Network-Centric Benchmarking of Operational Performance in Aviation

Karthik Gopalakrishnan[a,*], Max Z. Li[a], Hamsa Balakrishnan[a]

[a]*Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA*

Abstract

Performance analysis of the air traffic operations is challenging because of the need to account for weather impacts and network effects. In this paper, we propose a framework that uses network clustering to identify baselines for benchmarking airline on-time performance. We demonstrate our framework by computing cancellation and departure delay baselines using US flight data for the years 2014-16. Subsequently, we use these baselines to benchmark daily on-time performance at the system-wide, airline-, and airport-specific levels, for both mainline and regional carriers. This framework enables an airline to conduct self- and peer-comparisons, evaluate improvements over time, and diagnose causes of poor on-time performance. Furthermore, our framework can be used by system operators to identify long-term trends in traffic management initiatives.

Keywords: airline benchmarking, Ground Delay Programs, delay and cancellation networks, network performance, regional and mainline carriers, outlier analysis

1. Introduction

Large-scale, complex infrastructures, such as air transportation, are prone to operational inefficiencies which result in delays [1, 2]. In 2017, nearly 20% of domestic flights in the US arrived more than 15 minutes late. In addition to inconveniencing airline passengers, delays have significant environmental and economic consequences [3–5]. It is therefore important to analyze system performance, identify operational inefficiencies, and reduce flight delays.

The analysis of aviation system performance is confounded by several factors. Weather reduces airport and airspace capacities, resulting in flight cancellations and delays. Since no two days are identical in terms of the locations, timings and intensities of capacity impacts, it is difficult to distinguish between poor performance due to operational inefficiency and delays due to reduced capacity. Similarly, airport closures, security breaches, and equipment failures also impact capacity and cause delays. The complex dynamics of delay propagation

*Corresponding author

Email addresses: karthikg@mit.edu (Karthik Gopalakrishnan), maxli@mit.edu (Max Z. Li), hamsa@mit.edu (Hamsa Balakrishnan)

in the air transportation network also poses a challenge to performance analysis. More than one-third of flight delays in the U.S. occur because the previous flight operated by the same aircraft was delayed, making it the most common reason for a delayed flight [6]. By contrast, primary sources of delay such as weather, volume of demand, equipment failures, and security incidents, account for less than one-third of all delayed flights.

Benchmarking is a standard technique used to evaluate system performance and identify potential operational deficiencies. It involves the comparison of an observed operational metric with a pre-computed reference value, or *baseline* [7, 8]. The baseline values may be different for different operational scenarios or contexts. Airlines have applied benchmarking principles in many aspects of their operations, including the analysis of financial performance, environmental impact [9], quality-of-service metrics [10], and airport performance [11, 12]. For example, by benchmarking their refueling and flight rotation practices with Formula 1 racing, Southwest Airlines reportedly reduced their refueling times by 70% [7].

This paper addresses the need for performance baselines that are context-dependent, i.e., are conditioned on the locations, timings, and intensities of the weather impacts. Such baselines allow us to control for extraneous weather effects when analyzing the operational efficiency of different stakeholders. We use network clustering to develop context-dependant benchmarks to evaluate the operational performance of airlines and airports. The proposed framework has been deployed at a major US airline for conducting post-analysis of its daily on-time performance.

The main contributions of our work are as follows:

1. We cluster airport capacity impacts based on their timings, locations, and intensities, and use these clusters to develop conditional performance baselines.
2. We construct network-theoretic feature vectors for clustering a time-series of weighted and directed graphs, and use these feature vectors to identify characteristic delay and cancellation behavior.
3. We demonstrate how our baselines can be used by an airline to monitor its performance, and to draw comparisons to its competitors. We also show how our approach can be used to compare long-term trends in airport capacity impacts, and their effect on on-time performance.

The remainder of this paper is organized as follows: We review prior literature in Section 2. The methodology for computing baselines, benchmarking, and clustering network-wide impacts is presented in Section 3. In Section 4, we present the results of airport capacity impact (reflected by Ground Delay Programs or GDPs), delay, and cancellation clustering, and discuss their significance. We discuss an application to airline performance benchmarking in Section 5, and a cluster-based analysis of long-term trends in Section 6. We conclude with a summary and some directions for future work in 7.

2. Literature review

2.1. Clustering approaches to computing baselines

Clustering has been previously used to develop appropriate case-specific baseline metrics from non-homogeneous observations [13–16]. In air transportation networks, there is

significant non-homogeneity due to the strong dependence of operational performance on the spatio-temporal occurrences of capacity reductions. However, prior works on clustering for identifying similar days in the National Airspace System (NAS) have typically focused on weather impacts and capacity constraints at a small subset of airports, or have only considered aggregate metrics such as total delays, cancellations, or traffic volume [17, 18]. Furthermore, these prior studies have not considered applying their identification of similar NAS days to stakeholder-specific on-time performance baselines.

An on-time performance benchmarking framework for air transportation using network-level clustering has not previously been considered. One concern with clustering approaches for benchmarking is that there may not be perfect homogeneity between days within the same cluster [19, 20]. We address this concern by incorporating the within-cluster variance to obtain statistically significant confidence bounds for our baselines.

2.2. Clustering similar GDPs

Efforts to identify similar GDPs have either focused exclusively on individual airports (e.g. Newark EWR [21–23], San Francisco SFO [23]), or a small subset of airports [20, 24]. Although these methods could be extended to the scale of the entire system, they do not address two key factors: the temporal parameters of the GDP (i.e. when it was issued), and the relative magnitudes of capacity reduction at different airports (i.e. how do we quantify the relative "intensity" of GDPs). For instance, in [25], the authors present a method that can be scaled to the system-level, but they do not consider temporal variation in airport capacity or its intensity. Weather has been the focus of several clustering efforts [26, 27], but convective conditions in and of themselves do not represent the actual reduction in airport or airspace capacity. Since GDPs are a more direct measure of the disruption that is induced by convective weather and other irregular operations, we will use these traffic management initiatives (TMIs) within our benchmarking framework.

In our work, we propose clustering network-centric *feature vectors* (that is, vectors of real numbers that represent the characteristics of similarity and dissimilarity between data observations) that scale for the entire system. In addition, these feature vectors also incorporate a normalized metric quantifying the intensity of the GDP, along with temporal information regarding GDP impact.

2.3. Clustering delay networks

Networks have been used to model several aspects of air transportation, including connectivity and delay dynamics [28–30]. The increased prevalence of data from networks has led to the development of techniques for clustering graphs [31–38]. We emphasize the distinction between the *clustering of graphs*, or finding groups of similar graphs (considered in this paper), and *community detection*, or finding clusters of nodes within a graph [39]. The choice of an appropriate feature vector for clustering strongly influences the quality of clusters. Prior work on clustering graphs has used feature vectors that include network-theoretic measures of centrality, such as degree centrality, between-ness centrality, and eigencentrality [32–34]; although the lengths of such features scale linearly with the size of the graph, they are best-suited only for unweighted, undirected graphs. By contrast, the clustering of

days in the NAS requires feature vectors that can compare time-series of weighted, directed graphs; we therefore develop a suitable network-theoretic feature vector whose length scales linearly with the size of the graph.

2.4. Airline performance monitoring

Prior work on airline performance analysis has compared the aggregate statistics [6, 40], economic measures [41], or indicators of consumer satisfaction and level-of-service of different airlines [42]. Methods for airlines to perform self- and peer-comparison of operational parameters such as on-time performance have not been considered. To the best of our knowledge, the explicit consideration of network state in benchmarking has also remained unexplored. While earlier studies have explored the role of vertical integration in the economic sense [43], there has been limited work on the network [44] and operational benefits [45] of regional airline subsidiary partnerships. Our analysis enables the comparison of the operational performance of regional carriers and their mainline counterparts.

3. Benchmarking framework and clustering methodology

3.1. Benchmarking framework

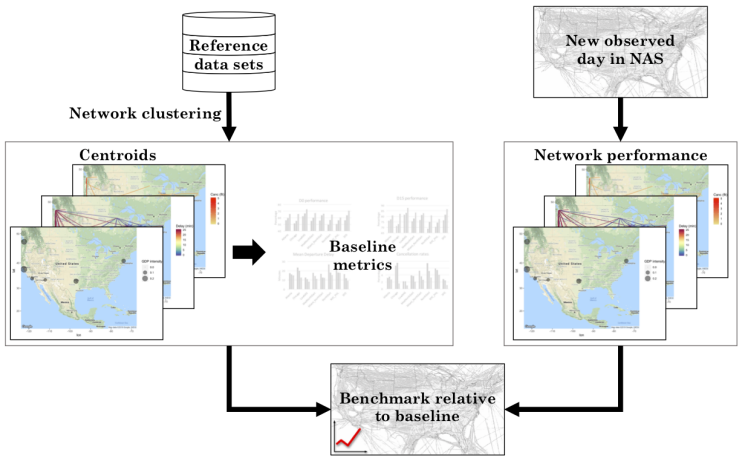

Figure 1: Flowchart representation of our benchmarking methodology framework.

Figure 1 shows a schematic of our benchmarking framework. The left side of the figure represents the computation of *baselines*, which are representative metrics. The *reference data set* contains information about all flights that operated within a certain time interval, e.g., a 2-year period. Clustering is used to categorize data points based on some parameter of interest; e.g., we cluster based on the GDP impacts in the system on that day. In this paper, each data point represents a day, and we identify representative days. Days that belong to a certain clusters will be used to compute baseline statistics for that cluster. This means that we can define any metric of importance, say mean departure delays, or number of canceled flights, or average diversions from Atlanta, or even be as specific as cancellations

on the Boston-San Francisco route for United Airlines. In other words, we can define any metric of importance, and compute baselines for each of the representative groups. On the right-hand side of Figure 1, we describe how we would perform *benchmarking* for a new day. *Benchmarking* is the process by which we compare the new day, with baselines corresponding to an appropriate cluster (or representative group) that the day maps to. Depending on the metrics chosen during the clustering and baseline computation, the benchmarking can be airline-, airport-, or route-dependent.

3.2. Clustering methodology

We approach the identification of similar GDP, delay, and cancellation networks for a given day as a classical clustering problem. The k-means algorithm [46] is a widely-used approach in unsupervised learning that splits the data points into k disjoint clusters such that the within-cluster sum-of-squares, i.e. the within-cluster error, is minimized. Our primary motivation for choosing k-means clustering is its simplicity, and the ability to directly control the number of clusters. Other clustering approaches are not as well-suited for our application: e.g., DBSCAN requires a hyperparameter that sets the minimum number of points in a cluster [47], k-medoids was found to give similar results as k-means clustering [30], spectral clustering is not applicable for a time-series of weighted directed graphs [48], and hierarchical clustering methods identify relations between clusters [49], a different objective from ours.

The k-means algorithm considers two data points to be similar if the 2-norm of the difference between their corresponding feature vectors is small, and proceeds to group these two data points within the same cluster. The mean of all data points that belong to a cluster, called the *centroid*, is representative of the entire cluster that it belongs to.

Feature selection, or the choice of vectors that best characterize the similarities and differences between groups of data points – in our case, capacity constraints, delays, and cancellations within the air transportation network – is a critical aspect of clustering. We differentiate between the network-centric feature vectors we construct for airport capacity reductions versus delays and cancellations. Specifically, GDP intensities are represented as node weights, whereas origin-destination delays and cancellations are represented as edge weights in our network. Next, we describe the construction of these feature vectors.

3.3. Feature engineering for GDP impacts

A GDP limits the rate at which aircraft can arrive at the impacted airport during a specified period of time, by delaying flights at their departure airports. Several factors can lead to a demand-capacity imbalance at an airport, potentially resulting in a GDP; causes include inclement weather events such as low ceilings, reduced visibility, and convective activities, construction activities, high demand volume, equipment outages, and even security and safety incidents. Each GDP issuance at an airport specifies the duration and a reduced capacity profile for that time interval.

