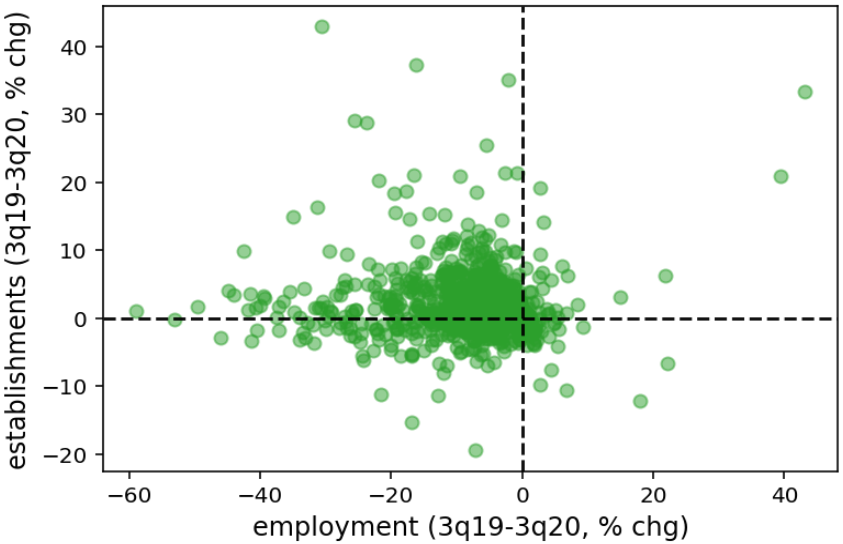
(YOY % CHANGE AND FREQUENCY)



*Employment per establishment. *Wages per employee. Source: BBVA Research, BLS and BEA.

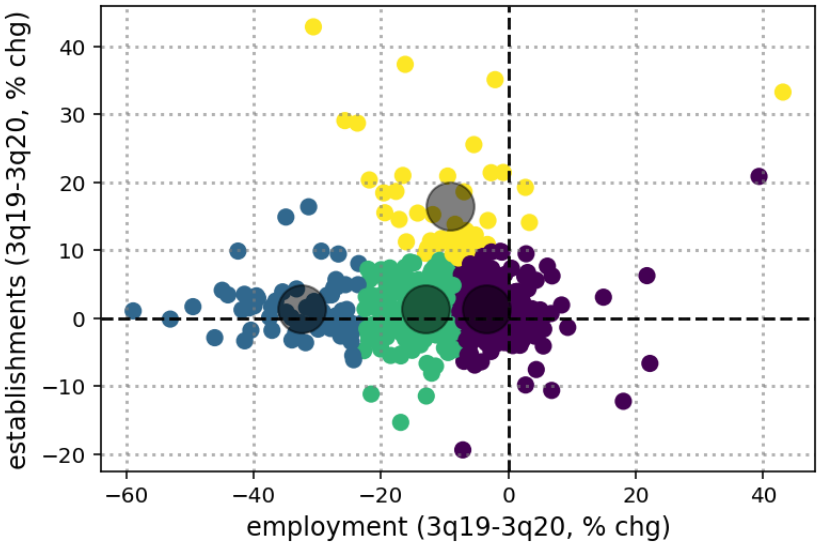
Unsupervised ML can help identify patterns in industry performance at the most granular level (6-digit NAICS codes)

6-DIGIT NAICS EMPLOYMENT*AND NUMBER OF ESTABLISHMENTS



6-DIGIT NAICS SEGMENTATION USING K-MEANS WITH 4 ARBITRARILY SELECTED CLUSTERS

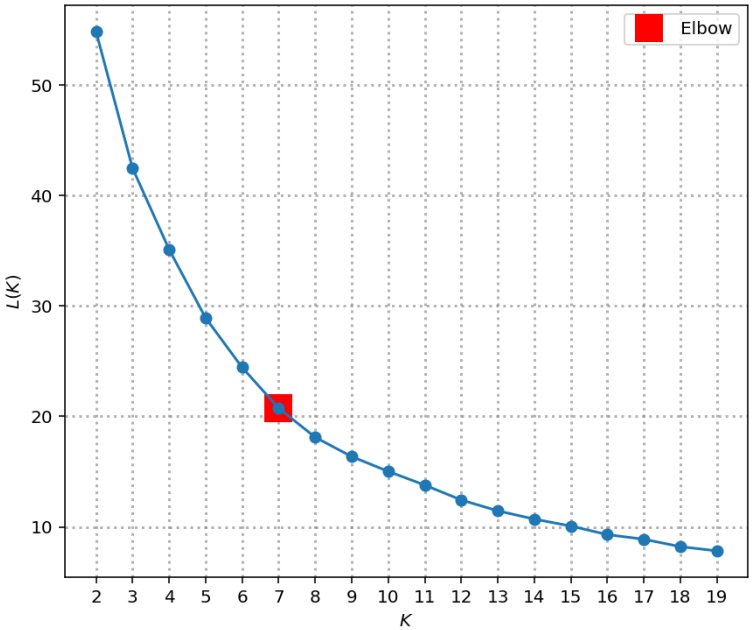
(K =4, CENTROIDS IN SHADED CIRCLES)



