

Analysis of US Domestic Air Travel Cost Using GIS and Spatial Analysis

Gang Gong
Department of Geography
Boston University
Email: ggong@bu.edu

INTRODUCTION

Deregulation movement has brought enormous changes to the US domestic airline industry in the past a quarter century since the Airline Deregulation Act was signed into law in 1978. It has been widely hailed as a success. The total numbers of enplanements and passenger miles have more than doubled since then, and the overall airfare has been considerably lower than it would have been had regulation continued. The traveling public in general has enjoyed a larger network of connections, greater service choices and lower fares on average. However, this overall success in deregulation should not mask the fact that all these changes in airline industry have not been even across space. There are significant geographic variations in the cost of US domestic air travel. While many people benefit from the decrease of airfares, some communities claim that they have been suffering deteriorating air travel services, mostly in the form of high airfares or short of services, namely the problem of "pockets of pain". Given the important place of air travel in the whole economy, the geographic variations of air travel cost may have significant implications for regional economies and overall public welfare.

After the removal of the restrictions posted on airline industry in regulation years, airfares have taken a more and more complex structure. Traditional bond between airfares and distance has broken down. Instead, airfares are heavily influenced by factors such as scale economies, competition, airport congestion and airline marketing strategies. In general, longer flights tend to have lower average cost because the fixed costs associated with each flight can be spread over a longer distance. Average cost may also be expected to be lower in markets with larger passenger volume since airlines in those markets are able to use larger planes and achieve higher load factors. The level of competition is another factor that may influence air travel cost. Because airlines may compete each other in different forms, and competition can happen both at the airports and in the route markets, the influences on cost from competition have been quite complicated. Whether competition has impact over airfares had been a controversial issue even before the implementation of the airline deregulation. One theoretical foundation for deregulation was the contestable theory, and the contestability theory in its pure form suggests that the number of actual competitors should have no effect on prices. However, many studies have found that the number of airlines actually competing on a route has a significant effect on the price level (Bailey, Graham, and Kaplan, 1985; Call and Keeler, 1985; Morrison and Winston, 1987; Borenstein, 1989; Hurdle et al., 1989). If one airline establishes dominance in a specific city-pair market, it tends to have the monopoly power to set high fares. In the similar way, if one airline sets fortress hub dominance at certain airport, it can also exert considerable influence on the airfares charged on the flights to and from that airport. One recent study by the US Department of Transportation (2001) found clear evidence that carriers charge higher fares in the absence of effective competition.

Studies about airfares have been plenty. Spatial analysis techniques, however, have been used rarely. Most of the previous studies have adopted a standard multivariate linear regression approach by regressing airfare on a series of pertinent variables and using ordinary least squares method to estimate the parameters (Anderson et al., 2002; Black, 1992; Borenstein, 1989; Evans and Kessides, 1993). Due to the fact that airfares in city-pair markets have been a spatial phenomenon happening in a network configuration (airfares can be regarded as a link attribute), spatial factor may come into

play. For example, the airfare in one market may be correlated with the airfares in some connected markets (markets sharing one common endpoint). This phenomenon is called network autocorrelation (Black, 1992). Network autocorrelation is special kind of spatial autocorrelation where spatial dependence exists among random variables associated with the links of a network. The phenomenon of spatial autocorrelation has been recognized and studied for years (Cliff and Ord, 1973, 1981; Goodchild, 1987; Griffith, 1987). However, most researches have focused on the analysis of spatial data in which the basic unit of observations is point or area. Study of spatial autocorrelation happening in a network context has been rare. Network autocorrelation concerns the dependence of variable values on given links to such values on other links to which it is connected in a network context (Black, 1992). Airline network is a typical example in this context. Since airfare is a spatial variable associated with links in the airline service network, spatial dependence is likely to happen among the connected links. In statistical analysis, this dependence will violate the independence assumption on which the classic regression analysis is based, and may lead to biased results, some of the optimal properties of the OLS estimates may no longer hold. To solve this problem, one solution is to construct a spatial regression model with the spatial dependence structure incorporated. Once the spatial autocorrelation effect is accounted for, the regression model will provide satisfactory results.