In our benchmarking framework, we quantify the severity of the airport capacity reduction as the *intensity* of the GDP. For example, consider San Francisco International Airport (SFO), where the nominal airport arrival rate (AAR) is 60 aircraft per hour. At 2 pm,

weather forecasts indicate thunderstorms in the region from 4 to 7 pm. As a result of this potential capacity reduction, a GDP is issued, reducing the AAR at SFO to 20, 30, and 45 aircraft per hour from 4 to 7 pm, respectively. Our proposed GDP intensity metric is defined for each hour as:

$$\text{GDP intensity} = \frac{\text{Nominal rate} - \text{Reduced rate}}{\text{Nominal rate}}. \tag{1}$$

In the SFO example, the GDP intensities for the hours starting at 4 pm, 5 pm, and 6 pm are $\frac{60-20}{60} = 0.67$, $\frac{60-30}{60} = 0.5$, and $\frac{60-45}{60} = 0.25$, respectively. By computing the fractional reduction in capacity rather than the absolute value, the metric allows for a fair comparison of GDP impacts at different airports, accounting for airports with differing nominal AARs. A capacity reduction of 5 aircraft per hour will result in a higher GDP intensity at an airport with a nominal AAR of 30 aircraft per hour (the GDP intensity would be $(30-25)/30 \approx 0.17$), than at an airport with a nominal AAR of 80 aircraft per hour (the GDP intensity in this case would be $(80-75)/80 \approx 0.06$). The GDP intensity, as we have defined it, is well-defined at the limits: when no GDP is issued, the GDP intensity is 0; on the other hand, the most severe GDP, with an assigned AAR of zero, has an intensity of 1.

Airports that are geographically proximate to each other oftentimes experience similar weather conditions. However, varying traffic demand at these proximate airports may result in GDPs being issued at one airport but not at the other. Furthermore, even when two proximate airports experience a GDP of similar intensity, their network impacts may be completely different depending on the network connectivity of the airlines servicing those airports. For example, the three major commercial airports in New York (EWR, LGA, and JFK) frequently illustrate these network effects. Hence, we consider the GDP impacts at each airport independently from GDP impacts at other airports when constructing the feature vector for a given day. Every day of operation has an associated time series of GDP intensities for each airport in the NAS. These values are stacked together to form a feature vector that describes the capacity reduction at all the airports across each time step during that day. For each day, the feature vector is a $(24 \times N)$-dimensional vector of real numbers in the interval $[0, 1]$, where N is the number of airports in the network. Specifically, let $\mathbf{g}_{d,h} \in \mathbb{R}^{N\times 1}$ be a vector of the GDP intensities at all the airports on day d at hour h. Then, to cluster based on airport capacity impacts, we construct the feature vector for day d as:

$$\mathbf{f}_d^{\text{GDP}} = \begin{pmatrix} \mathbf{g}_{d,1} \\ \vdots \\ \mathbf{g}_{d,24} \end{pmatrix} \in \mathbb{R}^{24N\times 1}. \tag{2}$$

3.4. Feature engineering for delay and cancellation networks

We now detail the construction of feature vectors to perform network-centric clustering of delay and cancellation networks. We utilize a network representation of delays for constructing the feature vectors. A delay network is defined for each hour h of the day d, with nodes corresponding to airports, and the weight on each directed edge corresponding to the

origin-destination delay. The edge weight for a network, a_{ij}, is the median departure delay of flights traveling from the origin airport i to the destination airport j within hour h. We collect all such a_{ij} weights into the adjacency matrix $A = [a_{ij}]$ for the delay network. Since the delays change with time, the edge weights of the delay network change as well. Each day d is represented by a time series of 24 weighted and directed networks $(G_{d,1}, G_{d,2}, \ldots, G_{d,24})$, where each of the delay network graphs $G_{d,h}$ has its unique adjacency matrix $A_{d,h}$.

We then construct a feature vector $\mathbf{f}_d$ of network delays for a given day of operations d. One approach is to stack all the edge weights sequentially for each of the 24 networks that correspond to a day [38]. This approach has two limitations. First, stacking all the edge weights completely disregards the network structure, ignoring any possible airport-to-airport interactions embodied by the existence of an edge within the graph. This renders the feature vectors as well as the identified clusters to not be truly representative of the system-wide network impacts that we wish to capture. Second, the feature vector may be very high dimensional (N^2, for a graph with N nodes), causing the clustering algorithms to be computationally expensive. By contrast, our approach uses *weighted hub and authority scores* to incorporate network effects, and is scalable.

Hub and authority scores for weighted directed networks were first introduced in the context of Hyperlink Induced Topic Search [50, 51]. The idea was that there are two notions of importance for a node in directed networks: one indicating the strength (weights) of edges pointing into a node (authority score), and the other for the strength of edges that are pointing out of the node (hub score). Specifically, a node with a high hub score has a lot of edges with large weights pointing out of it. The hub score of a node also increases if these edges point to nodes that are important and have a lot of other inbound edges with large weight values. A node with a lot of inbound edges from nodes with high hub scores has a high authority score. This mutually reinforcing definitions for hub and authority scores is primarily based on node connectivity, and is a very representative metric for the role of a node in the network [50, 51]. In terms of air traffic networks, having significant outbound delays to major airports will increase the hub score of an airport and having significant inbound delays from major airports will increase the authority score of an airport. These eigencentrality-based scores can be computed as follows [50, 51]:

$$A^T A\mathbf{a} = \lambda_{max}\mathbf{a}, \tag{3}$$

$$AA^T\mathbf{h} = \lambda_{max}\mathbf{h}, \tag{4}$$

where $\mathbf{h}$ and $\mathbf{a}$ indicate hub and authority score vectors, respectively. We construct $\begin{pmatrix}\mathbf{h}\\ \mathbf{a}\end{pmatrix}$ as a feature vector for our air traffic delay networks, thereby incorporating network effects in a compact representation using only $2N$ entries. However, since $\mathbf{h}$ and $\mathbf{a}$ are arbitrarily normalized eigenvectors, two networks whose weights differ by a scalar constant would have the same feature vector. In other words, delay networks during high congestion periods and periods of low traffic movements (e.g. late at night) could have the same feature vector if the relative proportion between the flights delays on different routes is the same. To address this issue, we use a *weighted* hub and authority feature vector, where the total system delay of each hour is used to re-scale $\mathbf{h}$ and $\mathbf{a}$.

Thus, for a day d defined by $(G_{d,1}, G_{d,2}, \ldots, G_{d,24})$, the feature vector f_d^{delay} is given by:

$$\mathbf{f}_d^{\text{delay}} = \begin{bmatrix} \alpha_1 \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{a}_1 \end{pmatrix} \\ \vdots \\ \alpha_{24} \begin{pmatrix} \mathbf{h}_{24} \\ \mathbf{a}_{24} \end{pmatrix} \end{bmatrix}, \tag{5}$$

where $\alpha_h = \sum_{ij} a_{ij}^h$ is the total delay at the h^{th} hour and $A^h = [a_{ij}^h]$ is the corresponding adjacency matrix.

We define cancellation networks in a similar manner as delay networks, with the only difference being that the edge weights are the number of canceled flights. The number of flight cancellations is typically small compared to the number of total scheduled flights. Furthermore, we note that the reasons for cancellations range from operational (e.g. aircraft maintenance issues, crew scheduling mishaps) to system-wide irregular operations resulting in excessive delays. Not all cancellations are reflective of system-wide capacity imbalances, and the small number of cancellations every hour would result in undesired sensitivity of the clusters to exogenous factors. For these reasons, in order to achieve a more robust clustering of cancellation networks, we ignore hourly variations and construct only one cancellation network to represent an entire day of operations. Thus, we construct the cancellation feature vector for each day d as:

$$\mathbf{f}_d^{\text{canc}} = \sum_{ij} a_{ij} \times \begin{pmatrix} \mathbf{h} \\ \mathbf{a} \end{pmatrix}. \tag{6}$$

3.5. Outcome of the clustering procedure

The NAS impacts in terms of GDPs, delays, and cancellations for a given day of operations d are reflected in the feature vectors $\mathbf{f}_d^{\text{GDP}}$, $\mathbf{f}_d^{\text{delay}}$ and $\mathbf{f}_d^{\text{canc}}$, respectively. We use these feature vectors as inputs into the k-means clustering in order to identify days with similar GDP, delay, and cancellation networks. Consequently, each day used for the training would have three labels associated with it: One describing the type of capacity impacts in terms of system-wide GDP issuances; one describing the propagation of delays within the network; and one for the distribution of flight cancellations within the network. Any new day in a testing set can now be mapped to the closest cluster centroid with regards to its GDP, delay, and cancellation networks.

The clusters of pertinent GDP impacts that we identify are primarily used as performance baselines for our benchmarking framework, particularly from the perspective of airlines. Since the effects of issued GDPs are targeted towards airports, this sets up the system constraints within which the airlines operate. We present case studies and discussions of using our benchmarking framework from both the airline and the system operator (e.g., the Federal Aviation Administration or FAA) perspective in Sections 5 and 6, respectively.

3.6. Discussion

The reference data sets that are used to generate clusters and baseline metrics can correspond to any time period, or could be restricted to an airline. They may also be scoped to include only flights operating through a particular subset of airports (e.g., only hubs). In practice, all baselines could be re-computed periodically to include more recent data; such re-computations can be performed by appending new days of operations to the reference data sets. Furthermore, one could examine hourly (as opposed to daily) clusters of GDP, delay, and cancellation networks as well.

We reiterate that our focus is not the clustering of nodes within a graph, i.e., we are not performing node-based community detection; we are instead taking entire graphs to be the object that we are trying to cluster with other graphs based on their topology (e.g., connectivity patterns, edge weights, etc.) Furthermore, different airspace monitoring applications may require different performance measures and feature vectors (e.g., [25, 38]).

4. Identification of baselines for benchmarking

In this section, we apply k-means clustering algorithm using the feature vectors described in Section 3 to identify representative, or characteristic type-of-days.

4.1. Description of data sets

Data on airport capacity impacts was obtained from the FAA Advisory Database, which publishes information on the reduction in airport capacity due to GDPs [52]. We merge multiple notifications for the same GDP, including revisions, ground stops, extensions, and early cancellations, to obtain an hourly time series of the reduced capacity at a specific airport. We restricted our analysis to the FAA's Core 30 airports. We obtained nominal capacities from the FAA Airport Capacity Profiles report in order to compute fractional reductions in AAR [53]. GDP data used in this analysis was available for the period spanning 2014 through 2018. Flight schedules, delays, and cancellation information were obtained for 2014-2016 from public data sources such as the FAA ASPM [40], the DOT's BTS database [6], and other third-party providers.

The data set included 20,012 GDP advisory notifications specified for the FAA Core 30 airports, as well as records of 22.9 million domestic US flights operated by 55 airlines, through 426 airports, and on 6,600 unique origin-destination (OD) pairs. Approximately 500,000 (2.1% of the total) of the scheduled flights were cancelled. We removed 5,000 diverted flights (0.2% of the total) from our baseline calculations and subsequent benchmarking. To improve robustness and eliminate outliers, the clustering was restricted to the top 1,000 OD pairs based on the average traffic. All of these OD pairs had at least 6 flights per day (defined as between 0900Z and 0859Z) on average per day. Data from 2014-2015 was used as the reference.

4.2. Selecting the number of clusters

An input parameter for k-means clustering is the number of clusters k that the user wants the data observations to be partitioned into. The choice of this k parameter depends

on the application, and can be evaluated using a number of different criteria [54]. We use the within cluster sum of squares (WCSS), the silhouette score, and the Davies-Bouldin index for this purpose. However, we note that the data does not naturally form clusters, and there is no clear choice for k. Nevertheless, the underlying delay data is non-uniform [55], and clustering is still useful to identify interpretable and representative baselines.

The choice of k is now guided by two competing factors: lower k improves interpretability, while higher k captures more variability and is more accurate. We select values of k such that we achieve a balance between the number of clusters and the within-cluster population, while taking into account input from airline subject matter experts. For our analysis, we chose $k = 8$, 6, and 5, for the number of GDP, delay, and cancellation type-of-day clusters, respectively. A more detailed discussion of the selection of the number of clusters can be found in the Supplementary Material.