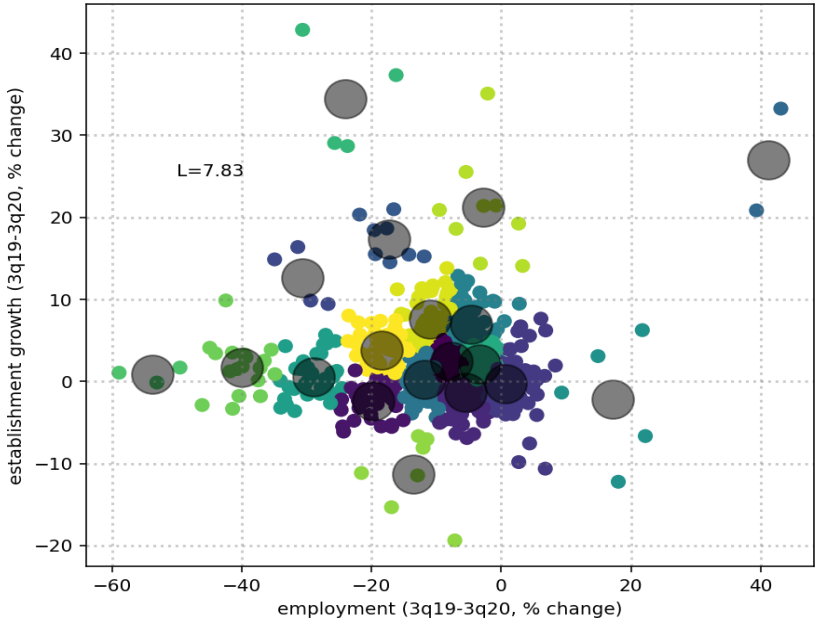
*Employment per establishment. Source: BBVA Research, BLS and BEA.

To prevent overfitting, we find the optimal number of clusters using tools like the Elbow Method

ELBOW METHOD USING INERTIA
(K= NO. OF CLUSTERS, L(K) = COST FUNCTION)



K-MEANS OVERFITTING EXAMPLE WITH K= 19
(CENTROIDS IN SHADED CIRCLES, L = INERTIA COEFFICIENT)



Source: BBVA Research with BLS data.

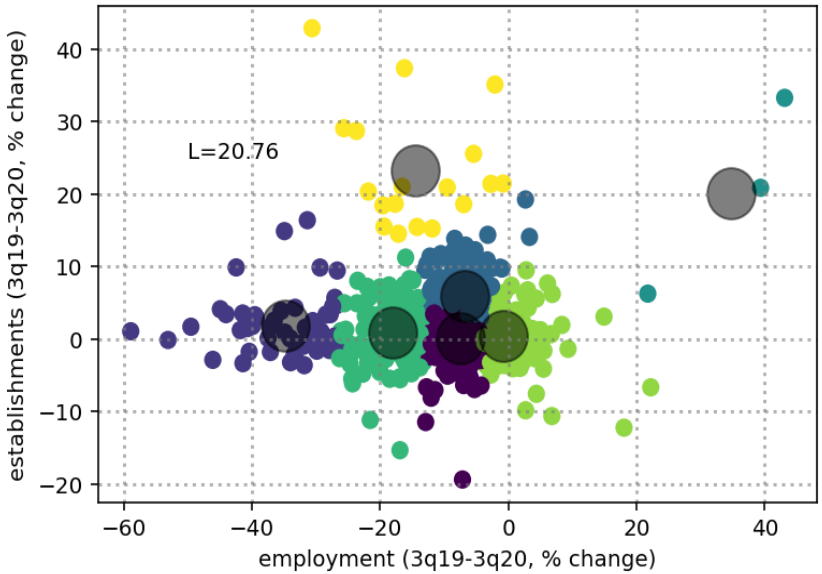
K = 7 seems to be the optimal solution for our ICS, but at the expense of interpretability

CLUSTER CENTROIDS
(3Q19-3Q20 % CHANGE, K = 7)

Cluster	Avg. employment* growth	Avg. establishment growth
cluster_1	-7.5	0.1
cluster_2	-34.8	1.9
cluster_3	-6.8	5.9
cluster_4	34.7	20.1
cluster_5	-18.1	0.9
cluster_6	-0.8	0.5
cluster_7	-14.6	23.3

*Employment per establishment. Source: BBVA Research and BLS.

6-DIGIT NAICS SEGMENTATION USING K-MEANS
(K= 7, CENTROIDS IN SHADED CIRCLES, L = INERTIA COEFFICIENT)