DATA

Domestic Airline Fares Consumer Report was the principal data source used for both descriptive and statistical analysis in this paper. The Report has been issued quarterly by the Department of Transportation since the third quarter of 1996. Currently the latest issue is for the second quarter of 2002. The report provides data on air travel activities and fares for the 1000 largest city-pair markets in the contiguous 48 states. The 1000 largest city-pair markets generally account for about 75% of all domestic passengers flying between destinations in the contiguous 48 states. Starting in the fourth quarter of 1998, coverage was expanded to include every city-pair market with an average of at least 10 passengers per day. Thus the most recent data covers nearly 100% of all passengers. For each city pair, the average number of daily passengers are reported along with three fares: the average fare, the fare of the airline with the largest market share, and the lowest fare of any airline serving the market with at least a 10% share. All the information in the report is drawn from an underlying database comprising a 10% sample of all domestic flight coupons (DOT DataBank 1A).

DESCRIPTIVE ANALYSIS

Figure 1 shows the average one-way fare and yield¹ for the top 1000 largest city-pair markets from the 3rd quarter of 1996 to the 2nd quarter of 2002. The average one-way fare showed an increasing trend with a clear seasonal pattern before the year of 2001. Then in 2001, due to the impact of 9/11, average one-way fare plummeted. The average one-way yield curve echoes the fare curve mostly. It has a relatively small seasonal variation.

To study the air travel cost, I start with examining the relationship between airfares and distance. Using the data for the 2nd quarter of 2002 as an example, I plot the average one-way fare against the flight distance. The result is showed in Figure 2. Although longer trips tend to have higher fares, the relationship between fare and distance is very weak. This confirms an earlier claim that the old link between fare and distance in the regulation years has broken down.

Figure 1 Average one-way fare and yield for the top 1000 largest city-pair markets

