

Paper 162-2012

Using SAS[®] to Measure Airport Connectivity: An Application of Weighted Betweenness Centrality for the FAA National Plan of Integrated Airport Systems (NPIAS)

Hector Rodriguez-Deniz, University of Las Palmas de Gran Canaria, Canary Islands, Spain

ABSTRACT

The US Federal Aviation Administration (FAA) estimates that \$52.2 billion will be available over the next five years (2011-2015) to fund airport infrastructure developments. Because one of the main objectives is to reduce congestion and delays, there is a need to acknowledge the importance of connectivity (measured with a centrality indicator) when establishing funding priorities. Currently, the FAA does not do this. In this paper, we expand an existing SAS/IML[®] implementation of betweenness centrality to handle passenger-weighted airport networks, construct a graphical representation of the US air transport network from airline ticketing data, test the module to identify hub airports, and produce stylish output using SAS[®] GMAP. Performance and complexity considerations of the new algorithm are addressed.

INTRODUCTION

The US Federal Aviation Administration (FAA) estimates that \$52.2 billion will be available over the next 5 years (2011-2015) to fund infrastructure developments for all segments of civil aviation under the Airport Improvement Program (AIP). The National Plan of Integrated Airport Systems (NPIAS) is used by the FAA in administering the AIP. In the NPIAS (FAA, 2011), investment requirements and funding priorities are set according to a decades-old airport typology based on the proportions over total US passenger enplanements¹. While the merit (and simplicity) of such an approach are not questioned, the existing literature widely agrees that the importance of a single airport within a network needs to take into account its hubbing potential (i.e. connectivity), which the FAA does not. Since one of the main objectives of the AIP is to fund airport capacity expansions in order to reduce congestion and delays, from a social perspective it seems only reasonable that funding priority should be given to airports playing a central role in the network, not just because they process a significant proportion of US traffic but also because passengers and airlines are connecting through them to other destinations, which will also benefit from any delay reductions at the hub. Hence, there is a potential for the FAA, as a public agency, to optimize the social benefits from any AIP investment by improving the NPIAS airport classification with a hubbing/connectivity measure, such as, for example, the weighted betweenness centrality.

Taking into account the significant amount of data involved in such calculations (i.e. down to flight coupon level), SAS is clearly a suitable software for the task. In this study, we will expand a previous SAS implementation (Ellis, 2009) of the betweenness centrality to deal with weighted networks, construct a graph-like structure representing the US air transport network using detailed airline ticketing data, test the implementation to discover relevant airports in terms of centrality and finally produce a stylish output using the PROC GMAP graphical utility and Annotate data sets. Many other SAS tools such as the SQL and IML procedures were used in order to accomplish other stages of the process.

Since Freeman (1977) defined a set of indices to discover relevant actors within a social network, centrality measures have been increasingly used in a variety of academic and practitioner circles. In the air transport literature, a wide range of models for measuring airport connectivity have been proposed (See e.g. Burghouwt and Redondi, 2009), and betweenness centrality has become a popular approach in this context. In spite of that, the applicability of this topological indicator to the proposed NPIAS problematic is very limited because it does not take into account the importance of the different passenger markets served through each airport, which, in the end, is the key factor to determine how “central” the airport is to the US network. In order to solve that, this paper proposes a slight variation of betweenness centrality to incorporate market-based weightings. An extension of the Ellis (2009) SAS implementation, which allow the computation of a weighted betweenness centrality, is presented based on the suggestions of Brandes (2008). This new approximation will allow us to take into account the relevance of each market in terms of connecting passengers without worsening the algorithm time complexity with respect to the original Ellis (2009) module. A performance test on a set of synthetic networks, and a few brief remarks on complexity conclude the paper.

¹ Large hubs are those airports that each account for at least 1% of total US passenger enplanements. Medium and Small hubs are defined as airports that each account for between 0.25-1 and 0.05-0.25 %, respectively (FAA, 2011).

DATABASE DESCRIPTION

Data comes from the Bureau of Transportation Statistics of the Research and Innovative Technology Administration² (US Department of Transportation). The Airline Origin and Destination Survey (BTS, 2011) is a sample of airline ticket information from more than 30 US carriers such as American Airlines, JetBlue and Delta, among others³. The survey covers about 10% of tickets from reporting carriers. Detailed information of each domestic itinerary (e.g. origin and destination airport, miles flown) is provided in several tables to analyze air traffic patterns, airline dominance or passenger flows. Data is available on a quarterly basis since 1993. In our case we focus on identifying potential airline hubs (i.e. airports) using network centrality measures, so detailed stopover information between origin and destination is required. The DB1BMarket⁴ table contains such information, which comes in the form of a string variable that indicates the origin and destination airport for every flight segment within a single itinerary, allowing us to proceed with the calculation of centrality measures in a straightforward way. Our sample period (fourth quarter 2010 to third quarter 2011⁵) contains more than 20,000,000 records representing individual itineraries. In order to process this amount of data, several Macro definitions, DATA steps, PROC SQL and PROC IML were used throughout the data load and preparation process in SAS.

BETWEENNESS CENTRALITY OF THE US AIR TRANSPORT NETWORK

THEORETICAL BACKGROUND

As noted before, betweenness centrality quantifies the prominence of an actor in terms of connectivity within a network (Freeman, 1977; Brandes, 2001). It is common to represent networks as a graph $G = (V, E)$, where V is the set of vertices or nodes, and E the set of edges, which represent connections between nodes. The number of vertices and edges in the network are given by $n = |V|$ and $m = |E|$, respectively. For the sake of simplicity, undirected connections are assumed, and this clearly applies to our problem since we are looking for potential hubs regardless the direction of individual flights. We denote w as the weight function of a network. In a weighted network, we assume that $w > 0$ for all $e \in E$, while $w = 1$, $e \in E$, in the case of unweighted networks. Consequently, the length of a path between any two vertices of the network will be either the sum of the weights of its edges, or the number of steps in an unweighted network. The shortest distance $d_G(s, t)$ between two nodes s and t is the minimum length of any path in G that connects s and t . The betweenness centrality measure is given by:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where σ_{st} is the number of minimum length paths connecting nodes $s \in V$ and $t \in V$, and $\sigma_{st}(v)$ is the number of such paths in which some $v \in V$ lies on. It is also clear that $\sigma_{st} = \sigma_{ts}$, and $\sigma_{ss} = 1$ by convention.