4.3. GDP type-of-day

Eight GDP type-of-days are identified through k-means clustering using the feature vector $\mathbf{f}^{\text{GDP}}$. Qualitative descriptions and frequencies of occurrence of these clusters are presented in Table 1. We visualize the centroid, or the representative data point, of each GDP type-of-day cluster in Figure 2.

Day type	**Qualitative description**	**Frequency (2014-15)**
Atlanta (`ATL`)	High GDP intensity at ATL; moderate to low GDP activity elsewhere.	6.3%
Chicago (`CHI`)	Very high GDP intensity at ORD and MDW; moderate to low GDP activity elsewhere.	7.8%
Medium Northeast (`MedNE`)	Medium-high GDP intensities in the Northeast (NYC airports, PHL and BOS). Moderate to low GDP activity elsewhere.	15.3%
Low NAS (`LowNAS`)	Low GDP intensities nationwide.	48.1%
Miami, Northeast (`Miami_NE`)	High GDP intensity in MIA, and medium-high GDP intensities in the NYC/PHL area.	5.1%
Northeast (`NE`)	Very high GDP intensities in the NYC, PHL, and Washington DC airports; medium-high at BOS and Chicago. Moderate to low GDPs elsewhere.	2.2%
New York City, Philadelphia (`NYC_PHL`)	Very high GDP intensities at the NYC airports (LGA, JFK and EWR) and PHL; medium-high GDP activity at BOS. Moderate to low GDPs elsewhere.	5.1%
San Francisco (`SFO`)	High GDP intensity in SFO; moderate to low GDP activity elsewhere.	10.1%

Table 1: Characteristic GDP type-of-day identified through clustering.

Each centroid in Figure 2 for the GDP type-of-days is visualized as the average of the corresponding time-series of GDP intensities at the FAA Core 30 airports; the full time-series is a sequence of 24 node-weighted networks. A larger circle indicates a higher average GDP impact at the airport. For ease of visualization, we use one image to represent the

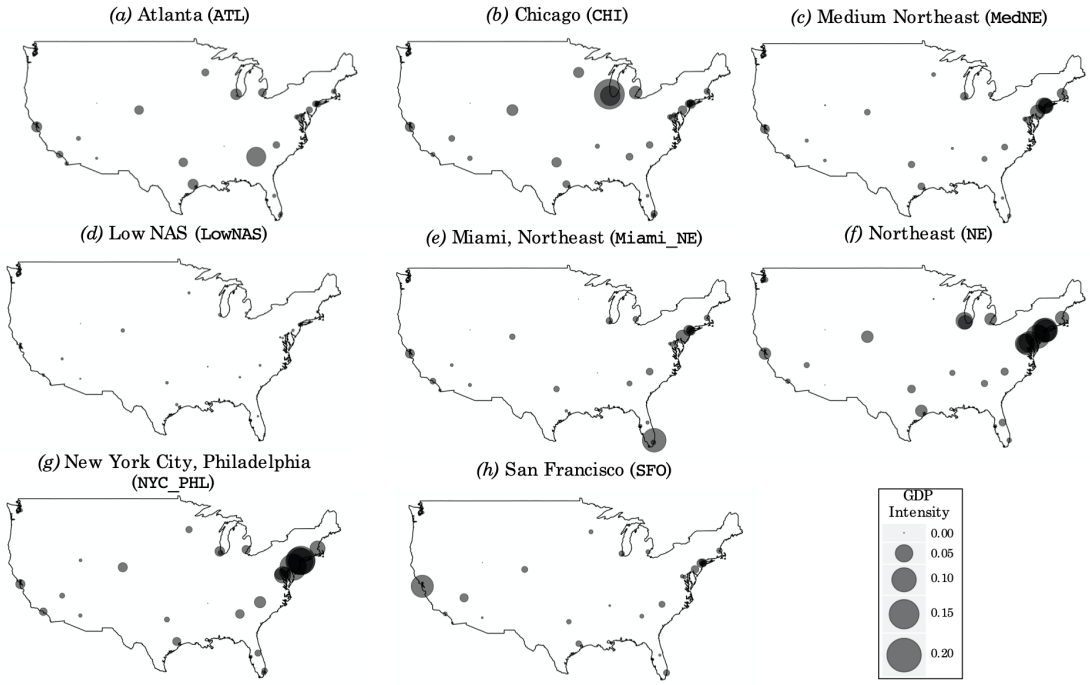

Figure 2: Visualization of the centroids for all GDP type-of-days. The size of the circle reflects the GDP intensity at an airport.

entire time series, although the clustering considers temporal variations. We note that peak GDP intensities and impacts may occur at different times for different centroids.

As expected, `LowNAS` GDP type-of-days occur most frequently: 48% of days in 2014-2015 are classified as having low GDP intensities throughout the NAS. However, the GDP types affecting the Northeast, i.e. `NE`, `MedNE`, `Miami_NE`, and `NYC_PHL`, together account for a significant fraction of days (28%), and have higher GDP intensities and impacts. We note the multitude of distinct GDP type-of-days that impact the Northeast and Mid-Atlantic region of the US, reflecting both the frequency and diversity of GDP activity at airports in the East Coast.

4.4. Delay type-of-day

We present the six different delay type-of-days identified via network-centric clustering. We list each delay type-of-day, along with its qualitative description and the frequency of occurrence, in Table 2. We visualize the centroid for each cluster (Figure 3) as a network, where the edge weight is the median OD pair delay for all flights on that OD pair during the days of operations belonging to that particular cluster. We note that even though the edges are directed, for ease of visualization, the edges are depicted as undirected and weighted by the maximum of the two directions. Although our centroid visualizations shows a single delay network, recall that it is actually representative of a 24-hour time series, reflecting the temporal evolution of delays. Analogous to the case with the GDP centroid visualizations, in order to maintain simplicity and visual interpretability, we project out the temporal

Day type	Qualitative description	Frequency (2014-15)
Atlanta (ATL)	Delays centered around ATL, with a mean departure delay of 19 min. 71% of flights arrive within 15 min of their scheduled arrival times. Largest avg. arrival delays: ATL, MEM, ORD, IAD, EWR.	3.6%
Chicago (CHI)	Delays centered around Chicago, with a mean departure delay of 17 min. 74% of flights arrive within 15 min of their scheduled arrival times. Largest avg. arrival delays: ORD, MDW, MEM, EWR, IAD.	8.1%
High NAS (HighNAS)	Delays widespread and high, with a mean departure delay of 29 min. Only 61% of flights arrive within 15 min of their scheduled arrival times. Largest avg. arrival delays: EWR, DFW, IAH, LGA, ORD.	2.9%
Low NAS (LowNAS)	Delays low nationwide, with a mean departure delay of only 9 min. Over 85% of flights arrive within 15 min of their scheduled arrival times. Largest avg. arrival delays: SFO, EWR, JFK, ORD, IAD.	49.3%
Northeast (NE)	Delays centered in the Northeast, with a mean departure delay of 18 min. 72% of flights arrive within 15 min of their scheduled arrival times. Largest avg. arrival delays: LGA, EWR, JFK, PHL, BOS.	11.0%
West Coast, Medium NAS (WC_MedNAS)	Delays centered in the West coast & moderate elsewhere; mean departure delay of 14 min. 77% of flights arrive within 15 min of their scheduled arrival times. Largest avg. arrival delays: SFO, DFW, DEN, LAX, LAS.	25.2%

Table 2: Characteristic delay type-of-day identified through clustering.

dimensions, even though they are still relevant in the actual clustering process.

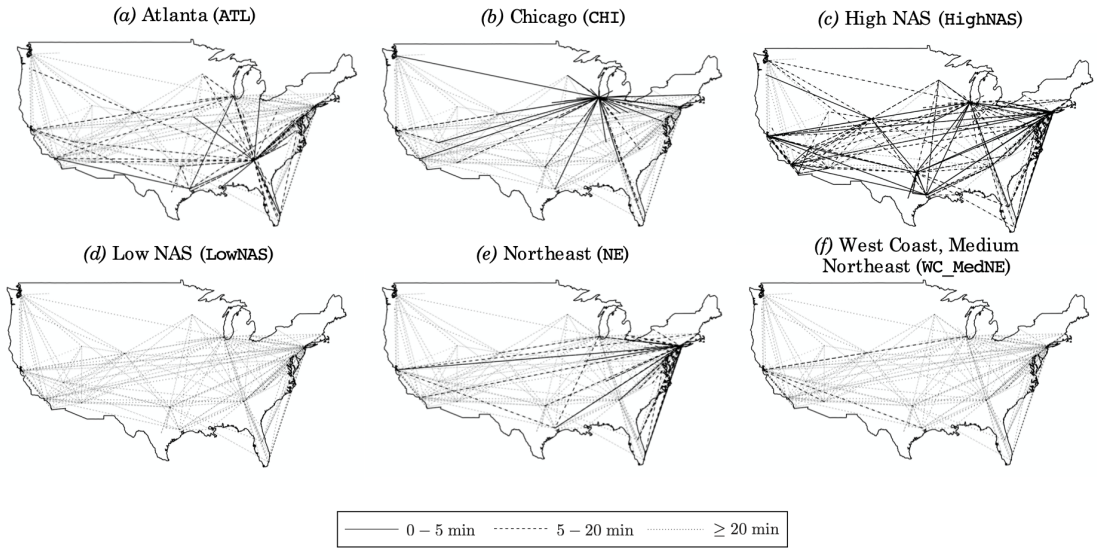

Figure 3: Visualization of the centroids for all delay type-of-days.

Similar to the case with GDPs, the most prevalent type-of-day for delays is LowNAS, with average departure delays of 9 minutes, and with over 85% of flights arriving within fifteen minutes of their scheduled arrival times (i.e. the A15 percentage metric). We can also isolated the airports with the largest arrival delays for each delay type-of-day to identify

airports that are typically delayed when a particular delay type-of-day occurs. Certain airport nodes begin to stand out in terms of their participation in many delay type-of-days. For example, Chicago O'Hare International Airport (ORD), a major dual hub that also shares its terminal airspace with the proximate airport Chicago-Midway, has significant delays in `ATL`, `CHI`, and most prominently, `HighNAS` delay type-of-days.

4.5. Cancellation type-of-day

To complete our presentation and analysis of resultant cluster centroids, we qualitatively discuss the five type-of-days for cancellations in Table 3, and present the visualizations of the centroids in Figure 4. We constructed these simplified visualizations analogous to those for delay type-of-days, with the key difference being that the edge weights now correspond to the number of cancellations observed on an OD pair.

Day type	**Qualitative description**	**Frequency (2014-15)**
Atlanta (`ATL`)	Very heavy cancellations on routes to/from ATL (75% of all scheduled arrivals at ATL cancelled). Systemwide cancellation rate of 18%.	0.4%
Chicago, Northeast (`CHI_NE`)	Heavy cancellations in Chicago and parts of the Northeast (37% of all scheduled arrivals at LGA and ORD, 35% of scheduled arrivals at EWR, and 32% of scheduled arrivals at MDW, cancelled). Systemwide cancellation rate of 14%.	2.3%
Medium Chicago, Medium Northeast (`Med_CHI_NE`)	Moderate cancellations of flights to/from Chicago and the Northeast (14% of all scheduled arrivals at LGA, 13% at EWR, 12% at ORD, and 10% of scheduled arrivals at PHL, cancelled). Systemwide cancellation rate of 6%.	13.2%
Low NAS (`LowNAS`)	Most frequently-occurring day, with a systemwide 1.3% cancellation rate. EWR has the largest (2.5%) rate of cancelled arrivals.	83.5%
High Northeast (`HighNE`)	Very high cancellations of flights to/from the Northeast (83% of all scheduled arrivals at LGA, 79% at EWR, 66% at PHL, 64% at BOS, 56% at DCA, and 51% of scheduled arrivals at IAD and JFK, cancelled). Systemwide cancellation rate of 25%.	0.5%

Table 3: Cancellation type-of-day centroids

Due to their operational rarity, cancellation type-of-days are more difficult to distinguish compared to delay and GDP type-of-days. This rarity is exemplified in the fact that 83.5% of operational days within the 2014 to 2015 time frame are classified as `LowNAS` cancellation type-of-day, with an overall system-wide cancellation rate of 1.3%. Interestingly, in the `LowNAS` cluster, EWR had the highest rate of canceled arrivals, at 2.5%. Some of the cancellation clusters match those seen in delay and GDP type-of-days in terms of geographic distribution, centering on major hub airports in Atlanta, Chicago, and the Northeastern region of the US.