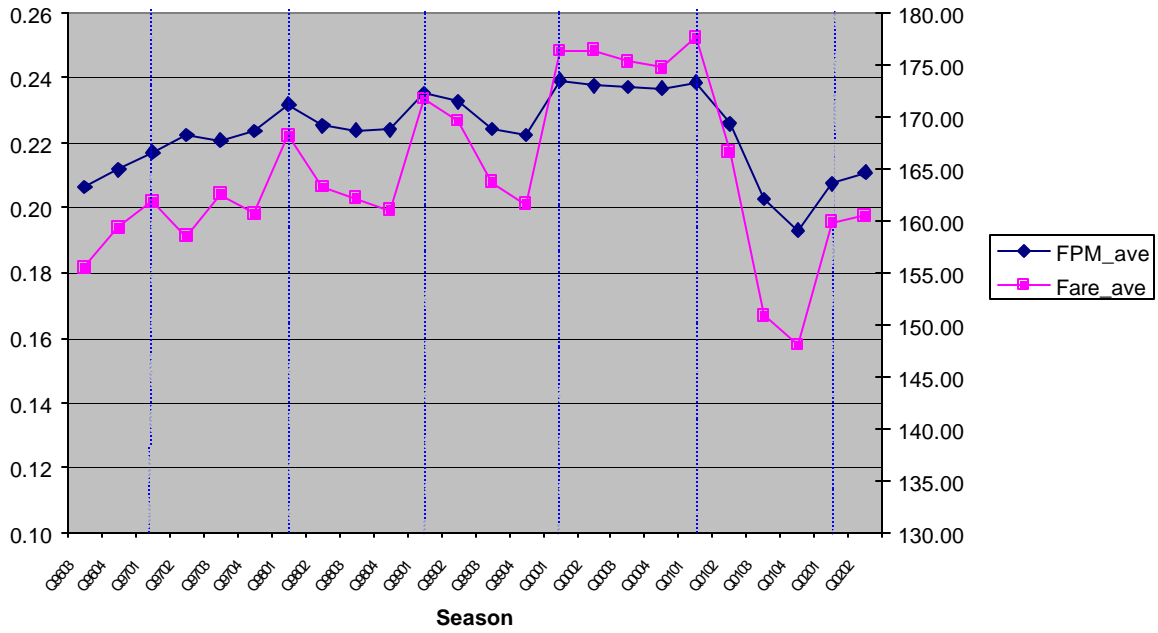


Figure 2 Average one-way fare vs. flight distance

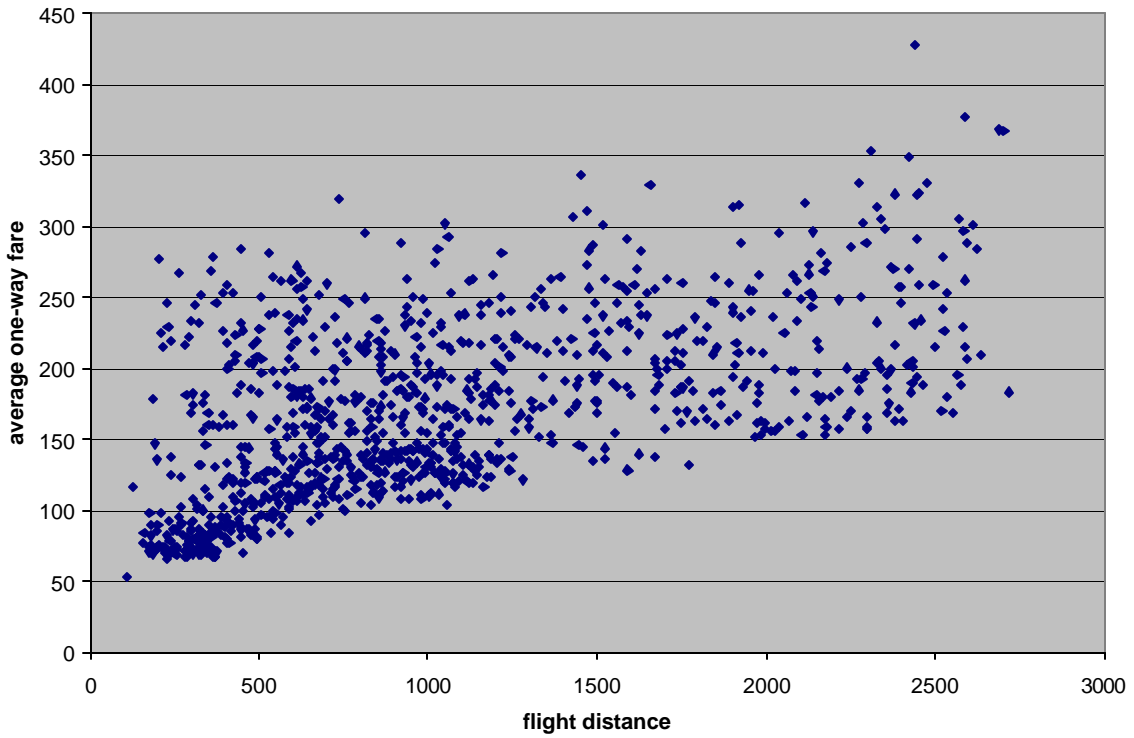


Figure 3 Average one-way yield vs. flight distance

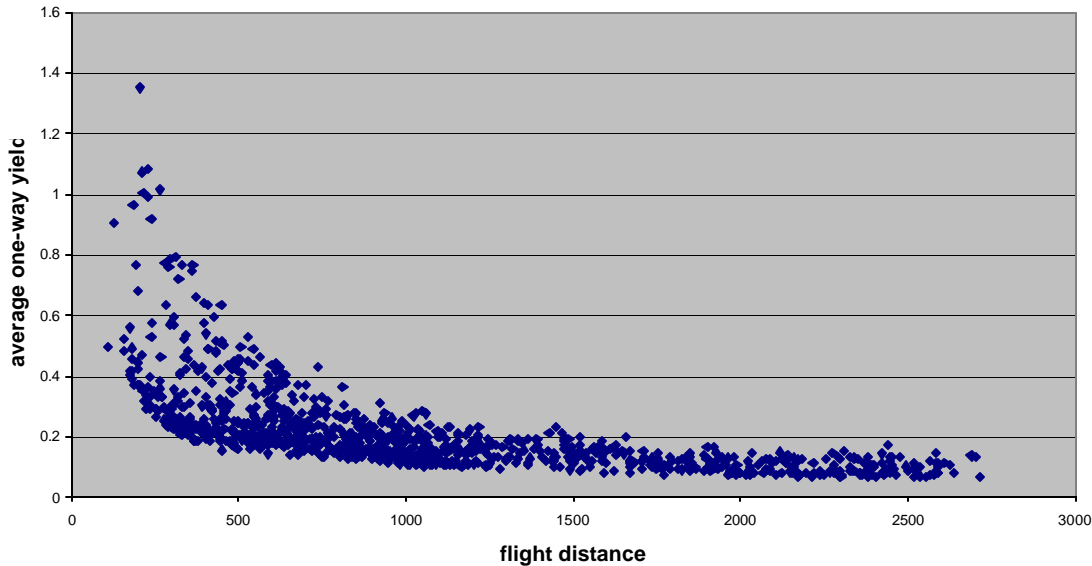


Figure 3 plots the average one-way yield against the flight distance. A clear inverse curvilinear relationship is displayed. This confirms the claim that longer trips tend to have lower yield because the fixed costs associated with each flight can be spread over a larger number of miles.