From an air transport perspective, this measure presents some drawbacks. Since the calculation of the index for a vertex $v \in V$ involves all the shortest paths between any combination of nodes $s, t \in V$ except v on the graph, each origin-destination pair will contribute to a vertex v centrality regardless of either distance or relationship between them, as long as v lies on any shortest path. Consequently, a well positioned but irrelevant airport (e.g. in terms of passenger traffic) could be highly ranked even though it lies on paths which do not represent any real market⁶. Another difficulty, even when considering weighted networks, comes up when it comes to quantifying the relative prominence of a specific route. Certainly, weights can be employed to represent distances, travel time and many other interesting features in a network, but once the length and number of shortest paths have been calculated, each of them contribute identically to the betweenness centrality of v , given some $s, t \neq v$. Note that the use of weights is restricted to finding the shortest paths but not to estimate its relative importance with respect to e.g. the number of enplanements (Skiena, 2008). Therefore, that means that well-established markets (e.g. New York LGA to Orlando via Atlanta) and routes on which scheduled traffic is insignificant would sum up equally to the centrality index if they lie in the shortest paths, provided certain criteria such as origin-destination distance were met.

² <http://www.bts.gov>

³ For a comprehensive list of the recent reporting carriers, visit: <http://www.transtats.bts.gov/ReleaseInfo.asp?tb=247&display=data>

⁴ http://www.transtats.bts.gov/TableInfo.asp?Table_ID=247&DB_Short_Name=Origin%20and%20Destination%20Survey&Info_Only=0

⁵ Latest data available at February 2012.

⁶ We define a market as a specific origin-destination journey, thus aggregating all possible routings.

These shortcomings are closely related to the fact that betweenness centrality has a strong topological motivation. However, we need some mechanism to identify key airports in an air transport network according to both their topological position (i.e. connectivity potential) and the relevance of the markets they serve in terms of traffic density⁸. Considering that our market definition is not point-to-point oriented, traditional weighted betweenness centrality does not seem to be a suitable choice in this case. A number of variants of the standard algorithm have been proposed to tackle the problem of shortest paths weighting (see e.g. Borgatti and Everett, 2006; Brandes, 2008). A common approach is to weight all shortest paths inversely proportional to their length, so the contribution of the shorter shortest paths to centrality index is increased, to the detriment of the longer ones. Then, the new indicator is defined as:

$$C_{B_{dist}}(v) = \sum_{s \neq v \neq t \in V} \frac{1}{dist_{st}} \cdot \frac{\sigma_{st}(v)}{\sigma_{st}}$$

In the present study, the shortest paths between every origin-destination $s, t \in V$ will be weighted similarly, according to the "importance" of said market, which is calculated as the total number of passengers that travelled on that market (Q_{st}), divided by the total number of passengers in the sample (Q), as follows:

$$C_{B_{mkt}}(v) = \sum_{s \neq v \neq t \in V} \frac{Q_{st}}{Q} \cdot \frac{\sigma_{st}(v)}{\sigma_{st}}$$

This variation will allow us to identify major airline hubs (i.e. airports) by taking into account not only their central position within the network but also the importance of the markets they serve in terms of transit, providing an additional dimension to the betweenness centrality while the original properties of the index remains unchanged. As a result, top ranked airports are likely to play an important role within the US network by combining a central location with relevant market service. Airports lacking of either characteristic will be probably mid-ranked. Airports with similar traffic levels will be classified according to their centrality.

Ellis (2009) presented a PROC IML module for the calculation of an unweighted betweenness centrality, given an input adjacency matrix, using the Brandes (2001) algorithm. In this paper, we extended the original module to include the weight matrix (Q_{st}) as an additional input parameter, as well as to update the pair-dependencies accumulation consequently, following the guidelines specified in Brandes (2008). Minor changes were made to the previous code in order to complete this task, as is shown below in an extract of the final part of the new module. Additionally, this approach will permit us to apply weighting to the betweenness centrality and compute the algorithm in time $\mathcal{O}(nm)$, as is corroborated in the following sections.

```

...
do while (%isempty(stack) ^= 1); /* while stack is not empty */
  %pop(stack, ww); /* use double-w; this is distinct from */
  /* the w used for neighbors-to-v above */
  vv=loc(p[ww,]); /* indices of vertices on ww's list of predecessors*/
  /* use vv - this is a new v, too */
  do k=1 to ncol(vv);
    /* the addition of weights is straightforward in this case */
    /* Q[s,t] represents the relative importance of the S-T market */
    delta[vv[k]]=delta[vv[k]]+(sigma[vv[k]]/sigma[ww])*(Q[s,ww]+delta[ww]);
  end;
  if ww ^= s then cb[ww]=cb[ww]+delta[ww];
end;
end;
cb=cb/2; /* undirected -> divide by 2 */
cb=cb/max(cb); /* and normalize */
return(cb);
finish;

```

⁸ It is widely known that traffic connecting through hubs supports high-frequency services in a hub and spoke model (Doganis, 2001).

DATA PREPARATION⁹

Data from the BTS Airline Origin and Destination Survey (DB1Market table) needs some preprocessing prior to the calculation of the betweenness centrality using the PROC IML module mentioned in the previous section. We need to obtain a graph representation of the database, that is, an adjacency matrix, and in the present case, a weight matrix as well. The adjacency matrix is a zero-one square matrix of order $n = |V|$ which indicates which vertices are connected. The weight matrix has a similar structure but having positive values representing the relevance of each market (total transit passengers). Both matrices should be symmetric if the connections are undirected, as in the current approach.