K = 3 could do a better job from an interpretability point of view

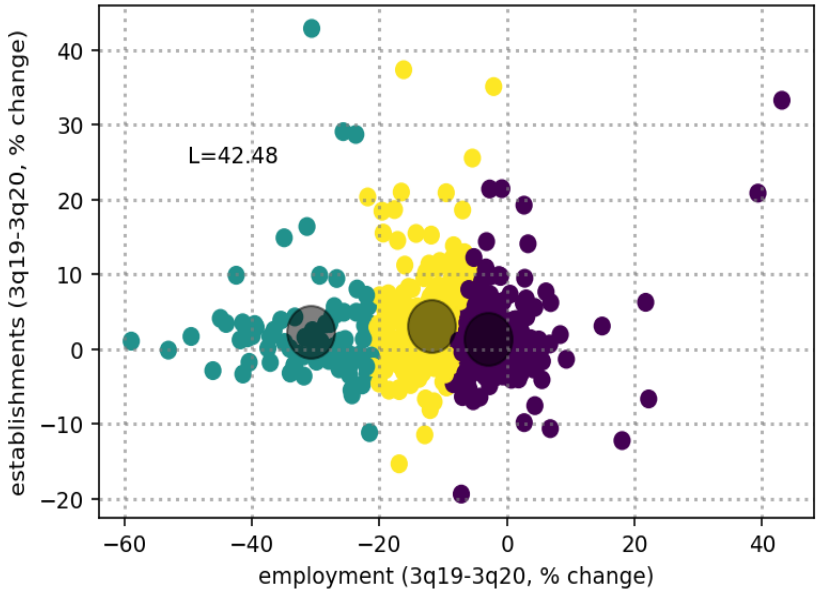
CLUSTER CENTROIDS
(3Q19-3Q20 % CHANGE, K = 3)

cluster	Avg. employment growth	Avg. establishment growth
cluster_1	-3.0	1.3
cluster_2	-30.7	2.3
cluster_3	-11.9	3.0

Interpretability based on employment figures		
Mild	-3.0	1.3
Severe	-30.7	2.3
Strong	-11.9	3.0

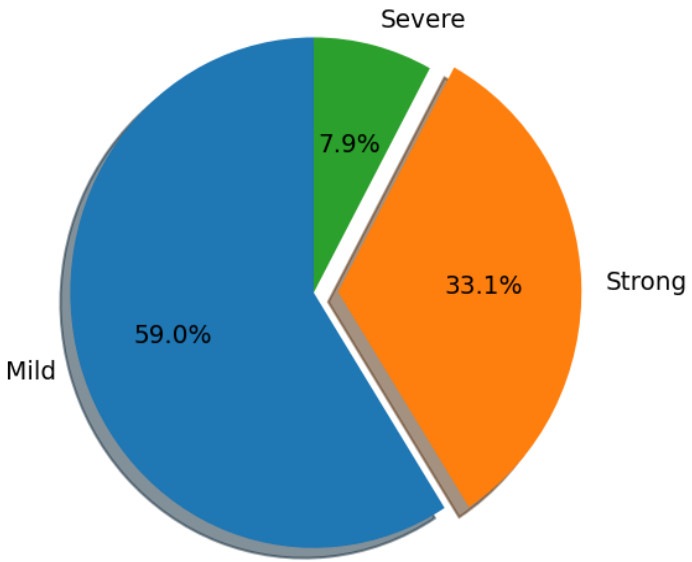
Source: BBVA Research and BLS.

6-DIGIT NAICS SEGMENTATION USING K-MEANS
(K =3, CENTROIDS IN SHADED CIRCLES, L = INERTIA COEFFICIENT)



Less than 10% of industries fall into the group that was severely impacted by the pandemic…

6-DIGIT NAICS DISTRIBUTION AMONG CLUSTERS



Severe impact (worst performers)

NAICS 483114 Coastal and great lakes passenger transport.
NAICS 512131 Motion picture theaters, except drive-ins
NAICS 711190 Other performing arts companies
NAICS 487110 Scenic and sightseeing transportation, land
NAICS 713110 Amusement and theme parks

Strong impact (worst performers)