Figure 3 shows the relationship between yield and travel distance, but it fails to reveal any information about the locational variations in air travel yield. Direct geographic display of yield is difficult because yield is an attribute associated with links in the airline service network, while showing 1000 links in one graph will make any pattern hard to discern. Therefore, for the display purpose, I calculated the average yield weighted by passengersⁱⁱ for each airport. Then for each season, I sorted all the airports by the average yield and picked up the highest ten and the lowest ten airports. Finally I aggregated the lists of the selected airports for all the seasons and displayed them in one map (as showed in Figure 4).

In Figure 4 one can easily discern a cluster of low yield airports in Florida, and a large high yield area in the Mid-Atlantic region. Airports around the Great Lakes region also tend to have lower air travel yields.

Figure 4 illustrates that the average air travel yields at the airports are not homogeneously distributed across the country. It shows some clustering patterns – places that are close to each other tend to have similar values. This regional variation of yields suggests a possible existence of spatial dependence among the airports. Since the yields at the airports are calculated through a veraging the city-pair market level yields, it is reasonable to speculate that the market level yields may also have some form of spatial associations.

REGRESSION ANALYSIS

To test the hypotheses about the factors underlying the variations in air travel yields, I first applied a simple linear regression model. To be consistent with the analyses undertaken in the later part, I only used the data for the top 100 largest city-pair markets in the second quarter of 2002.

The general specification of the regression equation is

$$f_{ij} = b_0 + b_1 d_{ij} + b_2 p_{ij} + b_3 S_{ij} + b_4 P_i + b_5 P_j + e$$

The dependent variable f_{ij} is the yield for the market connecting airport i and airport j. Independent variables d_{ij} , p_{ij} , and S_{ij} are the distance, passenger volume and largest carrier share for that market respectively. P_i (P_j) is the total number of passengers originating from airport i (j), and calculated as $P_i = \sum_j p_{ij}$.

The results of the OLS regression analysis are shown in Table 1.

Figure 4 Highest 10 and Lowest 10 Airports in Average Yield (All Seasons)