The data set resulting from merging data of the selected quarters (4th 2010 to 3rd 2011) contains nearly 23 million records¹⁰ and 39 attributes. However, we just need two variables as the starting point to construct the adjacency and weight matrices for our air transport analysis: AirportGroup and Passengers. The AirportGroup attribute is a record of each point (i.e. airport) in the itinerary in which a passenger has stopped, and fits our purposes perfectly. Nonetheless, it comes in the form of a colon-separated concatenation of IATA codes (see Figure 1) which needs to be partitioned into individual origin-destination segments. DATA steps, SQL and IML procedures were employed for the most part of this process, and the adjacency matrix was obtained easily. An extract of the PROC IML that creates an array of segments out of the previously divided AirportGroup attribute (using SUBSTR functions) is listed at the end of the page. Regarding the weight matrix, we just need a PROC SQL to group the data by origin and destination, aggregating the number of passengers of each itinerary without taking into account the number of segments on every route. After some additional operations, e.g. transposing the data, the final matrices were generated. They characterize a 446x446 low-density¹¹ (0.18) air transport network with vertices and edges representing airports and connectivity between airports, respectively. In the next section, we will put everything into practice measuring the degree of centrality of the US airports for the specified sample period.

| | DestStateFips | DestState | DestStateName | DestWac | AirportGroup |
|---------|---------------|-----------|---------------|---------|--------------|
| 1541936 | 06 | CA | California | 91 | JFK:SFO |
| 1541937 | 36 | NY | New York | 22 | SFO:ATL:JFK |
| 1541938 | 06 | CA | California | 91 | JFK:SFO |
| 1541939 | 36 | NY | New York | 22 | SFO:ATL:JFK |
| 1541940 | 06 | CA | California | 91 | JFK:SFO |
| 1541941 | 36 | NY | New York | 22 | SFO:ATL:JFK |

Figure 1. Sample data from the 2011's DB1Market table displaying the AirportGroup attribute

```
*Creates a data set of individual origin-destination segments;
proc iml;
  use sasgf.mkt20113;
  read all var _char_ into segmentMatrix;
  read all var _num_ into itineraryLength;
  nRow = NROW(segmentMatrix);
  output = J(nRow*3,2,'aaa');
  count = 1;
  do i=1 to nRow;
    *1 stopover itinerary (2 O/D segments);
    if itineraryLength[i]=11 then; do;
      output[count,1]=segmentMatrix[i,1];
      output[count,2]=segmentMatrix[i,2];count = count+1;
      output[count,1]=segmentMatrix[i,2];
      output[count,2]=segmentMatrix[i,3];count = count+1;
    end;
  end;
  ...
```

⁹ Refaat (2007) is an excellent reference for data preprocessing in SAS, as well as Delwiche and Slaughter (2008).

¹⁰ Point-to-point services were deleted from the sample since our study is focused on connecting traffic, resulting in 13 million observations after all.

¹¹ The global density of a network indicates how saturated it is in terms of number of edges and vertices. It is given by the number of edges, m , divided by $\frac{n(n-1)}{2}$ (Brandes, 2001).

RESULTS AND DISCUSSION

Results for the normalized centrality index ($C_{B_{mkt}}$) are shown in Table 1 and Figure 2. For comparative purposes, estimates of the standard betweenness centrality (topological index) are shown in Table 2. With a few exceptions, top-ranked airports combine large passenger traffic with a central location that allows them to appear in a significant number of shortest-path connections in high-density routes. As expected, the introduction of the market-based weighting has led to different results from previous implementations of betweenness centrality in airport networks. For example, while Anchorage (ANC) was ranked 2nd worldwide by Guimerà et al. (2005), the same airport moves down to the 92th position in the present study. The explanation is very simple: ANC's central location as sole gateway to Alaska's airport network led to a very high degree of "topological" centrality. This privileged location, however, does not translate into significant hubbing activity because very low density of the O/D markets served through ANC (e.g. Juneau-Fort Lauderdale). From a purely topological perspective, Denver was the most centrally-located airport in our study. Nevertheless, once the market-based weights were applied, Atlanta (ATL) claimed the first position as the largest and indeed most central airport in the US.

Table 1. Most central airports in the US air transportation network in 2010/2011.

| Rank | $C_{B_{mkt}}$ | IATA | State | City | Airport |
|------|---------------|------|-------|---------------|---|
| 1 | 1,00000 | ATL | GA | ATLANTA | HARTSFIELD-JACKSON ATLANTA INTL |
| 2 | 0,97810 | ORD | IL | CHICAGO | CHICAGO O'HARE INTL |
| 3 | 0,93606 | DFW | TX | DALLAS-FW | DALLAS/FORT WORTH INTL |
| 4 | 0,92675 | DEN | CO | DENVER | DENVER INTL |
| 5 | 0,85406 | LAX | CA | LOS ANGELES | LOS ANGELES INTL |
| 6 | 0,79663 | MSP | MN | MINNEAPOLIS | MINNEAPOLIS-ST PAUL INTL/WOLD-CHAMBERL. |
| 7 | 0,77810 | MCO | FL | ORLANDO | ORLANDO INTL |
| 8 | 0,77628 | DTW | MI | DETROIT | DETROIT METROPOLITAN WAYNE COUNTY |
| 9 | 0,77385 | LAS | NV | LAS VEGAS | MC CARRAN INTL |
| 10 | 0,75765 | IAD | DC | WASHINGTON | WASHINGTON DULLES INTL |
| 11 | 0,72849 | PHX | AZ | PHOENIX | PHOENIX SKY HARBOR INTL |
| 12 | 0,70818 | IAH | TX | HOUSTON | GEORGE BUSH INTERCONTINENTAL/HOUSTON |
| 13 | 0,69749 | BOS | MA | BOSTON | GENERAL EDWARD LAWRENCE LOGAN INTL |
| 14 | 0,69610 | DCA | DC | WASHINGTON | RONALD REAGAN WASHINGTON NATIONAL |
| 15 | 0,67078 | SEA | WA | SEATTLE | SEATTLE-TACOMA INTL |
| 16 | 0,66501 | CLT | NC | CHARLOTTE | CHARLOTTE/DOUGLAS INTL |
| 17 | 0,66049 | STL | MO | ST LOUIS | LAMBERT-ST LOUIS INTL |
| 18 | 0,65586 | BNA | TN | NASHVILLE | NASHVILLE INTL |
| 19 | 0,65487 | LGA | NY | NEW YORK | LA GUARDIA |
| 20 | 0,64985 | PHL | PA | PHILADELPHIA | PHILADELPHIA INTL |
| 21 | 0,64946 | BWI | MD | BALTIMORE | BALTIMORE/WASHINGTON INTL THURGOOD MR. |
| 22 | 0,64281 | EWR | NJ | NEWARK | NEWARK LIBERTY INTL |
| 23 | 0,64253 | TPA | FL | TAMPA | TAMPA INTL |
| 24 | 0,62367 | MCI | MO | KANSAS CITY | KANSAS CITY INTL |
| 25 | 0,61461 | SFO | CA | SAN FRANCISCO | SAN FRANCISCO INTL |