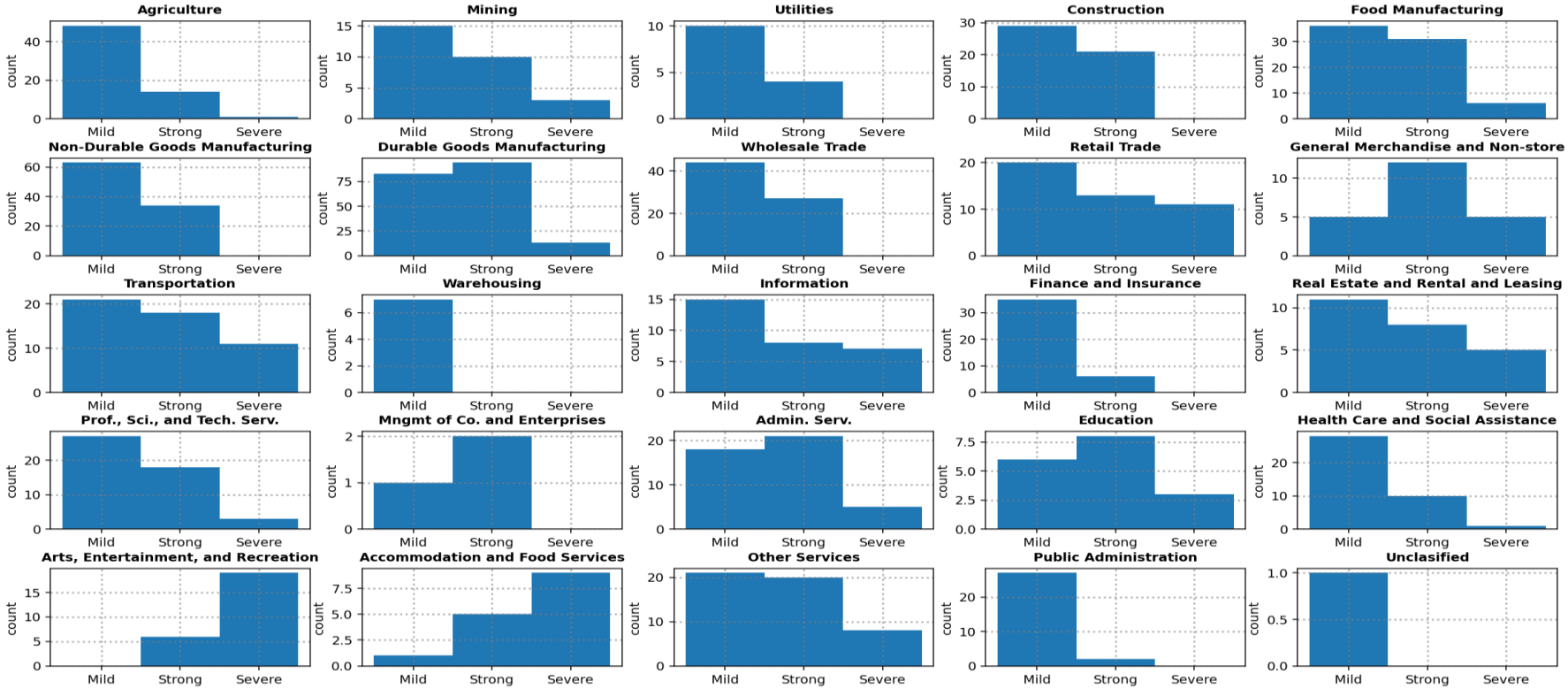
NAICS 561920 Convention and trade show organizers
NAICS 721214 Recreational and vacation camps
NAICS 711320 Promoters without facilities
NAICS 541921 Photography studios, portrait
NAICS 812930 Parking lots and garages

Mild impact (worst performers)

NAICS 813410 Civic and social organizations
NAICS 812199 Other personal care services
NAICS 713920 Skiing facilities
NAICS 532282 Video tape and disc rental
NAICS 522293 International trade financing

Source: BBVA Research and BLS.

…most of them in sectors that rely on physical proximity



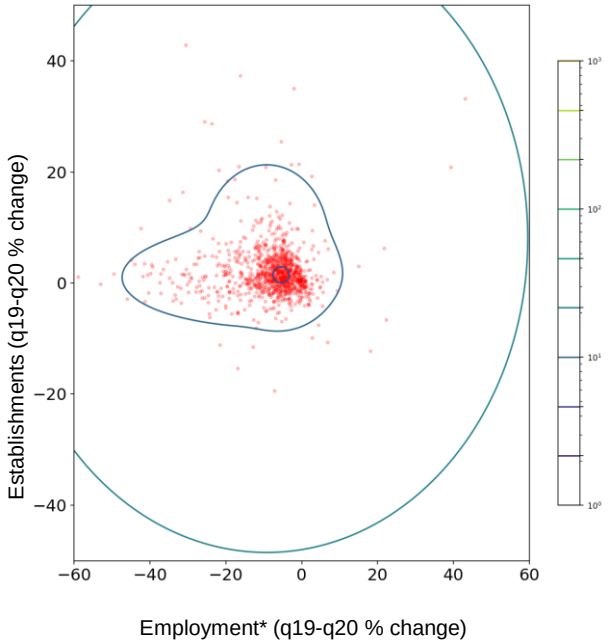
An ICS based on GMM adds more flexibility as it accounts for uncertainty

GMM PREDICTED PROBABILITIES
(SELECTED INDUSTRIES)

Industry code	Description	P(X ∈ Mild)	P(X ∈ Severe)	P(X ∈ Strong)
111160	Rice farming	0.94	0.04	0.02
312130	Wineries	0.74	0.22	0.05
532420	Office equipment rental and leasing	0.04	0.87	0.09
623110	Nursing care facilities, skilled nursing	0.97	0.02	0.01
485112	Commuter rail systems	0.35	0.04	0.61

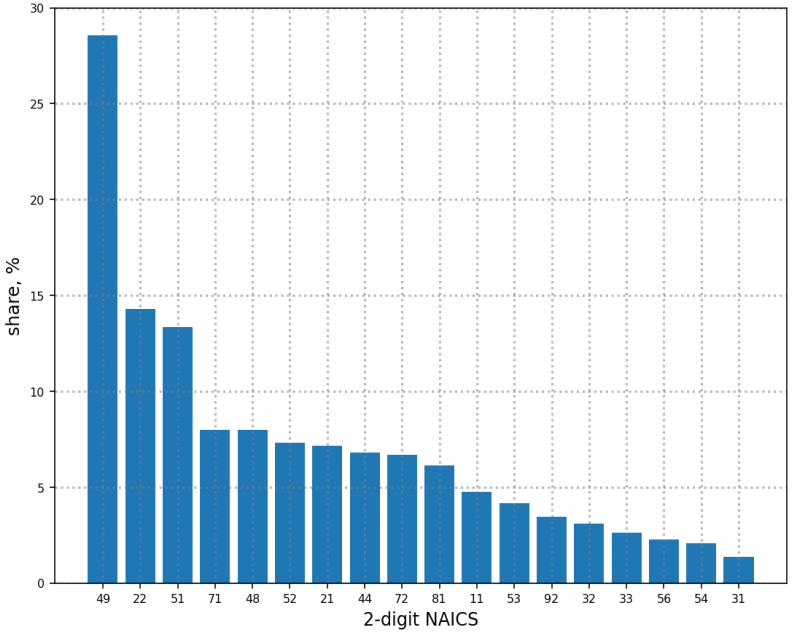
*Employment per establishment.
Source: BBVA Research with BLS data.