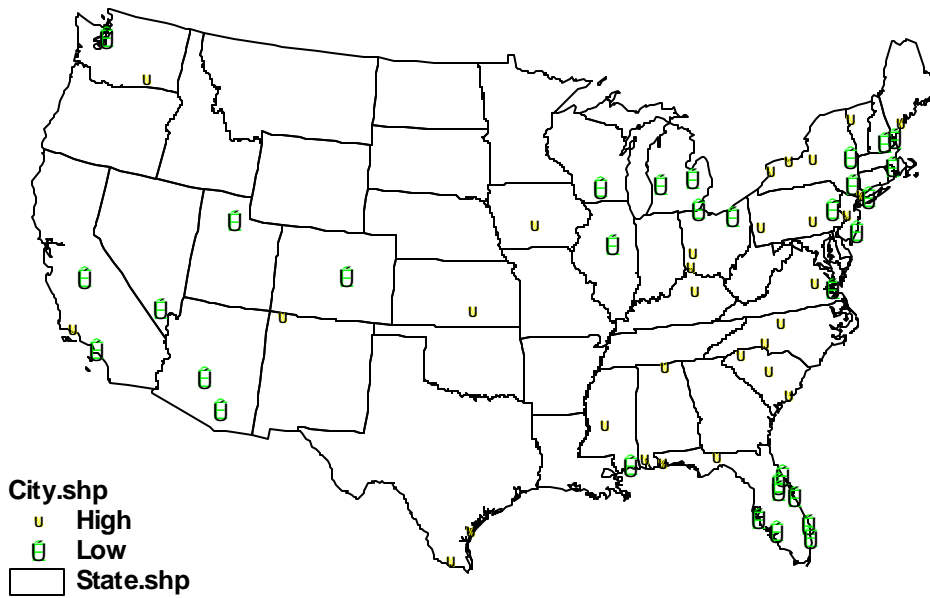
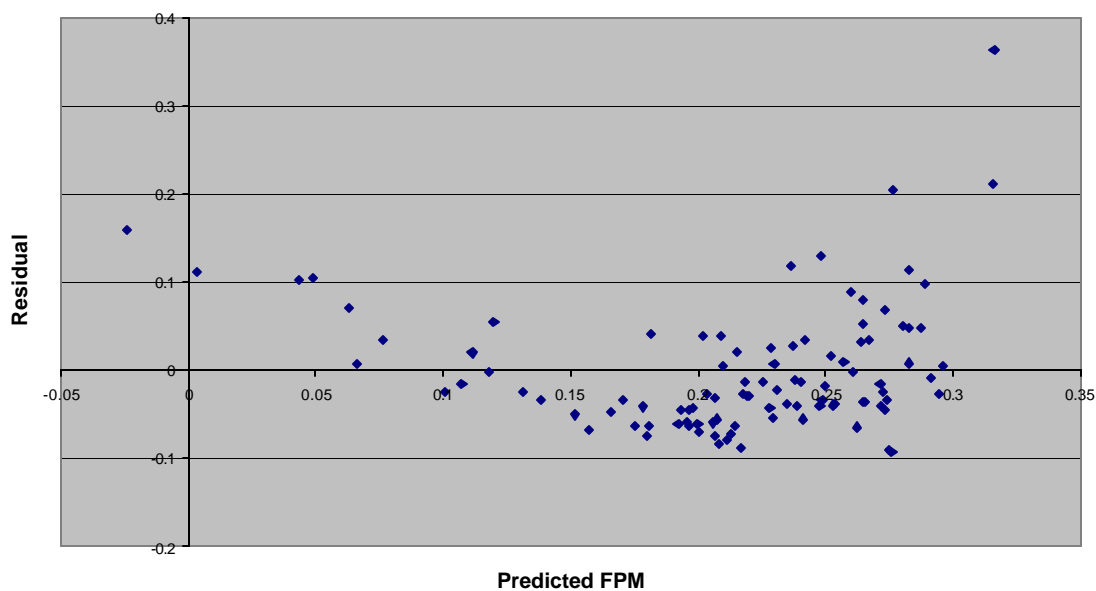


Table 1 Regression Results for the Top 100 markets from Q0202, Dependent Variable: Yield			
	<i>Coefficients</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	0.181384	3.13	0.0023
Distance d_{ij}	-0.000111	-8.05	0.0000
Passengers p_{ij}	-0.000007	-0.94	0.3482
Largest Carrier Share S_{ij}	0.000732	1.33	0.1883
Origin Size P_i	0.000001	2.51	0.0138
Destination Size P_j	0.000001	2.84	0.0056
Adjusted R Square	0.430		

Distance has the expected negative sign, showing that longer trips tend to have lower yields. The scale economy on each market is captured by the negative influence from the passenger volume. The corresponding p-value indicates that this influence is not statistically significant. One possible explanation for this insignificance comes from the way in which the data is selected. Since only the largest 100 markets are included in the regression, there lacks enough variation in the passenger volume variable to drive it significant. A separate regression using the same specification was conducted with equal number of records randomly selected. The results showed a significant scale economies effect. The largest carrier market share is a proxy measure of market dominance. The positive sign indicates that dominant markets tend to have higher airfares. This confirms the hypothesis that competition influences airfares and lack of competition drives airfares high. The origin and destination sizes variables both have positive effects, suggesting that airfares tend to be higher in the markets connected with large airports. This effect may have various explanations, each reflecting different unaccounted-for underlying forces. Larger airports usually have much more severe congestion problems. Costs associated with congestion may be simply passed on to the passengers in the form of higher fares. Another explanation relates to the way in which this analysis was formulated. The regression equation applied actually only explains the supply-side function of airfares. The demand-size effects – for example, airfares may be higher in some markets because travelers are willing to pay more, are not considered. A simultaneous equations model will be explored in further research.