Source: Own elaboration.

Table 3 shows the important differences between FAA's top-25 implicit ranking (i.e. those airports labelled as "large hubs" based on passenger enplanements - see Figure 3) and the one based on weighted betweenness centrality. First, note that JFK, MIA, and SFO drop quite significantly, 28 and 37 places, respectively, from the original FAA ranking. This is clearly related to their status as international gateways, which cannot be properly characterized by the proposed indicator because of the absence of international passenger movements in the BTS database. Another significant drop is that of Honolulu Airport (HNL). Similarly to ANC, HNL is the major gateway to the Hawaiian airport system and thus it scores very high in the topological component of $C_{B_{mkt}}$. However, given its isolated location, HNL does not participate in a large number of O/D markets between other airport pairs in the mainland, where the bulk of US traffic is located. Thus, even though it is one of the busiest airports in the US, HNL's low ranking makes sense from a policy perspective. The reason is that any investment aimed at reducing congestion and delays at this airport will not lead to maximum social benefits because only a small portion of the network will be affected, which is exactly what the proposed centrality indicator aims to measure.

The case of SLC, which dropped 6 places and moved out of the top-25, is particularly interesting because it shows the impact of airline networks on airport connectivity. Even though it scores high in the topological measurement, the introduction of market-based weights penalizes the airport for not being a "central" hub in Delta's network, which is by far the dominant carrier at SLC (approx. 73% market share). Thus, passengers travelling on Delta will be less likely to transfer at SLC than in e.g. Atlanta. The impact of hub competition between different airlines is clearly seen in the case of IAH, largest hub of United/Continental, yet sharing importance with its second largest hub ORD. This network configuration leads to a significant loss of centrality to DFW, which concentrates a significant share of the operations of American Airlines. The remaining airports that have dropped in the $C_{B_{mkt}}$ ranking are located in either coast and

thus, are evidently less central for their respective dominant carriers (LAX, CLT, EWR, PHL, FLL). The opposite applies to airports such as DTW or MSP which score very high in a topological sense. Finally, we can also identify a number of apparently non-central airports, such as BOS or IAD, that participate in a significant number of high density coast-to-coast markets. The resulting airport classification is proposed as a new reference to allocate funding for capacity investments as airports are now ranked according to their importance and influence over the rest of the network.

Table 2. Top 25 central airports according to standard betweenness centrality in 2010.

| Rank | C_B | IATA | State | City | Airport |
|------|---------|------|-------|----------------|---|
| 1 | 1,00000 | DEN | CO | DENVER | DENVER INTL |
| 2 | 0,73311 | ANC | AK | ANCHORAGE | TED STEVENS ANCHORAGE INTL |
| 3 | 0,59819 | MSP | MN | MINNEAPOLIS | MINNEAPOLIS-ST PAUL INTL/WOLD-CHAMBERL. |
| 4 | 0,37392 | BOS | MA | BOSTON | GENERAL EDWARD LAWRENCE LOGAN INTL |
| 5 | 0,33728 | ATL | GA | ATLANTA | HARTSFIELD - JACKSON ATLANTA INTL |
| 6 | 0,32180 | LAS | NV | LAS VEGAS | MC CARRAN INTL |
| 7 | 0,31153 | SEA | WA | SEATTLE | SEATTLE-TACOMA INTL |
| 8 | 0,28222 | IAH | TX | HOUSTON | GEORGE BUSH INTERCONTINENTAL/HOUSTON |
| 9 | 0,27379 | DTW | MI | DETROIT | DETROIT METROPOLITAN WAYNE COUNTY |
| 10 | 0,23330 | LAX | CA | LOS ANGELES | LOS ANGELES INTL |
| 11 | 0,23170 | IAD | DC | WASHINGTON | WASHINGTON DULLES INTL |
| 12 | 0,23038 | DFW | TX | DALLAS-FW | DALLAS/FORT WORTH INTL |
| 13 | 0,21537 | SLC | UT | SALT LAKE CITY | SALT LAKE CITY INTL |
| 14 | 0,21314 | MCI | MO | KANSAS CITY | KANSAS CITY INTL |
| 15 | 0,19984 | STL | MO | ST LOUIS | LAMBERT-ST LOUIS INTL |
| 16 | 0,19072 | ORD | IL | CHICAGO | CHICAGO O'HARE INTL |
| 17 | 0,18168 | SFO | CA | SAN FRANCISCO | SAN FRANCISCO INTL |
| 18 | 0,16859 | CLE | OH | CLEVELAND | CLEVELAND-HOPKINS INTL |
| 19 | 0,16121 | ABQ | NM | ALBUQUERQUE | ALBUQUERQUE INTL SUNPORT |
| 20 | 0,15632 | MEM | TN | MEMPHIS | MEMPHIS INTL |
| 21 | 0,14908 | PHX | AZ | PHOENIX | PHOENIX SKY HARBOR INTL |
| 22 | 0,12776 | BWI | MD | BALTIMORE | BALTIMORE/WASHINGTON INTL THURGOOD MR. |
| 23 | 0,12402 | EWR | NJ | NEWARK | NEWARK LIBERTY INTL |
| 24 | 0,11816 | HNL | HI | HONOLULU | HONOLULU INTL |
| 25 | 0,10848 | ALB | NY | ALBANY | ALBANY INTL |

Source: Own elaboration.