CONTOUR PLOT
(LOG-LIKELIHOOD PREDICTED BY A GMM)

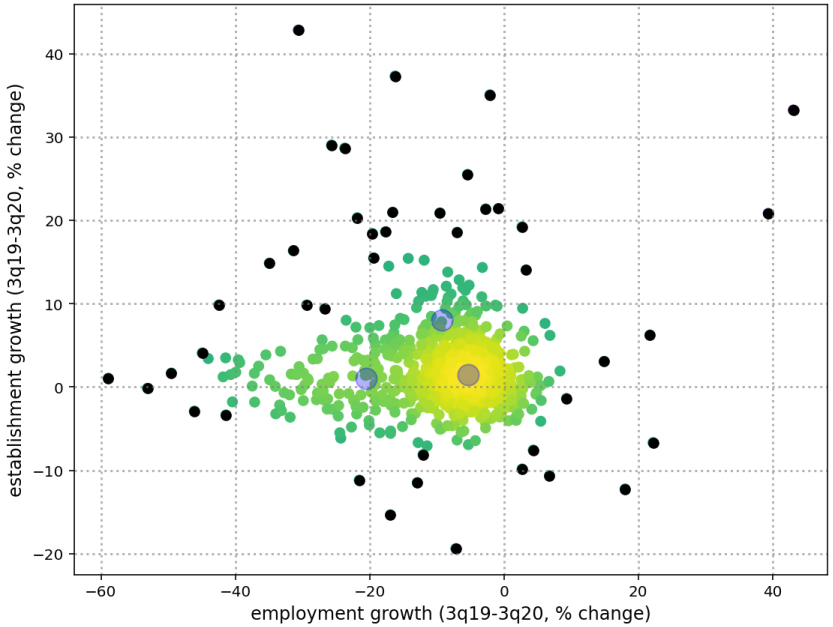


GMM can also help us spot extreme performers

2-DIGIT NAICS EXTREME VALUES DISTRIBUTION BASED ON GMM (%)



6-DIGIT NAICS EXTREME VALUES DETECTION BASED ON GMM



Source: BBVA Research with BLS data.

This can help identify opportunities and risks

TOP AND BOTTOM TEN INDUSTRIES BY EMPLOYMENT GROWTH

NAICS-6	NAICS-2	Top 10
926110	92	Administration of general economic programs
813940	81	Political organizations
212222	21	Silver ore mining
492210	49	Local messengers and local delivery
336419	33	Other guided missile and space vehicle parts
493110	49	General warehousing and storage
212392	21	Phosphate rock mining
322110	32	Pulp mills
111419	11	Other food crops grown under cover
322122	32	Newsprint mills

NAICS-6	NAICS-2	Bottom 10
485320	48	Limousine service
487210	48	Scenic and sightseeing transportation, water
722320	72	Caterers
561920	56	Convention and trade show organizers
711310	71	Promoters with facilities
713110	71	Amusement and theme parks
487110	48	Scenic and sightseeing transportation, land
711190	71	Other performing arts companies
512131	51	Motion picture theaters, except drive-ins
483114	48	Coastal and great lakes passenger transport.

Machine learning to analyze state economic recovery

OBJECTIVES:

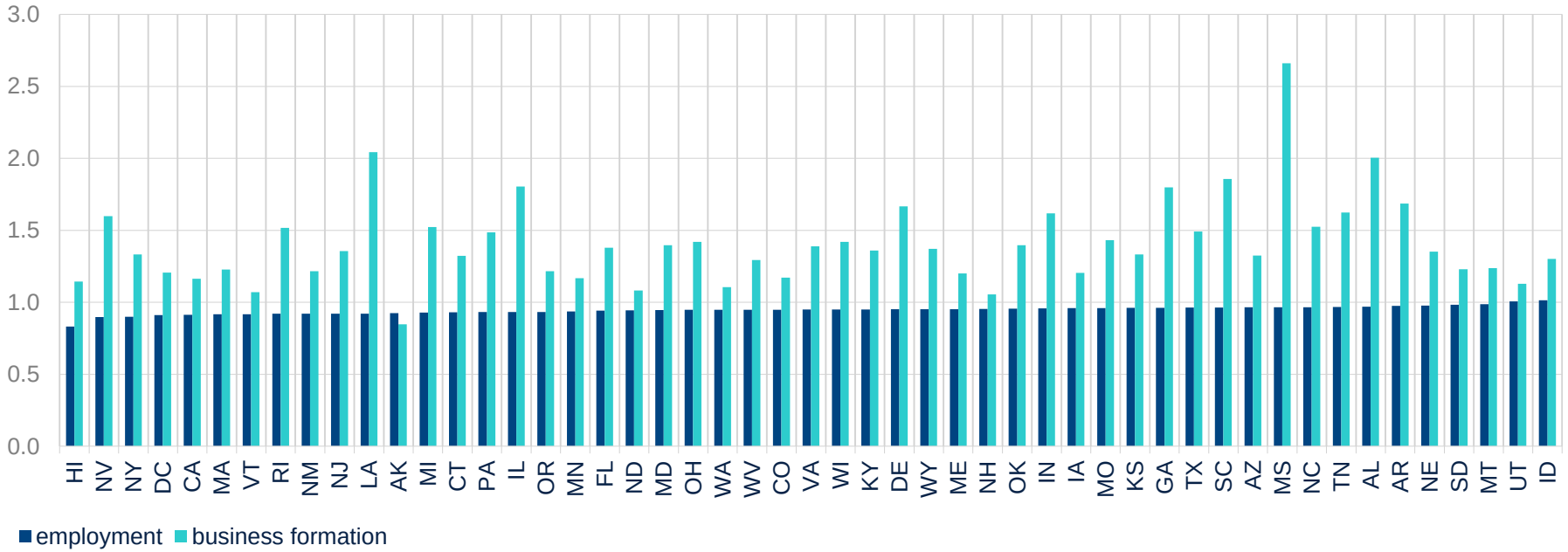
- Understand the link between pandemic control and economic recovery at the state level.
- Apply unsupervised machine learning techniques to classify states according to their level of “resiliency”.
- Support regional economic analysis and forecasting.

USERS:

- Business Intelligence.
- Strategy.
- Risk Department.
- CECL/CCAR Compliance.
- Corporate Responsibility.
- Marketing and Media.

American entrepreneurship counterbalanced labor market collapse

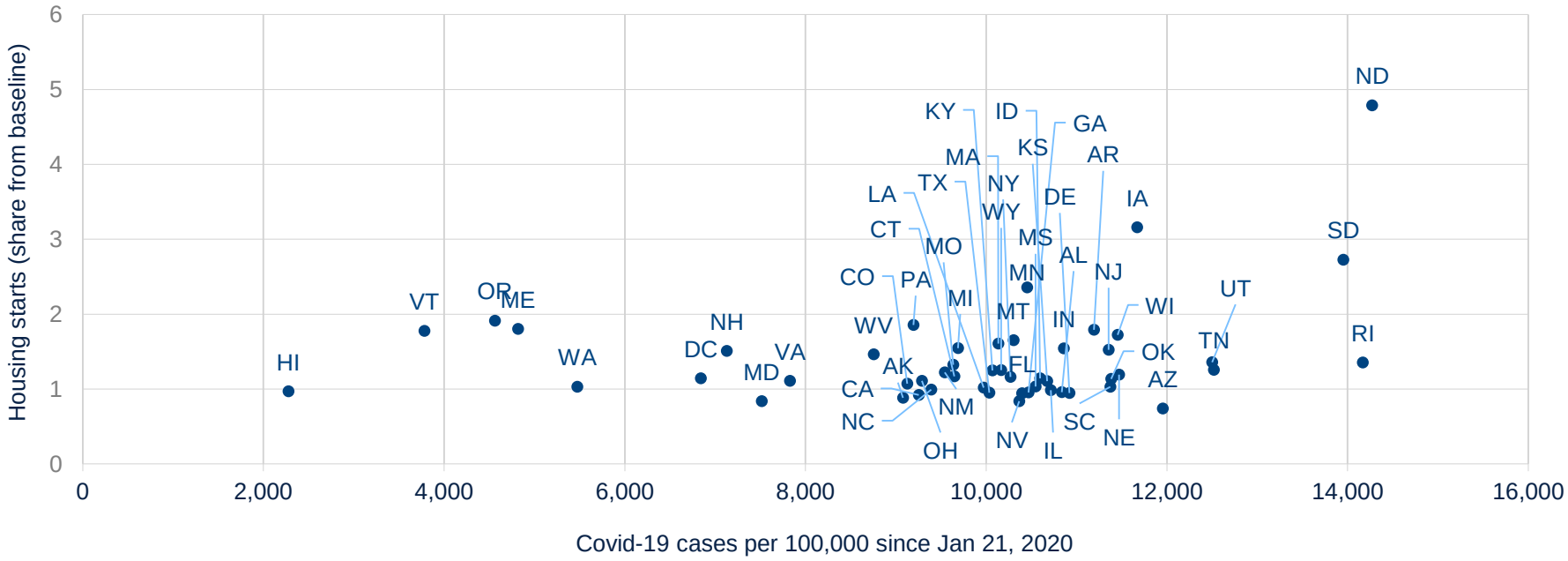
NONFARM PAYROLL AND BUSINESS FORMATION
(SHARE FROM BASELINE)



Source: BBVA Research and Haver Analytics.