5. Airline performance benchmarking

We present an on-time performance benchmarking application from the perspective of an airline. This use case was based on the implementation of these methods at a major US

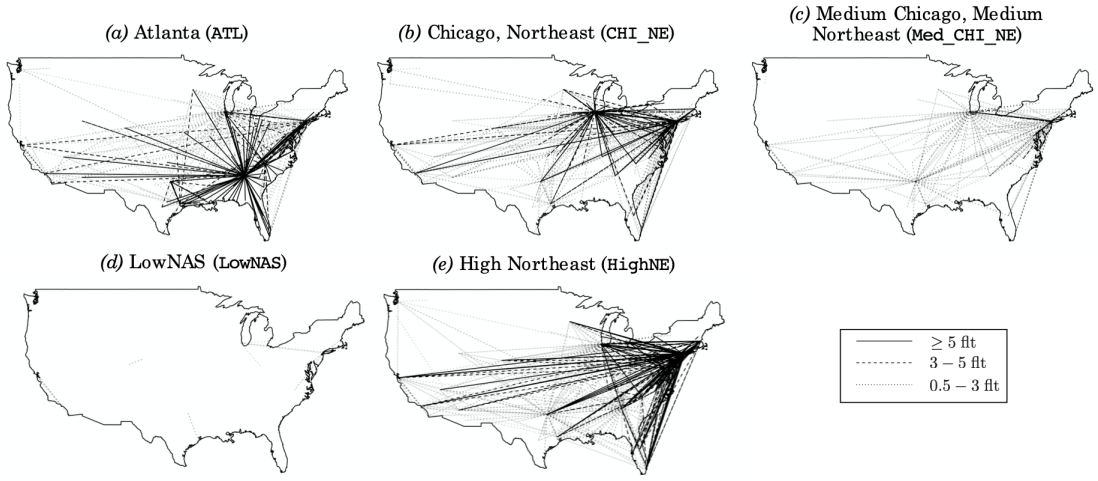

Figure 4: Visualization of the centroid for all cancellation type-of-days.

airline for post-hoc analysis of on-time performance. As illustrated in Figure 5, an airline can select from a variety of on-time performance metrics such as average departure and arrival delays, number of cancellations, D0/D15/A0/A14 percentages, block time adherence, and ground turn times, each at the airline-, airport-, or system-level. We consider the average departure delay (henceforth simply referred to as *delays*) per flight and the cancellation percentage for illustrative purposes.

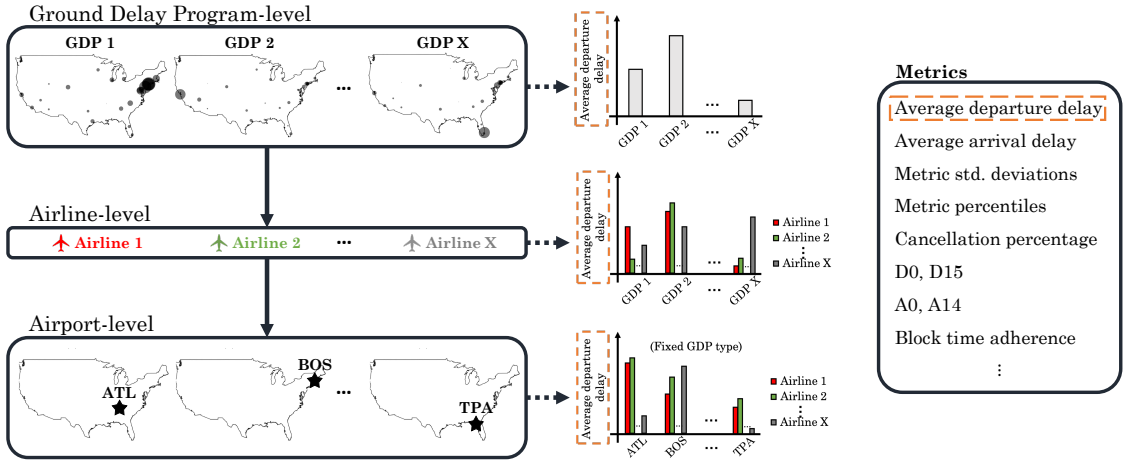

Figure 5: Performance baselines can be customized by airline, airport, or system-wide.Mainline and regional carrier breakdowns are not shown here for simplicity, but are considered in Section 5.3.

5.1. Confidence intervals for performance metrics

Let $\{x_d^{(i)}\}_{i=1}^{i=n(\mathcal{P})}$ represent the set of observations of performance metric d. Every observation $x_d^{(i)} \in \mathbb{R}$ is recorded by aircraft $i = 1, \ldots, n(\mathcal{P})$, where $n(\mathcal{P})$ is the $\mathcal{P}$-conditioned

sample population. $\mathcal{P}$ could represent any particular combination of airline, mainline or regional, airports, and cluster identity. We denote the sample mean and standard deviation of $\{x_d^{(i)}\}_{i=1}^{i=n(\mathcal{P})}$ by $\bar{x}_d^{n(\mathcal{P})}$ and $s_d^{n(\mathcal{P})}$, respectively. The two-sided confidence interval (CI) of level α is denoted by

$$CI\,(d, n(\mathcal{P})) = \bar{x}_d^{n(\mathcal{P})} \pm z_{\alpha/2} \frac{s_d^{n(\mathcal{P})}}{\sqrt{n(\mathcal{P})}} \tag{7}$$

For $\alpha = 0.01$, $z_{\alpha/2} \approx 2.576$. We only compute the CI for delays, and not for cancellations. The percentage of cancellations is a *discrete* performance metric with a single value (and no CI), since a flight is either cancelled or it is not.

The CI can be computed for the mean performance of the baselines, as well as for any particular day. When benchmarking an individual day (Section 5.3), we can assert that the average delay on the individual day deviates in a significant way from the mean cluster delay if the two CIs do not overlap. Lastly, we note that (7) is only meaningful when $n(\mathcal{P})$ is large; Tables 5 and 6 in the Supplementary Information show that this is a valid assumption.

5.2. On-time performance baselines

Figure 6 shows the baseline metrics computed using the 2014-15 data set. The baseline metrics are evaluated as an average over all flights on days classified as belonging to a particular GDP type. From the system-wide baseline values (Figure 6*(a)*), we see that `NE` and `NYC_PHL` GDP type-of-days result in the highest departure delays and cancellations, while `LowNAS` experiences the lowest delays and cancellations.

A1, A2, and A3 correspond to three major US airlines. We observe from Figure 6*(b)* that the baseline delays and cancellations for airline A2 are lower than those for A1 and A3, across all GDP types-of-day. Furthermore, the GDP type-of-days that most impact system performance (Figure 6(a)) may not necessarily be the most severe for a particular airline. For example, `NE` GDP types-of-day affect the performance of A2 and A3 less severely than A1. Not surprisingly, airlines differ in their baseline performance even in the absence of airport capacity reductions (`LowNAS` GDP type-of-day), due to differences in operational practice. Although baseline performance differs by airline for most GDP types-of-day, the `ATL` GDP type results in similar impacts for all three major airlines, even though it is the prominent hub airport of one of them.

We analyze the mainline and regional carrier performance for airline A1 in Figure 6*(c)*, and consider specific airports in Figure 6*(d)*. Although all airlines have high delays at EWR, JFK, LGA, and PHL during `NYC_PHL` GDP type-of-days, airline-specific impacts are also evident: the delays are the highest for A1 at LGA, and for A2 at EWR. The propagation effects of GDPs also vary by airline. For example, by controlling for the `NYC_PHL` GDP type-of-day (Figure 6*(d)*), the propagation of departure delays for airline A1 from the NYC-PHL airports to ATL becomes much more apparent.

5.3. Benchmarking case study from the airline perspective

We present a case study of benchmarking operational performance from an airline's perspective. We consider two days, July 1, 2016 and February 21, 2015, which were both

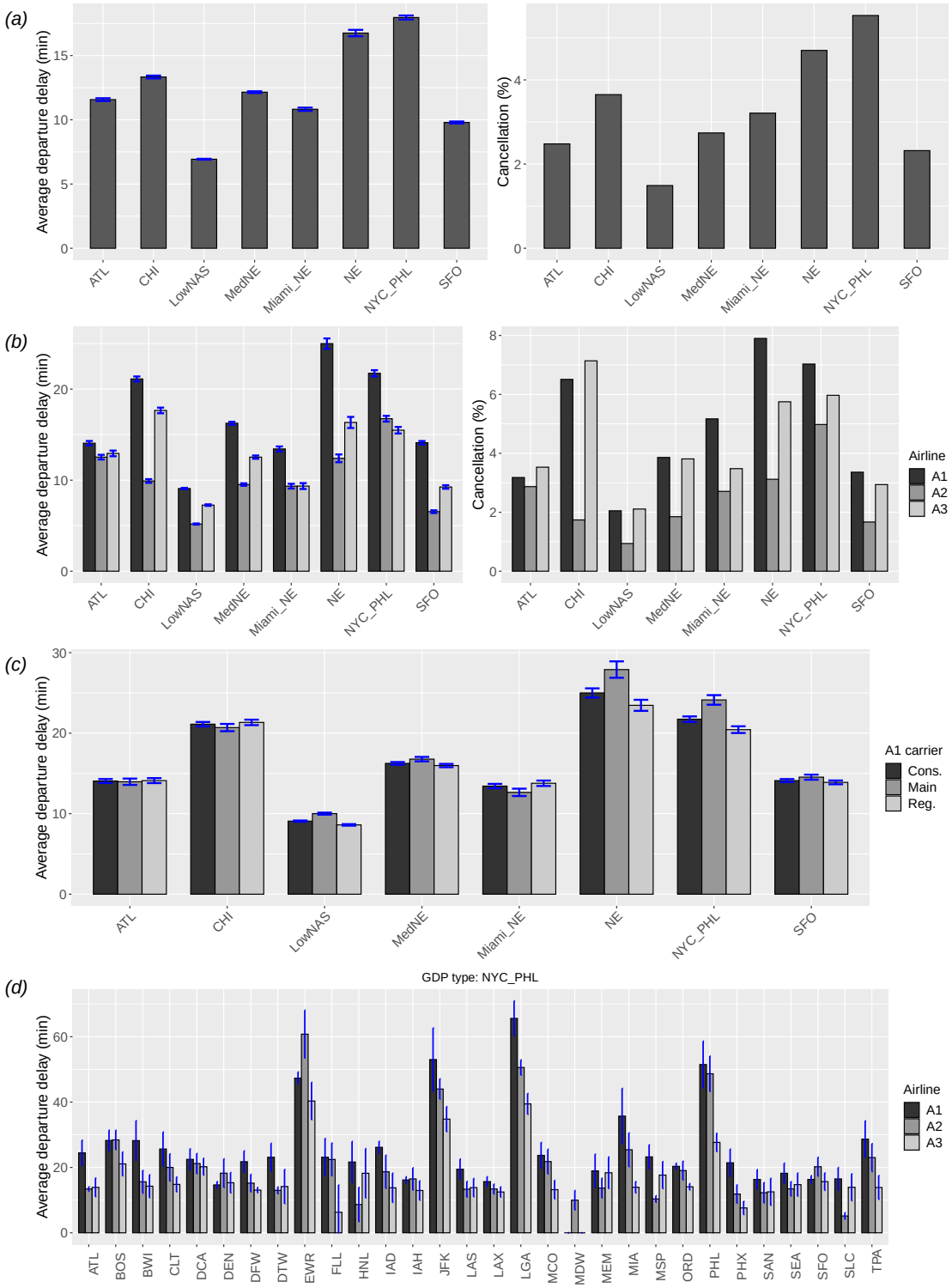

Figure 6: Performance baseline plots generated with average departure delays in minutes; baseline performance *(a)* across GDP types; *(b)* across three major US airlines for each GDP type; *(c)* across the consolidated fleet, mainline, and regional carriers of airline A1 for each GDP type; *(d)* across three major US airlines at the top 30 US airports, given a specific GDP type. The error bars indicate a 99% two-sided confidence interval; please see Section 5.1 for a discussion regarding the confidence intervals.