Figure 5 Residuals Plot (linear)

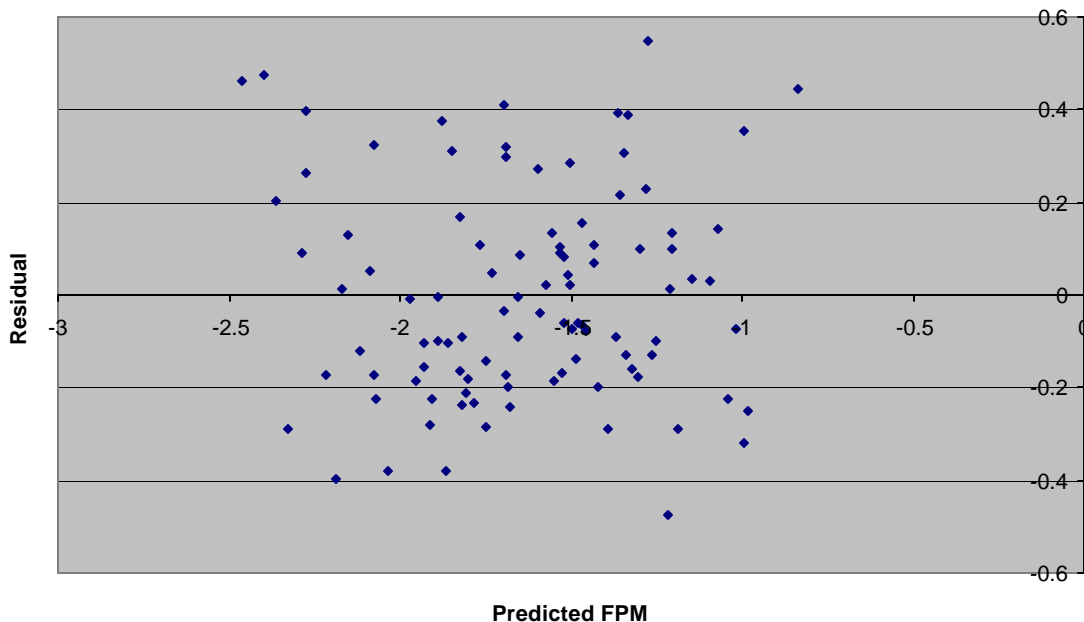


The simple linear regression residuals are plotted in Figure 5. It shows a clear curvilinear pattern, indicating that the regression specification is not adequate. To solve this problem, I performed another two regression analyses with different settings. In one of them I included the distance square term as one additional explanatory variable, and in the other one, I applied the logarithm-in-linear form – taking natural logarithm transformation on all the variables. Both cases produced improved results – increased R-square and well-behaved residuals pattern. The regression estimates and residual plot from the logarithm-in-linear model are showed in Table 2 and Figure 6.

Table 2 is consistent with Table 1 in many ways except for the intercept. The change of intercept from being significant in simple linear model to insignificant in the logarithm-in-linear model implies that the explanatory variables in the second model performed better than their counterparts in the first model. Table 2 also reveals valuable information about the relations between variables – the independent variables’ parameter estimates in the logarithm-in-linear form model can be interpreted directly as elasticities. For example, $\ln(\text{Distance})$ has an estimated coefficient of -0.5658 , which can be interpreted as 1% increase in the travel distance is associated with about 0.57% decrease in the yield.

Table 2 Regression Results for the Top 100 markets from Q0202, Dependent Variable: $\ln(\text{Yield})$			
	<i>Coefficients</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	-1.1578	-1.12	0.2668
$\ln(\text{Distance})$	-0.5658	-13.35	0.0000
$\ln(\text{Passengers})$	-0.0908	-1.26	0.2095
$\ln(\text{Largest Carrier Share})$	0.1159	1.33	0.1884
$\ln(\text{Origin Size})$	0.1278	2.72	0.0077
$\ln(\text{Destination Size})$	0.1971	4.11	0.0001
Adjusted R Square	0.7071		