Table 3. FAA Large Hubs 2010 versus Most Central Airports 2010/2011.

| Rank FAA | Rank $C_{B, mkt}$ | Change | IATA | State | City | Hub/Focus City for |
|----------|-------------------|--------|------|-------|----------------|-----------------------------|
| 1 | 1 | 0 | ATL | GA | ATLANTA | DELTA, SOUTHWEST |
| 2 | 2 | 0 | ORD | IL | CHICAGO | UNITED, AMERICAN |
| 3 | 5 | -2 | LAX | CA | LOS ANGELES | AMERICAN, SOUTHWEST, UNITED |
| 4 | 3 | 1 | DFW | TX | DALLAS-FW | AMERICAN |
| 5 | 4 | 1 | DEN | CO | DENVER | UNITED, SOUTHWEST, FRONTIER |
| 6 | 34 | -28 | JFK | NY | NEW YORK | JETBLUE, DELTA, AMERICAN |
| 7 | 12 | -5 | IAH | TX | HOUSTON | UNITED |
| 8 | 25 | -17 | SFO | CA | SAN FRANCISCO | VIRGIN AMERICA, UNITED |
| 9 | 9 | 0 | LAS | NV | LAS VEGAS | SOUTHWEST |
| 10 | 11 | -1 | PHX | AZ | PHOENIX | USAIRWAYS, SOUTHWEST |
| 11 | 16 | -5 | CLT | NC | CHARLOTTE | USAIRWAYS |
| 12 | 49 | -37 | MIA | FL | MIAMI | AMERICAN |
| 13 | 7 | 6 | MCO | FL | ORLANDO | SOUTHWEST, DELTA, JETBLUE |
| 14 | 22 | -8 | EWR | NJ | NEWARK | UNITED |
| 15 | 8 | 7 | DTW | MI | DETROIT | DELTA |
| 16 | 6 | 10 | MSP | MN | MINNEAPOLIS | DELTA |
| 17 | 15 | 2 | SEA | WA | SEATTLE | ALASKA |
| 18 | 20 | -2 | PHL | PA | PHILADELPHIA | USAIRWAYS |
| 19 | 13 | 6 | BOS | MA | BOSTON | JETBLUE |
| 20 | 19 | 1 | LGA | NY | NEW YORK | DELTA, AMERICAN |
| 21 | 10 | 11 | IAD | DC | WASHINGTON | UNITED |
| 22 | 21 | 1 | BWI | MD | BALTIMORE | SOUTHWEST |
| 23 | 40 | -17 | FLL | FL | FT LAUDERDALE | SOUTHWEST, SPIRIT, JETBLUE |
| 24 | 30 | -6 | SLC | UT | SALT LAKE CITY | DELTA |
| 25 | 98 | -73 | HNL | HI | HONOLULU | HAWAIIAN |

Source: FAA (2011), Own elaboration.