classified as being NYC_PHL GDP type-of-days. We then benchmark the on-time performance of the three airlines A1, A2, and A3 with respect to the appropriate baselines for that GDP type-of-day. The left-hand panel of Figure 7 summarizes July 1, 2016 (see [56] for more details), a day when three distinct convective systems moved through the western, Midwestern, and East Coast regions of the US. The weather patterns in the West and Midwest persisted through the entire day, whereas the East Coast system moved offshore at around 1600Z. These weather systems resulted in intense GDPs in the Mid-Atlantic region, and less-intense GDP activity in the Los Angeles area. By contrast, the weather patterns on February 21, 2015 (Figure 7, right panel) consisted primarily of a large nor'easter-type storm that moved steadily from west to east, persisting through the entire day, resulting in GDPs in the Northeast. Our clustering framework helps compare these two complex weather phenomenon and their aviation impacts precisely: if the two days belong to the same GDP cluster, we can assume the weather impacts to be similar. However, the resultant delay networks for the two days are significantly different: July 1, 2016 was a NE delay type-of-day, while February 21, 2015 was a LowNAS delay type-of-day.

The average delay (and associated CI) for all airlines combined was lower on February 21, 2015 than the NYC_PHL GDP type-of-day baseline values. This trend was reversed for July 1, 2016. On July 1, 2016, A2 and A3 performed poorly compared to their baselines, whereas A1 saw a statistically significant improvement over its baseline. This analysis illustrates the importance of conditioning on capacity impacts while benchmarking: without doing so, one would rate A1 and A2's delay performance to be similar. However, A1 actually performed better than its NYC_PHL GDP type-of-day baseline, while A2 performed worse. In other words, A1 not only outperformed its typical performance for NYC_PHL GDP type-of-days, but it also outperformed A2, even though A2's delay performance is typically better on similar GDP type-of-days.

Figure 7*(e)* compares the performance of mainline and regional carriers. On July 1, 2016, even though A1 saw a reduction in delays, most of the reduction could be attributed to their regional carrier. On February 21, 2015, both A1 and A3 saw reduced delays for their consolidated fleet; the improvement for A1 was its regional operations, while it was the mainline carrier for A3. The cancellation percentages (Figure 8) provide a more complete picture of the two days described in Figure 7. On February 21, 2015, A1 and A3 saw a statistically significant decrease in delays compared to their baselines. However, A1 cancelled significantly more flights (79 percentage points higher) compared to A3. We see that A1 cancelled heavily within its regional and mainline operations, whereas A3 selectively cancelled mainline operations. In fact, the cancellation rate in A3's regional operations were below average. Of course, different external factors (e.g., airline networks, schedules, etc.) could have been more favorable for one airline, even if they both experienced the same GDP impacts. Our analysis helps identify such days for closer inspection. The relative benefits and costs of cancelling mainline vs. regional operations for such a GDP type-of-day is a question for future research.

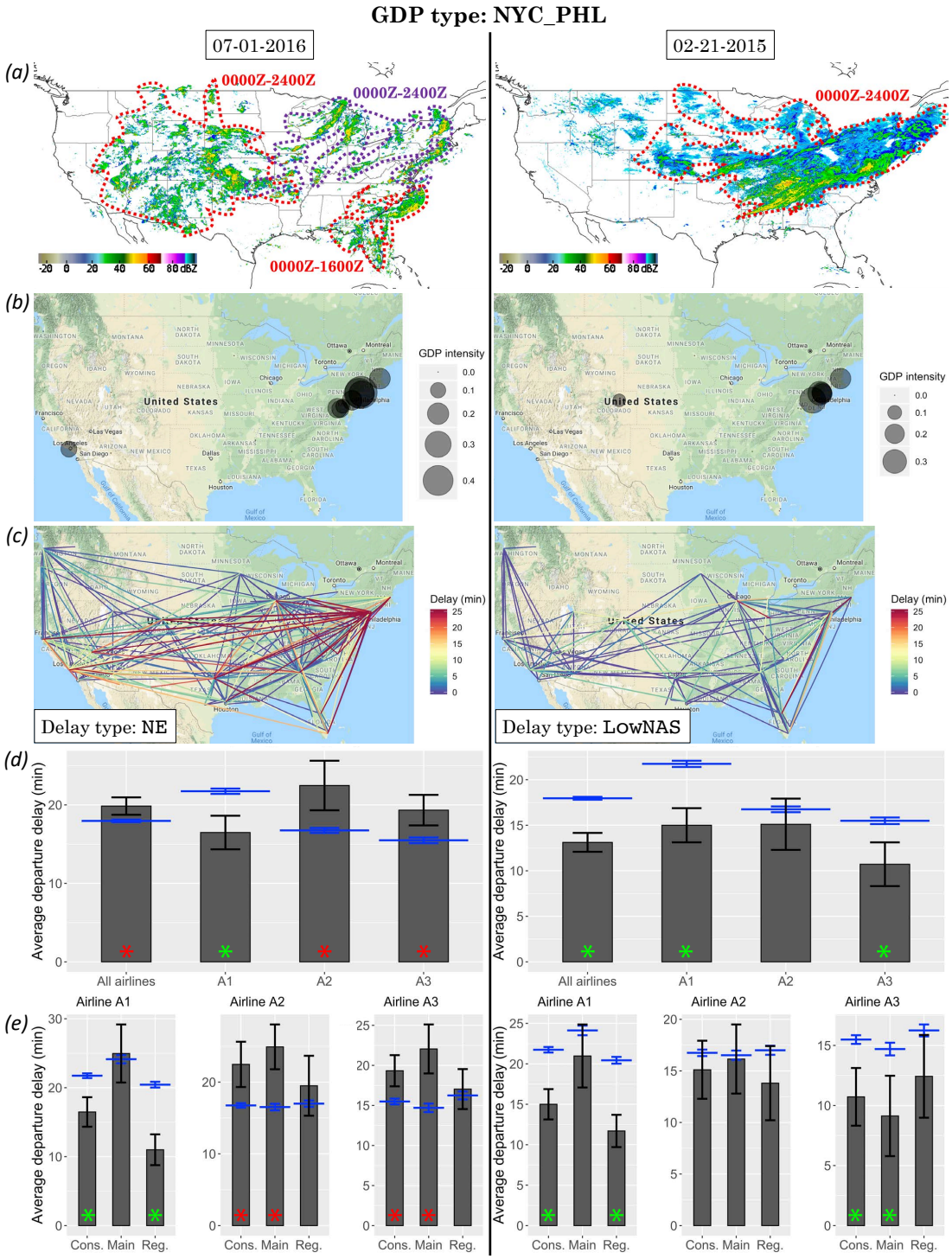

Figure 7: Example of a benchmarking panel generated for two NYC_PHL GDP type-of-days. *(a)* Composite weather radar returns from 0000Z to 2400Z. *(b)* Average GDP intensities by airport. *(c)* Delay networks. *(d)* Benchmarks of on-time performance across all airlines and three major US airlines. *(e)* Benchmarks of on-time performance across the consolidated fleets, mainline, and regional carriers of three major US airlines. Note that the horizontal blue lines in *(d)* and *(e)* represent baseline values. All the bars denote a 99% two-sided confidence interval for the mean. The colored asterisk indicates a statistically significant difference between the observed mean, and the benchmark mean; the color indicates if the observed delays on the particular day were higher or lower.

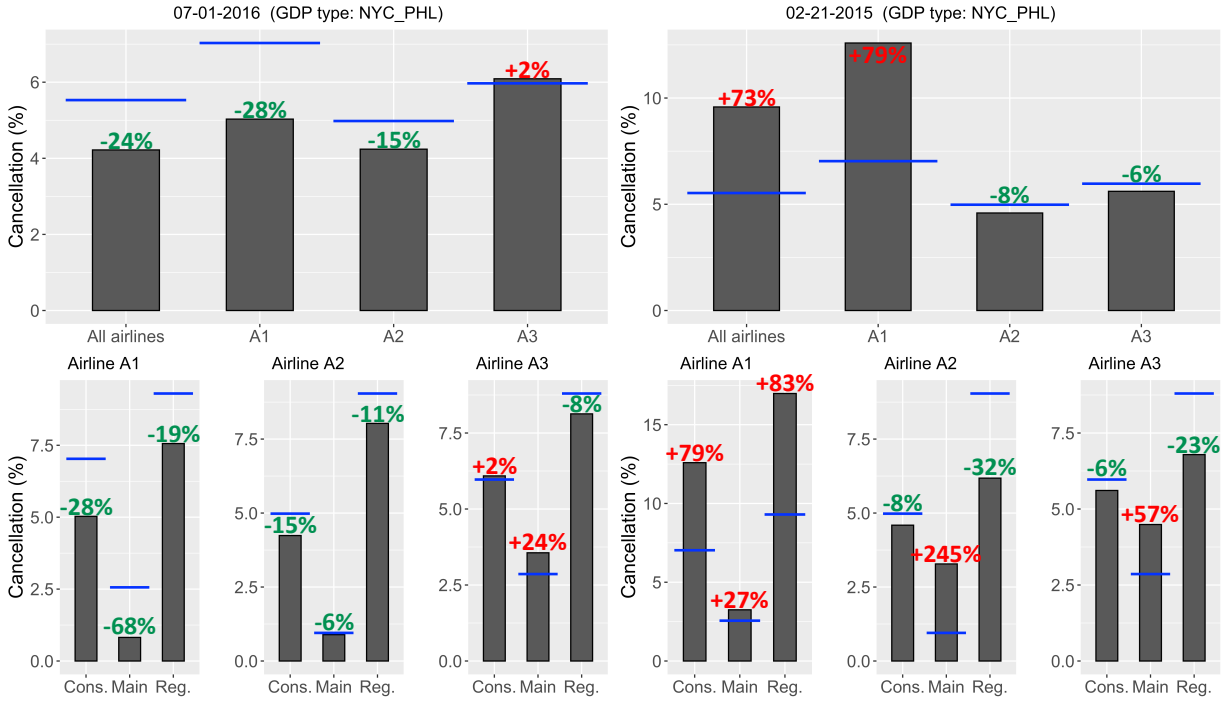

Figure 8: Cancellation percentage benchmarks for the two `NYC_PHL` GDP type-of-days considered in Figure 7. The horizontal blue lines represent baseline values.

6. Benchmarking for system operators

We present two applications of our clustering analysis for a system operator (the FAA, in the US): the analysis of seasonal and yearly trends in the occurrences of GDPs, and an evaluation of the correlation between GDPs and system impacts (delays and cancellations).

6.1. Frequency of occurrence of GDP type-of-days

Figure 9 shows the frequency of occurrence of different GDP type-of-days for each year in 2014-2018. First, we observe that nearly half the days in a year exhibit a significant amount of GDP activity. We also note that the `Miami_NE` GDP type-of-day was seen only in 2014, and very rarely thereafter. The increase in the occurrence of `LowNAS` days approximately compensates for the decrease in `Miami_NE` GDP type-of-days, indicating a change in terms of the timing, location, and intensity of GDPs.