Residential construction has thrived in markets severely impacted by the pandemic

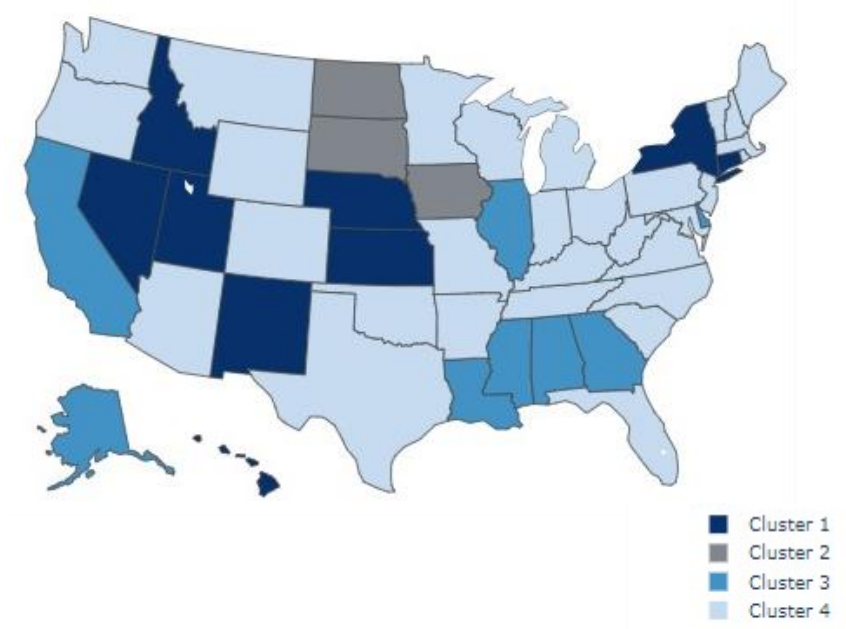
HOUSING STARTS AND COVID-19 CASES PER CAPITA



Source: BBVA Research, CDC and Haver Analytics.

GMM clusters highlight different levels of economic resiliency

GMM CLUSTERING BASED ON ECONOMIC RECOVERY AND CONTROL OF COVID-19



Source: BBVA Research, CDC and Haver Analytics.

CLUSTER 1
Weakest economic resiliency/mixed pandemic control
High exposure to severely impacted industries.

CLUSTER 2
Poor pandemic control/highest economic resilience
Large agricultural sector/low exposure to severely impacted industries.

CLUSTER 3
Poor pandemic control/highest business formation
High exposure to severely impacted industries and global economy.

CLUSTER 4
Average pandemic control and economic resiliency. Mixed industry exposure.

A machine learning application for MSA sustainability

OBJECTIVES:

- Identify sustainability risks and opportunities at the MSA and regional levels.
- Apply unsupervised machine learning techniques that can learn patterns from untagged data.
- Complement in-house MSA Sustainability Index with K-means and GMM models.

USERS:

- Commercial and Industrial Lending.
- Corporate and Investment Banking.
- Risk Department.
- Auto Lending.
- Residential Lending.
- Property Insurance.
- Strategy.

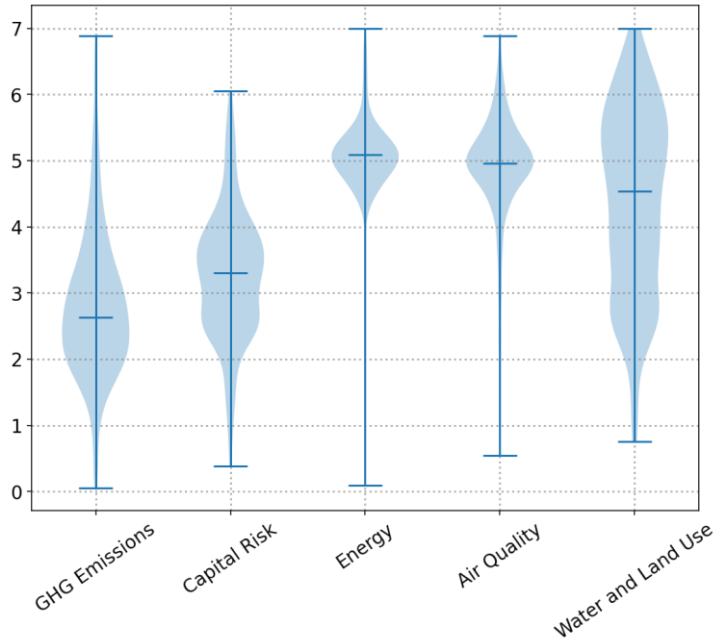
ML clustering serves to identify local and regional patterns

BBVA USA MSA SUSTAINABILITY INDEX
(MAIN COMPONENTS)

MSA	GHG Emissions	Capital Risk	Energy	Air Quality	Water and Land Use
Carson City, NV	4.9	5.2	5.1	4.6	6.5
Madera, CA	6.3	4.3	5.4	5.0	5.0
Yuba City, CA	4.1	5.6	5.7	4.9	5.3
Saginaw, MI	3.5	5.2	5.1	5.0	6.7
Abilene, TX	2.5	5.8	5.4	5.0	6.7
Fond du Lac, WI	3.1	5.5	5.1	5.3	6.2
Burlington, VT	5.5	2.9	5.1	6.1	5.6
Boise City, ID	4.3	5.3	5.4	5.9	4.3
Chico, CA	5.2	5.0	5.4	5.0	4.5
Wenatchee, WA	5.6	3.0	5.4	5.2	5.8

Source: BBVA Research.