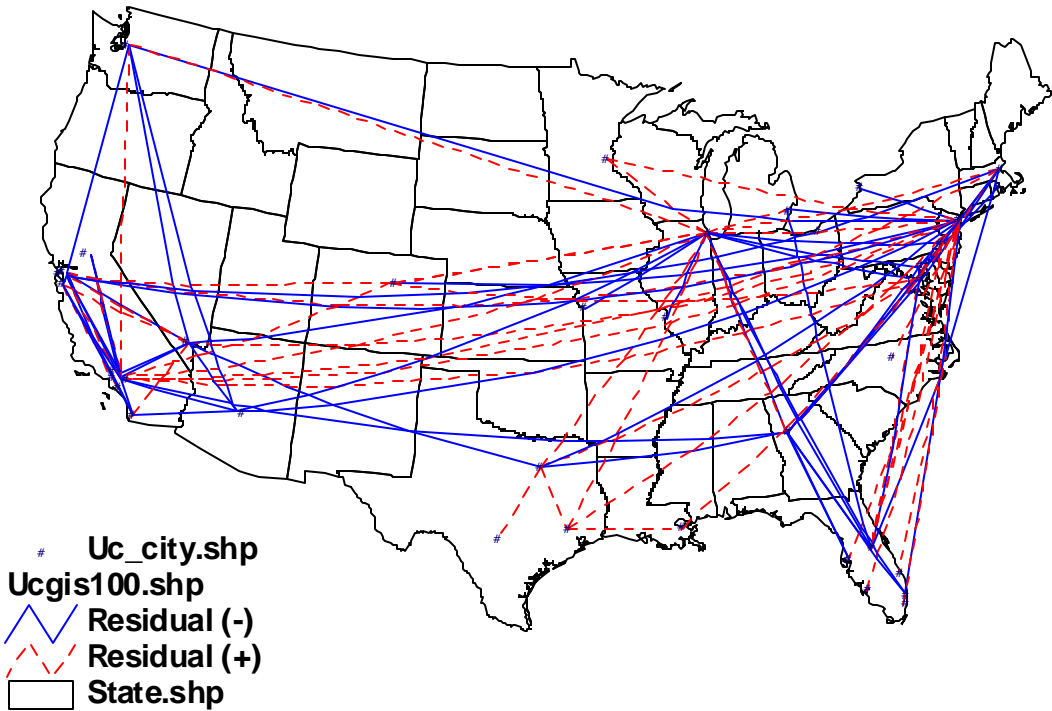
Figure 6 Residuals Plot (logarithm)



The residuals plot of the logarithm-in-linear regression (Figure 6) does not show any clear pattern, but this does not necessarily mean there is no spatial dependence structure. Since the residuals are associated with links in a network context, a geographic display of the residuals may

help to detect patterns. Figure 7 shows the top 100 largest city-pair markets displayed in different line patterns according to the regression residuals associated.

Figure 7 Residuals Plot for the top 100 markets



Although the overall pattern is a mixture, a few clusters could still be identified. For example, all the flights connecting Phoenix have negative residuals, while Houston has all positive ones. This geographic demonstration reveals the spatial structure of the regression residuals that otherwise cannot be discerned from the residuals plot. It further confirms our previous speculation about the existence of airline network autocorrelation. Visual interpretation of pattern or spatial clustering, however, only provides a qualitative and intuitive insight, in order to quantify this information and statistically verify the existence of spatial dependence, autocorrelation statistics needs to be calculated.

To statistically test the existence of network autocorrelation in air travel yield. I used Moran's I (Moran, 1948) to examine the regression residuals from the previous logarithm-in-linear regression. Moran's I is the most widely used measure for assessing spatial autocorrelation. When applied to OLS regression residuals, the test statistic I is calculated (Cliff and Ord, 1972 1981)

$$I = \frac{n}{S_0} \frac{e'We}{e'e}$$

where e is an N by 1 vector of regression residuals, W is the contiguity or spatial weights matrix.

The expectation and variance of the test statistic are

$$E(I) = \frac{n \cdot \text{trace}(MW)}{(n-k)S_0} \quad \text{where } M = (I - X(X'X)^{-1}X')$$

$$\text{VAR}(I) = \frac{n^2}{S_0^2(n-k)(n-k+2)} (\text{trace}(MWMW') + \text{trace}(MW)^2 + (\text{trace}(MW))^2) - E(I)^2$$

Table 3 Moran's I Testing Results				
I	E(I)	VAR(I)	ZI	p-value
0.223	-0.022	0.002	5.682	0.0000

Table 3 gives the results from the Moran's I testing. The Z score of 5.682 indicates a significant positive network autocorrelation among the regression residuals. The existence of positive network autocorrelation is quite reasonable if we examine the airfares interacting factors closer. Flights leaving from the same origin share the same airport facilities, have the same condition in weather, accessibility to the airport, regional economy, and other factors that may interplay with the costs but cannot be easily measured and included in the model. This part of costs is associated with local characteristics, therefore similar for all outgoing flights. When passed on to the passengers in form of fare, it works similarly on all the routes connecting to the airport. It reflects the variation in the air travel yield that cannot be explained by the model.