The year 2017 saw the highest frequencies of occurrence of `MedNE`, `NYC_PHL`, and `SFO` GDP types-of-days. A possible explanation for the observed increase in `Med_NE` and `NYC_PHL` GDP type-of-days in 2017 is the removal of slot controls at EWR in October 2016, and the subsequent congestion that resulted due to the increase in demand over the next year [57].

GDP occurrences also show seasonal trends (Figure 10), since weather is a major cause of capacity reductions. The months of January through March experience snowstorms in the Midwest and the Northeast, whereas summers exhibit increased thunderstorm activity near Atlanta and Chicago. As a result, we observe more frequent `ATL` and `CHI` GDP type-of-days

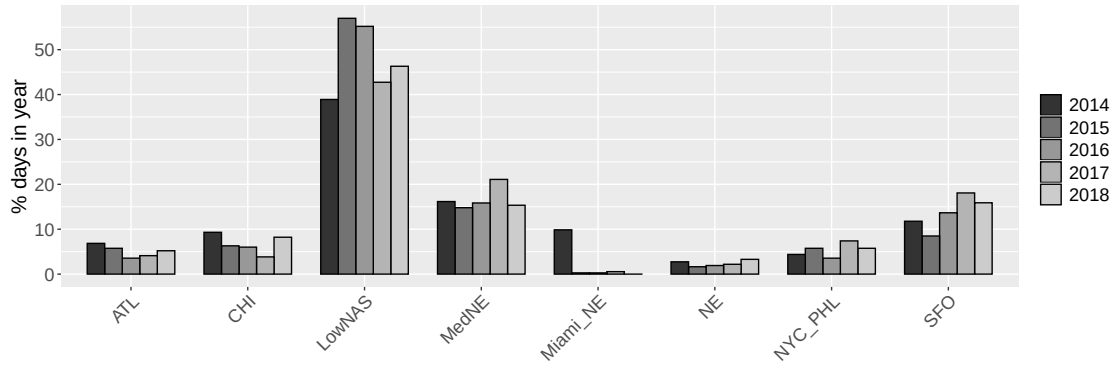

Figure 9: Frequency of occurrence of each GDP type-of-day in 2014-2018.

in the months of May and June. Somewhat surprisingly, we find that `LowNAS` GDP type-of-days are more common in winter than summer. A possible explanation is that proactive flight cancellations in advance of snowstorms reduce demand sufficiently, and eliminate the need for GDPs.

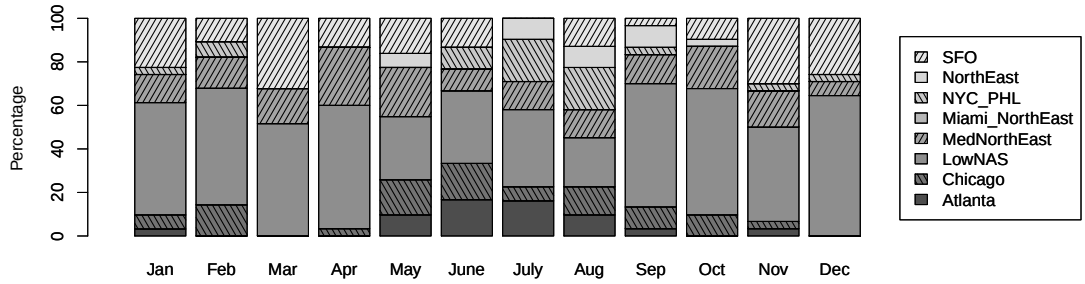

Figure 10: Frequency of monthly occurrence of different GDP type-of-days from 2014 through 2018.

6.2. Impact of GDPs on system-wide delays and cancellations

GDP occurrences are correlated with flight delays and cancellations. To analyze this correlation at the system-wide level, we consider the triplet (G, D, C), where G, D, and C represent the GDP, delay, and cancellation type-of-day, respectively. This triplet takes one of $6 \times 8 \times 5 = 240$ values. An example of a valid triplet would be (`LowNAS`, `LowNAS`, `LowNAS`), indicative of a day with little GDP activity, and low delays and cancellations.

6.2.1. Frequency of occurrence of (G, D, C) triplets

Even though the clustering and subsequent classification of GDP, delay, and cancellation type-of-days are conducted independently, it is reasonable to expect that they are correlated. Table 4 lists the ten most common triplet combinations, which together account for 75% of the days in 2014-2016. As expected from earlier results in Section 3, the most frequently-occurring triplet is (`LowNAS`, `LowNAS`, `LowNAS`), which accounts for 37% of the days. Furthermore, five out of the top ten most frequent triplets involve GDPs, delays, or both, on the West Coast, in conjunction with a `LowNAS` cancellation type-of-day. The

only triplet combination in the top ten that is not of LowNAS cancellation type is the triplet (NYC_PHL, NE, Med_CHI_NE), which corresponded to approximately 2% of all days.

GDP type	Delay type	Cancellation type	# of days (frequency)
LowNAS	LowNAS	LowNAS	402 (36.7%)
LowNAS	WC_MedNAS	LowNAS	101 (9.2%)
SFO	LowNAS	LowNAS	62 (5.7%)
MedNE	LowNAS	LowNAS	60 (5.5%)
SFO	WC_MedNAS	LowNAS	51 (4.6%)
MedNE	WC_MedNAS	LowNAS	42 (3.9%)
MedNE	NE	LowNAS	32 (2.9%)
CHI	CHI	LowNAS	26 (2.4%)
ATL	WC_MedNAS	LowNAS	23 (2.1%)
NYC_PHL	NE	Med_CHI_NE	21 (1.9%)

Table 4: Ten most frequently-occurring triplet combinations in 2014-2016.

Rare (G, D, C) triplets present a way to identify outlying days in terms of operational performance. We present one such rare (G, D, C) triplet here: January 8, 2014 was classified as a (SFO, CHI, Med_CHI_NE) type-of-day; in fact, this was the only day in 2014-2016 that was classified as such. Closer investigation revealed that while this day in itself saw little weather and GDP activity outside of SFO, the days preceding it had experienced bad weather and significant snow accumulation in Chicago and the Northeast. The triplet corresponding to this day reflects the slow recovery process after a major disruption. It is worth noting that only 87 out of 240 possible (G, D, C) triplet labels were assigned at least once in 2014-2016. Most combinations, e.g., (CHI, WC_MedNAS, NE) and (LowNAS, CHI, ATL), were never observed.

6.2.2. Scenario tree analysis using (G, D, C) triplets

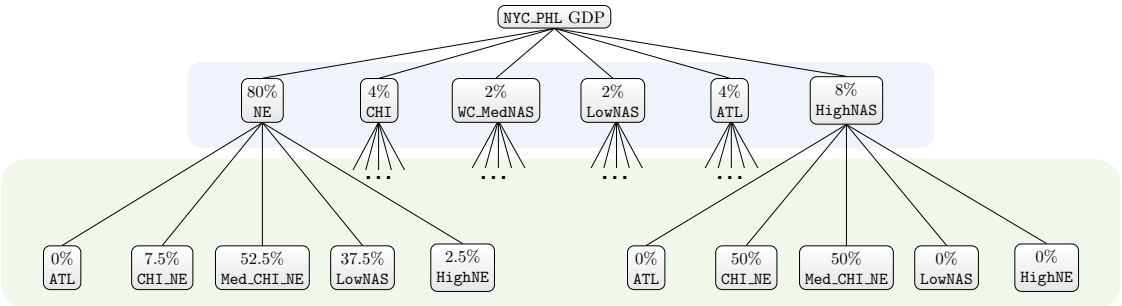

Figure 11: Scenario tree with NYC_PHL GDP type-of-day root; delay type-of-day stems (blue background) and cancellation type-of-day leaves (green background). Not all stems and leaves are shown.

Scenario trees present a natural way to interpret the historical correlations between the GDP, delay, and cancellation types-of-days (i.e., the (G, D, C) triplets). For a given triplet, the root node is the GDP type-of-day; the next row presents the probability for delay type-of-days conditioned on the GDP type, and the leaves present the GDP and delay type-of-day

conditioned probability of occurrence of a cancellation type-of-day. The scenario tree for a NYC_PHL GDP type-of-day is illustrated in Figure 11. We see that most (80%) NYC_PHL GDP type-of-days are NE delay type-of-days, a smaller fraction (8%) are HighNAS delay type-of-days, and it is rare (2%) for a NYC_PHL GDP type-of-day to be classified as a LowNAS delay type-of-day. Furthermore, 37.5% of days classified as NYC_PHL GDP and NE delay type-of-days map to LowNAS cancellation days. By contrast, days associated with the NYC_PHL GDP and HighNAS delay type-of-days will also see flight cancellations, as either CHI_NE or Med_CHI_NE cancellation type-of-days. In other words, similar GDP patterns can result in different cancellation patterns, depending on the magnitude and spatial distribution of delays in the system. Finally, we note that the scenario tree corresponding to a (G, D, C) triplet is not unique. For example, we can reverse the order of delay and cancellation types-of-day; the resulting scenario tree would help understand the correlations between GDP and cancellation patterns.

7. Concluding remarks

We presented a data-driven benchmarking framework that provides objective assessments of airline on-time performance. Using a set of reference airline operations data, we generated baseline metrics via clustering system-wide GDP, delay, and cancellation networks for each day. We used eigencentrality measures for directed graphs, specifically weighted hub and authority scores, to obtain computationally-efficient feature vectors for network clustering. The resultant baselines were used to benchmark the on-time performance for any given day of operations. We illustrated the proposed approach for use-cases of airlines and the FAA. Our framework has been deployed as part of a performance monitoring tool at a major US airline. In future work, we intend to explore the applicability of this approach to benchmarking the performance of other transportation and energy networks.

Acknowledgments

This work was partially supported by NSF CPS Award No. 1739505 and an NSF Graduate Research Fellowship (M. Z. Li). Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the U.S. Government.

[1] Bureau of Transportation Statistics, Traffic Data for U.S Airlines and Foreign Airlines U.S. Flights (2017).

[2] A. Cook, H. A. Blom, F. Lillo, R. N. Mantegna, S. Miccich, D. Rivas, R. Vzquez, M. Zanin, Applying complexity science to air traffic management, Journal of Air Transport Management 42 (2015) 149–158.

[3] M. Ball, C. Barnhart, M. Dresner, M. Hansen, K. Neels, A. Odoni, E. Peterson, L. Sherry, A. Trani, B. Zou, Total delay impact study (2010).

[4] Joint Economic Committee, US Senate, Your Flight has Been Delayed Again: Flight Delays Cost Passengers, Airlines, and the US Economy Billions (2008).

[5] S. Bratu, C. Barnhart, An Analysis of Passenger Delays Using Flight Operations and Passenger Booking Data, Air Traffic Control Quarterly 13 (1) (2005) 1–27.

[6] Bureau of Transportation Statistics, Airline On-Time Statistics and Delay Causes (23 2018). URL https://transtats.bts.gov/

[7] J. Fry, I. Humphreys, G. Francis, Benchmarking in civil aviation: some empirical evidence, Benchmarking: An International Journal 12 (2) (2005) 125–137.

[8] G. Francis, I. Humphreys, J. Fry, The nature and prevalence of the use of performance measurement techniques by airlines, Journal of Air Transport Management 11 (4) (2005) 207–217.

[9] P. D. Hooper, A. Greenall, Exploring the potential for environmental performance benchmarking in the airline sector, Benchmarking: An International Journal 12 (2) (2005) 151–165.

[10] H. Min, H. Min, Benchmarking the service quality of airlines in the united states: an exploratory analysis, Benchmarking: an International journal 22 (5) (2015) 734–751.

[11] J. Sarkis, S. Talluri, Performance based clustering for benchmarking of us airports, Transportation Research Part A: Policy and Practice 38 (5) (2004) 329–346.

[12] A. Odoni, T. Morisset, W. Drotleff, A. Zock, Benchmarking airport airside performance: Fra vs. ewr, in: 9th USA/Europe Air Traffic Management R&D Seminar, 2011.

[13] X. Gao, A. Malkawi, A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm, Energy and Buildings 84 (2014) 607–616.