Betweenness Centrality of U.S. Airports 2010/2011 - Top 25

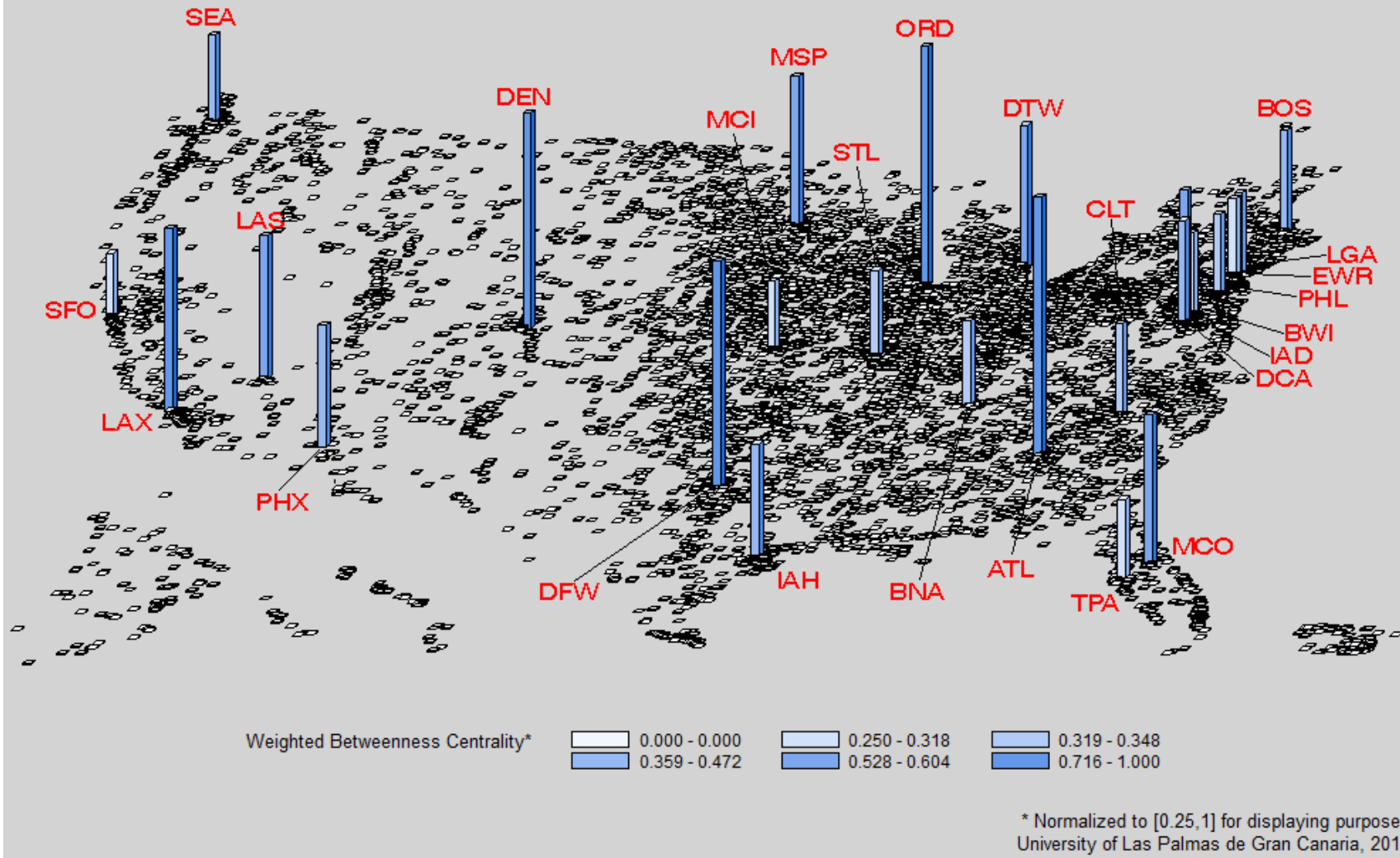


Figure 2. Block US City Map of the most central airports in 2010/2011 using PROC GMAP

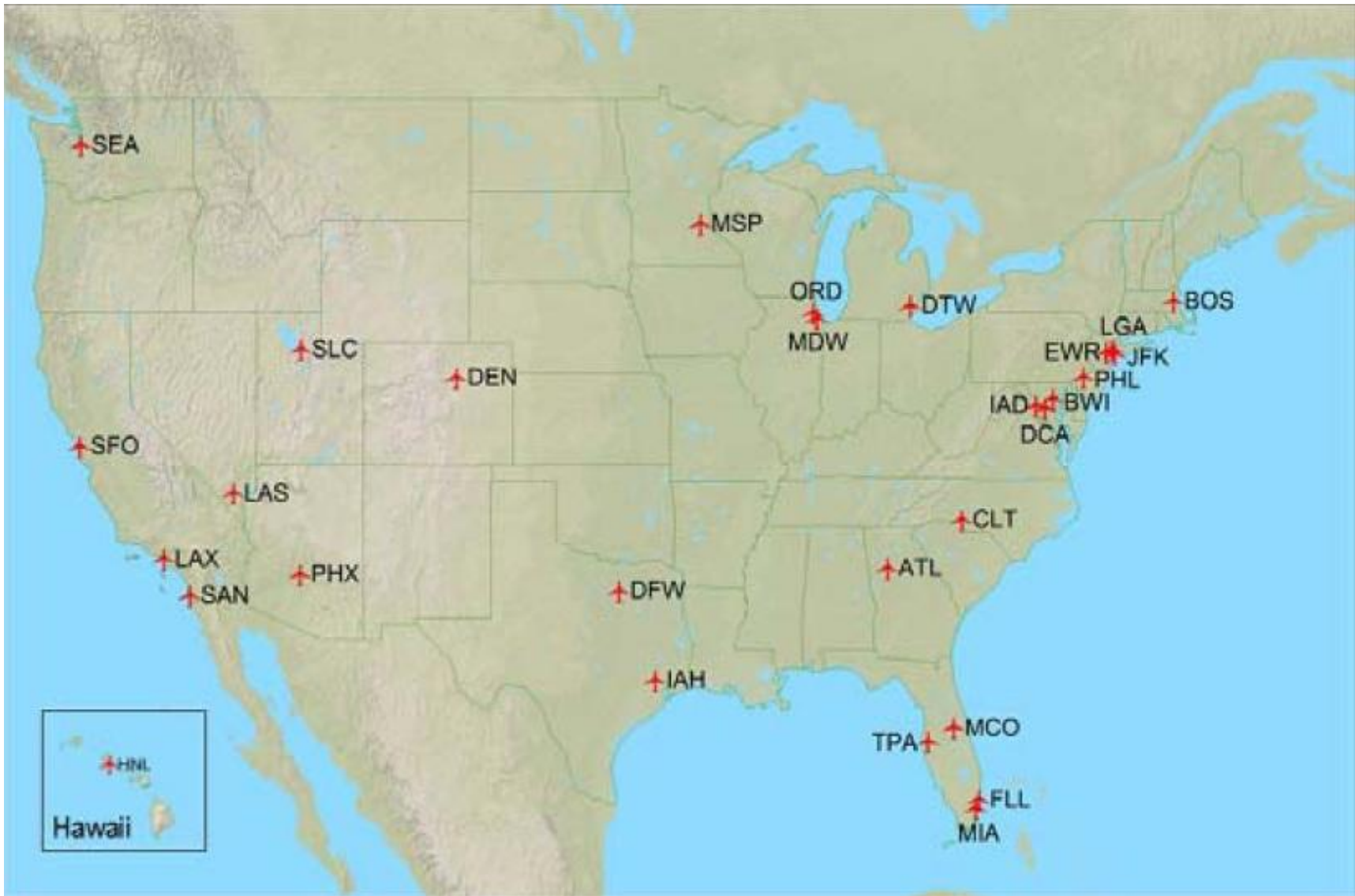


Figure 3. Large airport hubs in the US air transportation network (FAA, 2011)