MSA DISTRIBUTION ACROSS SUBINDIXES
(VALUE)



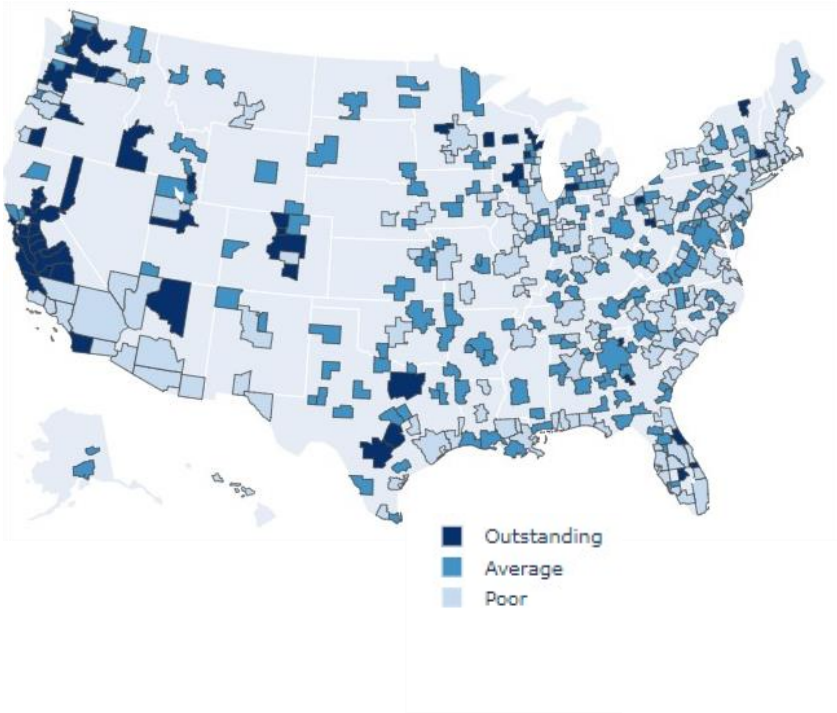
Western states have the largest number of sustainable MSAs

MSA SUSTAINABILITY CLUSTERS
(SELECTED MSAS)

MSA	Cluster
San Francisco-Oakland-Hayward, CA	Outstanding
Austin-Round Rock, TX	Outstanding
Denver-Aurora-Lakewood, CO	Outstanding
Columbus, OH	Average
Tuscaloosa, AL	Average
Atlanta-Sandy Springs-Roswell, GA	Average
Phoenix-Mesa-Scottsdale, AZ	Poor
Pittsburgh, PA	Poor
Houston-The Woodlands-Sugar Land, TX	Poor

Source: BBVA Research.

MSA SUSTAINABILITY CLUSTERS



The K-means model

Given a dataset $\{\mathbf{x}_n\}_{n=1}^N$, a **distance** metric between any two points $d(\mathbf{x}, \mathbf{x}')$ and a pre-defined number K of groups (clusters), K-means follows a series of steps to assign a point to any of the K posible groups:

1. Randomly select "centroids" $\mathbf{c}_k$ for each group k, $k = 1, \ldots, K$. These *centroids* will initiate the process.
2. For $n = 1, \ldots, N$, calculate the distance of all data points $\mathbf{x}_n$ to the centroids.
3. Assign data points to the closest cluster, according to a predefined metric. That is if, $A(\mathbf{x}_n) \in \{1, \ldots, K\}$ then:

$$A(\mathbf{x}_n) = \arg \min_{\{1,\ldots,K\}} d(\mathbf{x}_n, \mathbf{c}_k)$$

4. Recalculate the centroids of each cluster by taking the arithmetic mean of all data points in the cluster. For $k = 1, \ldots, K$,

$$\mathbf{c}_k = \frac{1}{N_k} \sum_{n:A(\mathbf{x}_n)=k} \mathbf{x}^{(n)},$$

where $\frac{1}{N_k}$ is the number of points assigned to grupo k.

5. Repeat steps 2, 3, and 4 until all points converge and the cluster stops moving.

Gaussian Mixture Models (GMM)

In a GMM, we fit our data to a probability density function as follows:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

where

- π_k the probability that a data point belongs to the k-th group (or cluster).
- We assume that all the data points assigned to the k-th group are normally distributed with mean μ_k and covariance matrix $\mathbf{\Sigma}_k$.
- $(\pi_1, \ldots, \pi_K)$, $(\mu_1, \ldots, \mu_K)$, $(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K)$ are the **parameters of the model**.

Training a GMM:

Parameters are chosen to maximize the probability of the observed data:

$$\max_{(\pi_1,\ldots,\pi_K),(\mu_1,\ldots,\mu_K),(\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K)} \sum_{n=1}^{N} \log p(\mathbf{x}_n)$$

The problem is solve using the Expectation-Maximization (**EM**) algorithm.

GMM can help us solve some of the limitations of K-means:

- K-means assumees that the shape of the clusters is "circular"
- K-means make "strong" assignments and does not allow for uncertainty

BBVA Research

Unsupervised Learning for Industry Classification, State Economies, and MSA Sustainability

Marcial Nava

May 2021

Creating Opportunities