[14] M. J. Sharma, S. J. Yu, Performance based stratification and clustering for benchmarking of container terminals, Expert Systems with Applications 36 (3) (2009) 5016–5022.

[15] X. Dai, T. Kuosmanen, Best-practice benchmarking using clustering methods: Application to energy regulation, Omega 42 (1) (2014) 179–188.

[16] E. G. Gomes, J. C. C. B. S. Mello, A. C. R. d. Freitas, et al., Efficiency measures for a non-homogeneous group of family farmers, Pesquisa Operacional 32 (3) (2012) 561–574.

[17] B. Hoffman, J. Krozel, S. Penny, A. Roy, K. Roth, A cluster analysis to classify days in the national airspace system, in: AIAA Guidance, Navigation, and Control Conference and Exhibit, 2003, p. 5711.

[18] G. Chatterji, B. Musaffar, Characterization of days based on analysis of national airspace system performance metrics, in: AIAA guidance, navigation and control conference and exhibit, 2006, p. 6449.

[19] K. Kuhn, A. Shah, C. Skeels, Characterizing and classifying historical days based on weather and air traffic, in: 2015 IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), 2015, pp. 1C3–1–1C3–12.

[20] K. D. Kuhn, A methodology for identifying similar days in air traffic flow management initiative planning, Transportation Research Part C: Emerging Technologies 69 (2016) 1–15.

[21] K. Ren, A. M. Kim, K. Kuhn, Exploration of the evolution of airport ground delay programs, Transportation Research Record: Journal of the Transportation Research Board (2018) 1–11.

[22] A. Estes, D. Lovell, Identifying Representative Traffic Management Initiatives, in: International Conference on Research in Air Transportation (ICRAT), 2016.

[23] Y. Liu, M. Hansen, Evaluation of the performance of ground delay programs, Transportation Research Record 2400 (1) (2014) 54–64.

[24] S. Gorripaty, Y. Liu, M. Hansen, A. Pozdnukhov, Identifying similar days for air traffic management, Journal of Air Transport Management 65 (2017) 144–155.

[25] S. R. Grabbe, B. Sridhar, A. Mukherjee, Similar days in the nas: an airport perspective, in: 2013 Aviation Technology, Integration, and Operations Conference, 2013, p. 4222.

[26] A. Mukherjee, S. Grabbe, B. Sridhar, Classification of Days using Weather Impacted Traffic in the National Airspace System, in: AIAA Aviation Technology, Integration and Operations Conference, 2013.

[27] S. R. Grabbe, B. Sridhar, A. Mukherjee, Clustering days with similar airport weather conditions, in: 14th AIAA Aviation Technology, Integration, and Operations Conference, 2014, p. 2712.

[28] R. Guimera, S. Mossa, A. Turtschi, L. A. N. Amaral, The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles, Proceedings of the National Academy of Sciences 102 (22) (2005) 7794–7799.

[29] M. Zanin, F. Lillo, Modelling the air transport with complex networks: A short review, The European Physical Journal Special Topics 215 (1) (2013) 5–21.

[30] K. Gopalakrishnan, H. Balakrishnan, R. Jordan, Clusters and communities in air traffic delay networks, in: American Control Conference (ACC), 2016, IEEE, 2016, pp. 3782–3788.

[31] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 2009, Ch. 2, pp. 68–125.
[32] L. A. Zager, G. C. Verghese, Graph similarity scoring and matching, Applied Mathematics Letters (2008) 86–94.
[33] B. Luo, R. Wilson, E. Hancock, Spectral clustering of graphs, in: 10th International Conference on Computer Analysis of Images and Patterns, CAIP, 2003.
[34] P. Bonacich, Factoring and weighting approaches to status scores and clique identification, Journal of Mathematical Sociology (2008) 113–120.
[35] S. P. Borgatti, Centrality and network flow, Social Networks (2005) 55–71.
[36] S. E. Schaeffer, Graph clustering, Computer Science Review I (2007) 27–64.
[37] P.-A. Champin, C. Solnon, Measuring the similarity of labeled graphs, in: Proceedings of the 5th international conference on Case-based reasoning (ICCBR'03): Research and Development, 2003, pp. 80–95.
[38] J. J. Rebollo, H. Balakrishnan, Characterization and prediction of air traffic delays, Transportation Research Part C (2014) 231–241.
[39] E. D. Kolaczyk, Statistical Analysis of Network Data: Methods and Models, 1st Edition, Springer Publishing Company, Incorporated, 2009.
[40] Federal Aviation Administration (FAA), ASPM database, http://aspm.faa.gov/ (2017).
[41] C. P. Barros, N. Peypoch, An evaluation of european airlines operational performance, International Journal of Production Economics 122 (2) (2009) 525–533.
[42] B. David Mc A, Service quality and customer satisfaction in the airline industry: A comparison between legacy airlines and low-cost airlines, American Journal of Tourism Research 2 (1) (2013) 67–77.
[43] D. Gillen, T. Hazledine, The economics and geography of regional airline services in six countries, Journal of Transport Geography 46 (2015) 129–136.
[44] S. J. Forbes, M. Lederman, Control rights, network structure and vertical integration: Evidence from regional airlines, Network Structure and Vertical Integration: Evidence from Regional Airlines (November 2005).
[45] S. J. Forbes, M. Lederman, Does vertical integration affect firm performance? evidence from the airline industry, The RAND Journal of Economics 41 (4) (2010) 765–790.
[46] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer-Verlag, New York, 2009.
[47] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, AAAI Press, 1996, p. 226231.
[48] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing 17 (4) (2007) 395–416.
[49] B. S. Everitt, S. Landau, M. Leese, D. Stahl, Cluster analysis 5th ed (2011).
[50] J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM (1999) 604–632.
[51] M. Benzi, E. Estrada, C. Klymko, Ranking hubs and authorities using matrix functions, Linear Algebra and its Applications (5) (2013) 2447–2474.
[52] Federal Aviation Administration, FAA Advisory Database (2017).
URL https://www.fly.faa.gov/adv/advAdvisoryForm.jsp
[53] Federal Aviation Administration, Airport capacity benchmark report (2004).
[54] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics (1987) 53–65.
[55] Y. Bai, Analysis of aircraft arrival delay and airport on-time performance, Master's thesis, University of Central Florida (2006).
[56] K. Gopalakrishnan, H. Balakrishnan, Control and optimization of air traffic networks, Annual Review of Control, Robotics, and Autonomous Systems 4 (2021) 397–424.
[57] Federal Aviation Administration, FAA Announces Slot Changes at Newark Liberty International (2016).
URL www.faa.gov/news/updates/?newsId=85309

Supplementary Materials

Selecting the number of clusters k

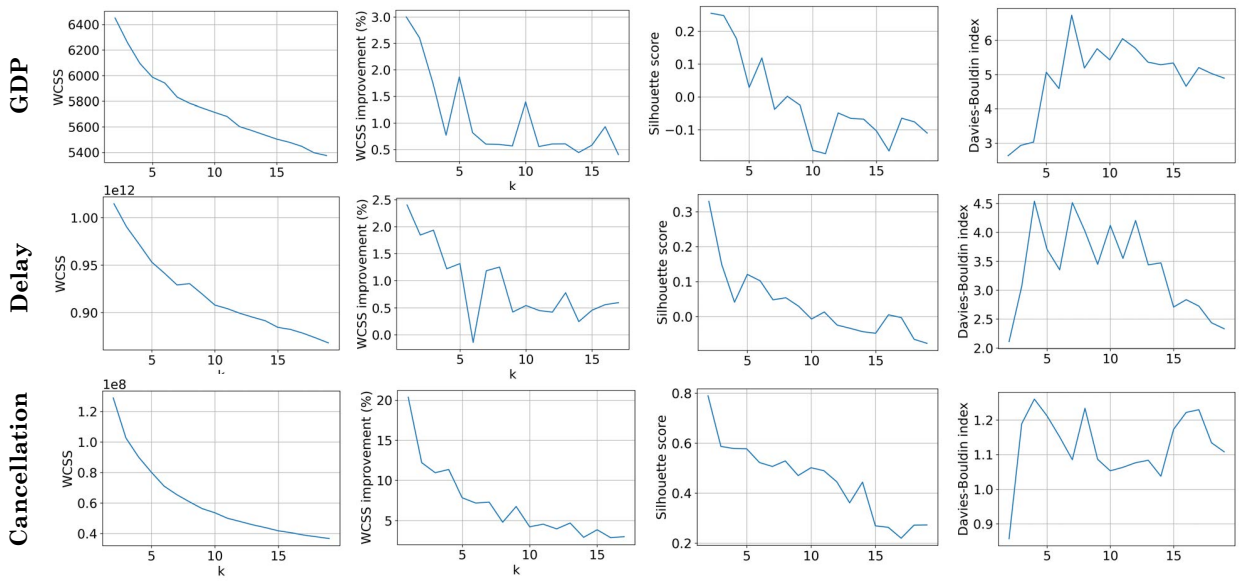

Figure 12: Four measures to guide the selection of k (WCSS, change in WCSS, the silhouette value, and the Davies-Bouldin index, all as a function of k), for k-means clustering of GDP, delay, and cancellation networks.

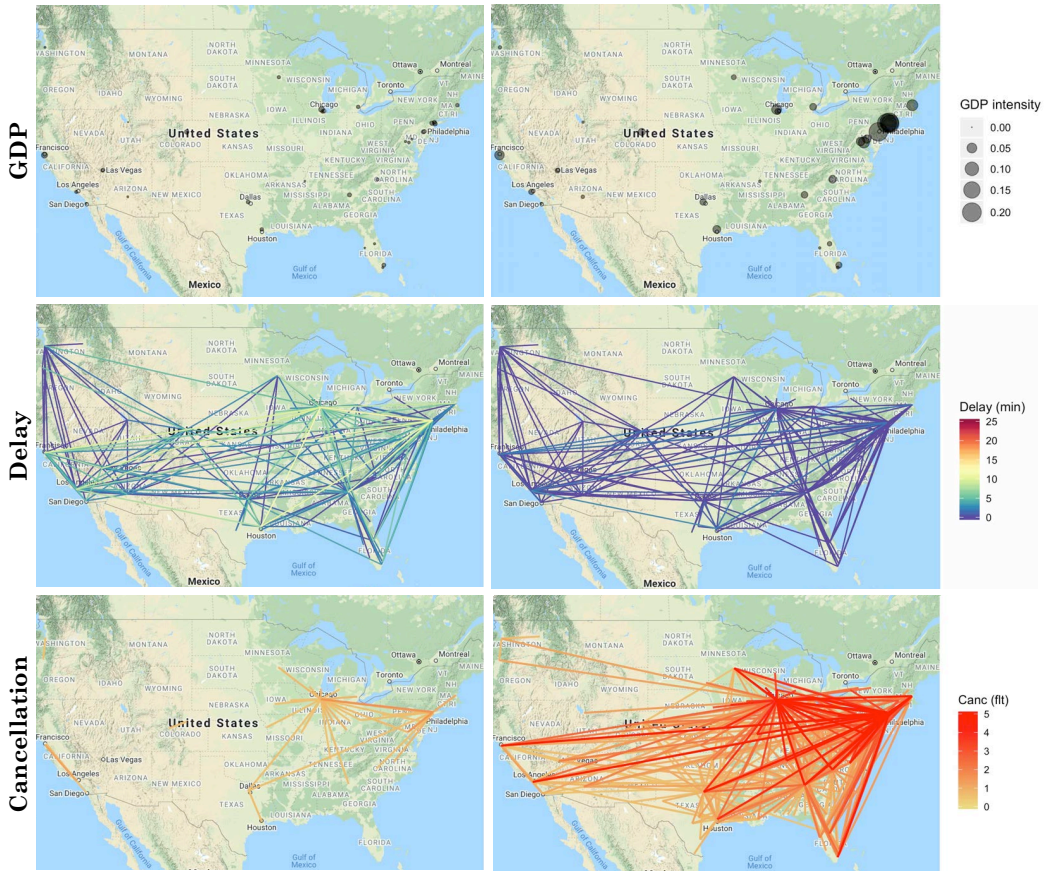

Figure 13: Resultant GDP, delay, and cancellation centroids if a small number of clusters was chosen (in this case, $k = 2$).