Figure 2 graphically depicts the results presented in Table 1. The PROC GMAP utility along with Annotate data sets were employed to represent the values of the betweenness centrality of the top 25 airports in a stylish manner. PROC GMAP is intended to produce two- or three-dimensional maps (e.g. block maps) that display the value of a variable between different areas. In this work we altered the SAS's US City map for our particular purposes in order to generate a block map representing the centrality index of each airport. A plain map without any additional indications would not be very informative so an Annotate data set containing airport IATA codes as labels was prepared to be used in the PROC GMAP. With the purpose of enhancing the visual discrimination between airports in the resulting map, the centrality values of the top 25 airports were normalized to the range [0.25,1] using the STDIZE procedure, which standardizes a set of variables using a variety of criteria following:

$$y = y_{min} + R \left(\frac{x - x_{min}}{x_{max} - x_{min}} \right)$$

Where x, y are the input and standardized value, x_{min}, x_{max} the minimum and maximum input values, y_{min} the minimum value for the standardized range and R the width of the standardized range (i.e. $y_{max} - y_{min}$). Sample code of range normalization, labels and map generation is presented below.

```
*Range Normalization to [0.25,1];
proc stdize data=sasgf.output2011 out=sasgf.output2011
            method=RANGE add=0.25 mult=0.75; var coll;
run;
...
*Top 25 airport labels and map;
data label;
  length function $ 8 text $ 20;
  retain function 'label' xsys ysys hsys '3' size 2;
  position='4'; color="RED"; style="SWISSX";
  x=93; y=72; text='BOS'; output;      x=50; y=71; text='MCI'; output;
  x=59; y=68; text='STL'; output;      x=63; y=79; text='ORD'; output;
  x=83; y=33; text='MCO'; output;      x=77; y=63; text='CLT'; output;
  x=94; y=59; text='LGA'; output;      x=93; y=57; text='EWR'; output;
  x=92; y=55; text='PHL'; output;      x=91; y=52; text='BWI'; output;
  x=90; y=50; text='IAD'; output;      x=89; y=48; text='DCA'; output;
  x=18; y=37; text='PHX'; output;      x=71; y=72; text='DTW'; output;
  x=16; y=62; text='LAS'; output;      x=38; y=29; text='DFW'; output;
  x=76; y=28; text='TPA'; output;      x=70; y=31; text='ATL'; output;
  x=63; y=29; text='BNA'; output;      x=55; y=30; text='IAH'; output;
  x=54; y=76; text='MSP'; output;      x=13; y=80; text='SEA'; output;
  x=3; y=54; text='SFO'; output;      x=7; y=44; text='LAX'; output;
  x=35; y=73; text='DEN'; output;
run;

goptions reset=all cback=lightgray;
pattern value=memory;
title1 'Betweenness Centrality of U.S. Airports 2010/2011 - Top 25'
height=10;
footnote1 j=r "* Normalized to [0.25,1] for displaying purposes";
footnote2 j=r "University of Las Palmas de Gran Canaria, 2012";

proc gmap data=sasgf.airportsmap map=sasgf.airportsmap;
  id city;
  block bcent /blocksize=0.8 levels=6 area=state relzero annotate=label;
run;
```

COMPUTATIONAL PERFORMANCE

In this section, we present a simple performance simulation of the betweenness centrality to ensure that our variant does not affect the $O(nm)$ running time of the original Brandes' algorithm for unweighted networks (Brandes, 2001). A set of 60 random networks (both adjacency and weight matrices) with 100 to 1500 vertices and {0.2, 0.4, 0.6, 0.8} densities were generated and tested. We used PROC IML procedures, encapsulated in a macro, to perform this task. Sample code of the macro and execution time results are shown below (Figure 4). It was easy to prove that execution time was on the order of $O(nm)$, with a linear OLS regression model. The tests were performed on an Intel i7-920 with 6 GB of main memory running SAS v9.2 (32 bits) on Windows 7 professional. This approximation has allowed us to take into account a significant feature of the field (the relevance of the routes in terms of transit) without worsening the algorithm time complexity (standard weighted betweenness is computed in $O(nm + n^2 \log n)$), with respect to the original approach.

```
%macro benchmark(dens=,nver=);
%put BENCHMARK STARTED FOR DENSITY = &dens AND NUMBER OF VERTICES = &nver;
proc iml;
  x = J(&nver,&nver,0);
  y = J(&nver,&nver,0);
  m = round((&dens*&nver*(&nver-1))/2);
  *Adjacency matrix;
  do k=1 to m;
    hit=0;
    do while (hit^=1);
      i = ceil(rand("Uniform")*&nver);
      j = ceil(rand("Uniform")*&nver);
      if x[i,j]=0 & (i^=j) then; do; *Add a link between a;
        x[i,j]=1; x[j,i]=1; *pair of nodes [i,j];
        hit=1;end;end;end;

  *Weight matrix;
  seed = 12345;
  w = j(&nver,&nver,seed);
  y = uniform(w); *Random;
  do i=1 to &nver;
    do j=1 to &nver;
      if i=j then y[i,j]=0;end;end;
  y=y/max(y); *Normalized weights;
  create benchadj from x; append from x;
  create benchwgt from y; append from y;
quit;
...
```