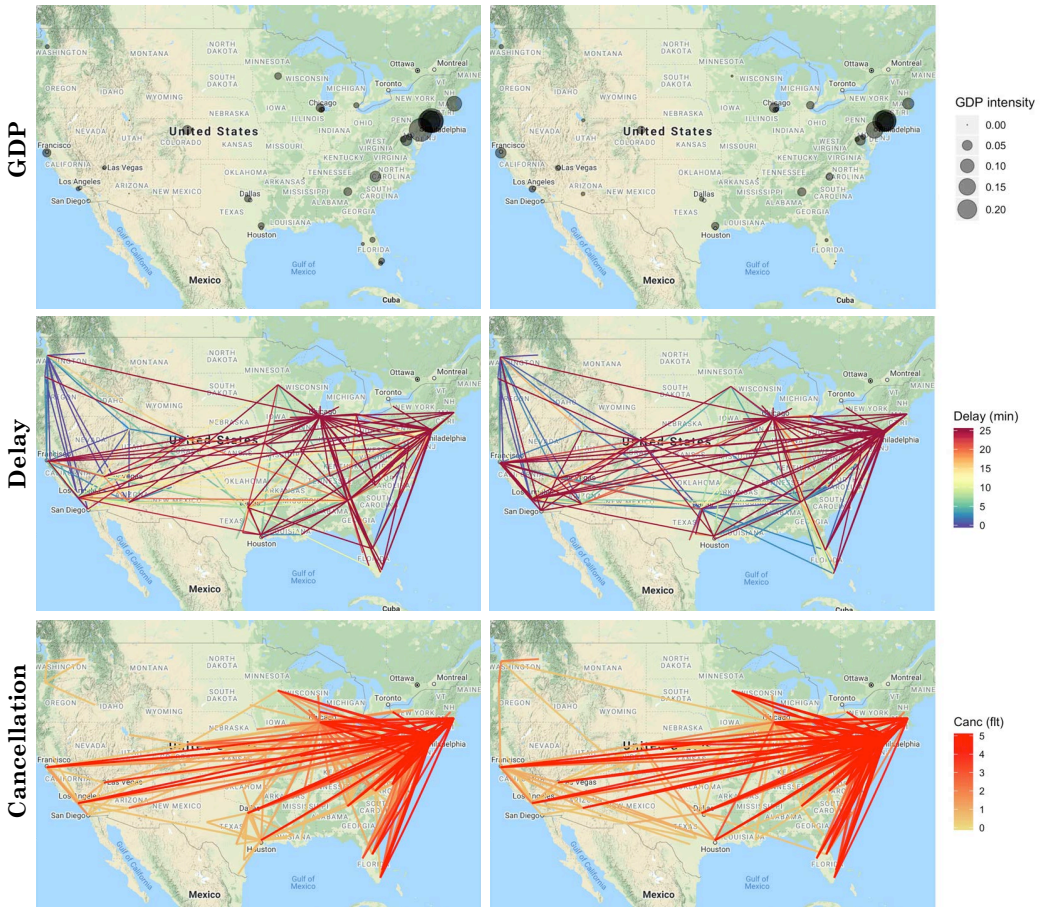

Figure 14: Similar (redundant) cluster centroid pairs for GDP, delay, and cancellation networks if a large number of clusters was chosen (in this case, $k = 15$).

Data statistics: Number of flights

GDP type	All Airlines	Airline A1	Airline A2	Airline A3
ATL	9.26×10^5	1.88×10^5	2.06×10^5	1.32×10^5
CHI	1.18×10^6	2.42×10^5	2.62×10^5	1.73×10^5
LowNAS	7.02×10^6	1.41×10^6	1.55×10^6	1.11×10^6
MedNE	2.32×10^6	4.72×10^5	5.17×10^5	3.61×10^5
Miami_NE	6.8×10^5	1.47×10^5	1.47×10^5	1.01×10^5
NE	3.23×10^5	6.68×10^4	7.29×10^4	4.48×10^4
NYC_PHL	7.61×10^5	1.54×10^5	1.71×10^5	1.18×10^5
SFO	1.51×10^6	3.07×10^5	3.37×10^5	2.42×10^5

Table 5: Number of flights per GDP type cluster, split by airlines, in the training set spanning January 1, 2014 through December 31, 2015.

GDP type	A1 Mainline	A1 Regional	A2 Mainline	A2 Regional	A3 Mainline	A3 Regional
ATL	6.19×10^4	1.26×10^5	1.04×10^5	1.02×10^5	6.46×10^4	6.78×10^4
CHI	7.94×10^4	1.63×10^5	1.3×10^5	1.32×10^5	8.39×10^4	8.92×10^4
LowNAS	4.69×10^5	9.4×10^5	7.89×10^5	7.57×10^5	5.36×10^5	5.79×10^5
MedNE	1.55×10^5	3.17×10^5	2.58×10^5	2.59×10^5	1.73×10^5	1.88×10^5
Miami_NE	4.4×10^4	1.02×10^5	6.84×10^4	7.83×10^4	5.02×10^4	5.08×10^4
NE	2.21×10^4	4.47×10^4	3.64×10^4	3.65×10^4	2.2×10^4	2.29×10^4
NYC_PHL	5.18×10^4	1.02×10^5	8.6×10^4	8.54×10^4	5.64×10^4	6.19×10^4
SFO	1.01×10^5	2.06×10^5	1.67×10^5	1.7×10^5	1.16×10^5	1.26×10^5

Table 6: Number of flights per GDP type cluster, split by mainline and regional carriers of airlines A1, A2, and A3, in the training set spanning January 1, 2014 through December 31, 2015.

Date (m/d/y)	All Airlines	A1	A2	A3	A1 Main	A2 Main	A3 Main	A1 Reg	A2 Reg	A3 Reg
7/1/16	23161	4197	5288	6186	1579	6186	2754	2618	2477	3432
2/6/15	20369	3953	4621	2948	1323	2948	1411	2630	2349	1537
2/21/15	15905	3170	3007	2566	1014	2566	1315	2156	1359	1251
1/8/17	19434	3409	4262	5389	1286	5389	2331	2123	1964	3058

Table 7: Number of flight operations during the two pairs of days (July 1, 2016 and February 6, 2015; February 21, 2015 and January 8, 2017) presented in the benchmark panels from Figures 7 and 15, respectively.

January 8, 2017 and February 6, 2015 (`SFO` GDP)

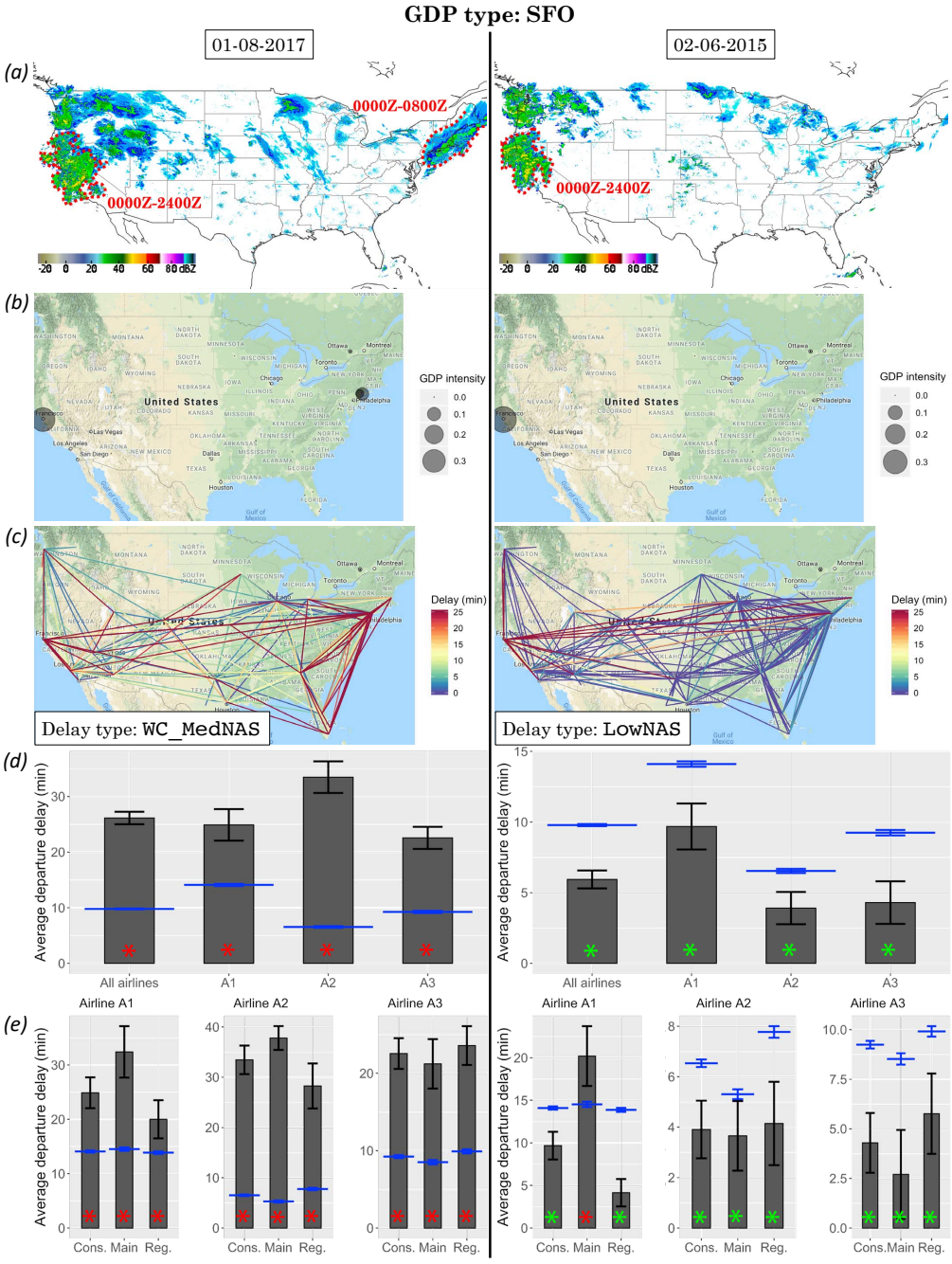

Figure 15: Sample benchmarking panel for two SFO-GDP type days; *(a)* composite weather radar returns from 0000Z to 2400Z; *(b)* the GDP intensity; *(c)* the delay networks; *(d)* benchmarks of on-time performance across all airlines and three major US airlines; *(e)* benchmarks of on-time performance across the consolidated fleets, mainline, and regional carriers of three major US airlines.

Additional examples of rare (G, D, C) triplets

- *September 6, 2014*: In terms of GDP, this day was a `NE` type-of-day, with significant impacts in the Washington, DC area, Philadelphia, and New York City. Despite the intense GDP activity in the Northeast, it was a `LowNAS` type-of-day both in terms of delay and cancellations. This type of triplet combinations only occurred twice in 2014 through 2016. Our hypothesis is that the low demand in terms of the number of scheduled flights on this particular day – almost 27% lower than what is typically seen during `NE` GDP type-of-days, possibly because it was the Saturday following Labor Day weekend – mitigated any congestion-related delays or cancellations.

- *March 2, 2014*: This day was classified as the triplet combination (`SFO`, `WC_MedNAS`, `Med_CHI_NE`). At first, it seems counterintuitive that capacity reductions and delays on the West Coast is associated with significant cancellations in the Northeast. A deeper analysis reveals the presence of a snowstorm over Chicago and the Northeast regions of the US. We hypothesize that this particular winter storm was well-predicted, resulting in extensive proactive cancellations. Since the demand at airports were reduced significantly due to the cancellations, the demand-capacity imbalance was less severe, and required little GDP issuances for the Northeast region.