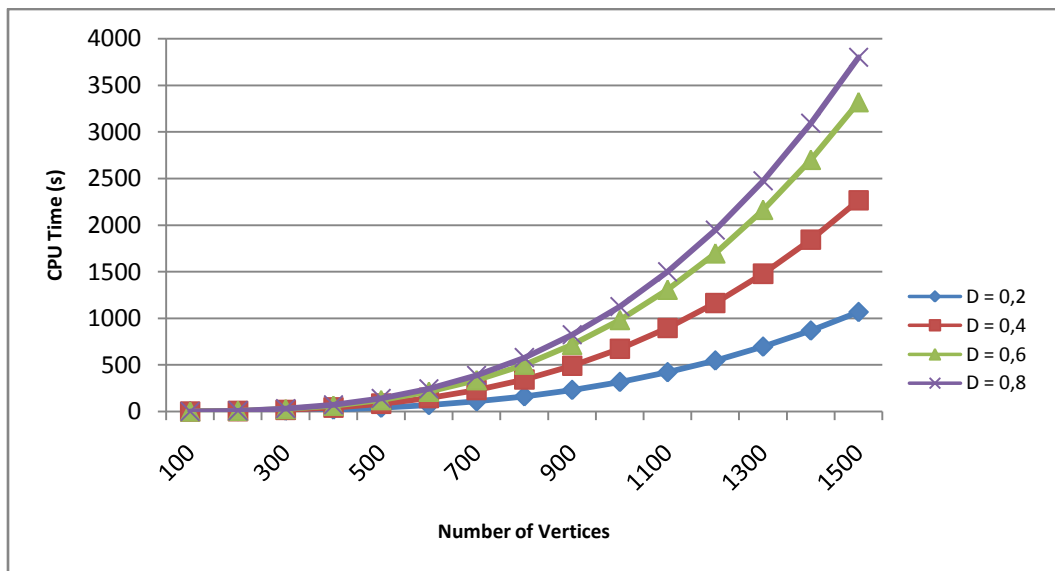


Figure 4. Benchmarking results for the betweenness centrality (D = network density)

SUMMARY

The US Federal Aviation Administration uses a very simple airport typology (based on passenger enplanements) in order to set investment requirements and funding priorities while administering the Airport Improvement Program. Taking into account that one of the main objectives of the AIP is to reduce congestion and delays, the consideration of the airport's hubbing potential would help in allocating said investments to the most important nodes in the network, leading to maximum social welfare. Thus, this paper aims at developing a new variation of the well-known betweenness centrality index in order to measure hub connectivity in the US airport network. Evidence from previous studies suggested that the purely topological measure of betweenness centrality was not an appropriate indicator of hubbing activity as it does not take into account the importance of the passenger markets served through the airport. This information was added to our indicator via market-based weights at a city-pair level.

Results show significant improvement from previous studies. Airports with a high level of centrality in isolated regions (such as Alaska or Hawaii) are no longer considered central hubs in the US network. By comparing the FAA's implicit ranking of top-25 "large hubs" with the one resulting from the modified betweenness centrality, several major differences can be observed. Many airports located in either coast, or otherwise serving as non-central hubs for their respective dominant carriers, dropped significantly in the ranking. The existence of multiple hub configurations and even hub competition between different airlines is also shown to reduce centrality. The resulting airport classification is proposed as a new reference to allocate funding for capacity investments as airports are now ranked according to their importance and influence over the rest of the network.

A range of SAS technologies and procedures (e.g. DATA step, PROC IML, PROC SQL, PROC GMAP, SAS Macro) were employed to integrate and preprocess a set of airline ticketing databases, extend a previous SAS implementation of the betweenness centrality index to analyze the hubbing potential of the US airports, produce a stylish output using SAS map graphical utilities and finally test the computational performance of the new algorithm. This project demonstrates how SAS can be used not only as a set of built-in analytical solutions, but also as a suitable framework for the implementation of state-of-the-art algorithms from almost any field of study.

Future research should study the evolution of these indicators during the last decades (data is publicly available online) in order to analyze the impact of airline decisions on airport hubbing and network connectivity. If airport rankings are not the preferred option to set funding priorities, the application of clustering techniques combining the topological and market-based measurements would be a suitable option to generate an alternative airport typology. All of these procedures can be easily implemented in SAS.

REFERENCES

- BTS, 2011. Origin and Destination Survey (DB1B).
http://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=125&Link=0
- Borgatti, S. P. and Everett, M. G., 2006. A graph-theoretic perspective on centrality. *Social Networks* 28, 466-484.
- Brandes, U., 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* 25 (2), 163-177.
- Brandes, U., 2008. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks* 30 (2), 136-145.
- Burghouwt, G. and Redondi, R., 2009. Connectivity in air transport networks: models, measures and applications, No 0901, Working Papers, Department of Economics and Technology Management, University of Bergamo.
<http://econpapers.repec.org/RePEc:brh:wpaper:0901>.
- Delwiche, L. D. and Slaughter, S. J., 2008. *The Little SAS Book: A Primer*, 4th Edition. SAS Institute Inc. Cary, NC.
- Doganis, R., 2001. *The Airline Business in the 21st Century*, 1st Edition. Routledge. London.
- Ellis, A. R., 2009. Using SAS to calculate betweenness centrality. *INSNA Connections* 29 (1), 26-32.
- FAA, 2011. Federal Aviation Administration National Plan of Integrated Airport System (NPIAS) 2011-2015 Report.
http://www.faa.gov/airports/planning_capacity/npias/reports/
- Freeman, L. C., 1977. A set of measures of centrality based on betweenness. *Sociometry* 40, 35-41.
- Guimerà, R., Mossa, S., Turtschi, A. and Amaral, L. A. N., 2005. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America* 102 (22), 7794-7799.
- Refaat, M., 2007. *Data Preparation for Data Mining Using SAS*. Morgan Kaufmann. San Francisco, CA.
- Skiena, S. S., 2008. *The Algorithm Design Manual*, 2nd Edition. Springer-Verlag, London.

ACKNOWLEDGEMENTS

I offer my sincerest gratitude to Dr. Casiano Manrique de Lara, who supported me with his knowledge and patience, introduced me to SAS and suggested me to participate in the Student Ambassador Program. I am always impressed by his knowledge and experience. I also want to thank Prof. Juan Méndez, who introduced me to the fields of Operations Research and Pattern Recognition, for his advice and constant support. Finally, I gratefully acknowledge the invaluable help of my friend and colleague Dr. Augusto Voltes-Dorta throughout all the stages of this research, which has greatly benefited from his comments and suggestions.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Hector Rodriguez-Deniz
School of Informatics
University of Las Palmas de Gran Canaria
Canary Islands, Spain
E-mail: hrodriguez@becarios.ulpgc.es
Linkedin: [linkedin.com/in/hrdeniz](https://www.linkedin.com/in/hrdeniz)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.