The existence of network autocorrelation violates the OLS assumptions, therefore may cause biased or inefficient results. To remedy this problem, a spatial regression model with interdependence structure explicitly incorporated was proposed. Since there exists significant autocorrelation among regression residuals, the autocorrelated error model is applied. This model deals with the situation in which there is spatial error dependence, and the regression error term follows a spatial autoregressive process. The autocorrelated error model is formulated as

$$y = X\mathbf{b} + u$$

$$u = \mathbf{r}Wu + \mathbf{e}$$

It can be considered as the combination of a standard regression model and a spatial autoregressive model in the error term u . The other error term \mathbf{e} is assumed to be well-behaved ($E(\mathbf{e}) = 0, E(\mathbf{e}\mathbf{e}') = \mathbf{S}^2I$). W is the spatial weight matrix.

The model can be rewritten as

$$Y = X\mathbf{b} + \mathbf{r}Wu + \mathbf{e}$$

$$= X\mathbf{b} + \mathbf{r}W(Y - X\mathbf{b}) + \mathbf{e}$$

$$= X\mathbf{b} + \mathbf{r}WY - \mathbf{r}WX\mathbf{b} + \mathbf{e}$$

Hence Y is expressed as a response to several influences: the general trend in Y explained by independent variables ($X\beta$), the surrounding Y (ρWY), and the neighboring trend ($\rho WX\beta$).

Due to the presence of spatial dependence, the usual framework for estimation that is based on a random sample of independent observations cannot be used for spatial regression models.

Estimation of spatial regression models is typically carried out by means of a maximum likelihood approach, in which the probability of the joint distribution (likelihood) of all observations is maximized with respect to a number of relevant parameters. The spatial dependence is formally incorporated in the joint probability density of the observations.

The autocorrelated error model has a nonspherical error variance. When the autoregressive coefficient ρ is known, the model could be estimated by means of generalized least squares (GLS). However, ρ is typically unknown, and needs to be estimated together with the other parameters (the β and σ^2). Therefore, an iterative procedure involving estimated generalized least squares (EGLS) is used instead. First the β coefficients are estimated conditional upon the ρ , and the ρ are estimated conditional upon the β , and the procedure switches back and forth until convergence is achieved. (For technical details, see Anselin, 1988).

Table 4 shows the results of the autocorrelated error model estimation.

	<i>Coefficients</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	-0.312849	-1.009	0.313
Ln(Distance)	-0.000000	-17.053	0.000
Ln(Passengers)	-0.028796	-2.186	0.029
Ln(Largest Carrier Share)	0.031189	2.155	0.031
Ln(Origin Size)	0.002487	3.025	0.002
Ln(Destination Size)	0.000310	3.607	0.000
Adjusted R Square	0.7749		

Table 4 shows some interesting results when compared with Table 2. First the adjusted R-square increased, indicating the spatial regression model performs better than the original one in explaining the variations in the dependent variable. Secondly, the coefficients for all the explanatory variables have dropped sharply, and all become significant at 0.05 level. This means that after the spatial error dependence is properly taken account, the influences from explanatory variables have decreased in terms of magnitude. It implies that the previous OLS model inflated the importance of the explanatory variables due to the spatial autocorrelation effects.

CONCLUSION AND DISCUSS

Through descriptive and statistical analyses of the US domestic air travel data, this paper confirms some conventional wisdom in the airfares research. It verifies that the city-pair market yield is affected by factors such as distance, scale of economies and competition. The market level air travel yield has a significant positive network autocorrelation over the airline service network. The spatial dependence among the links violates the assumptions of the OLS procedure commonly used in regression analysis, leading to biased estimation results. A spatial regression model integrated with the spatial dependence structure proves to be a remedy for this problem. After taking the network autocorrelation effects into account, the autocorrelated error model applied not only achieves a better explaining power, but also shows that the simple linear regression model substantially overestimated the influences of the explanatory variables. This is a clear illustration of the misleading conclusions one can make when regression analysis fails to address the spatial structure inherent in observations. Caution should always be used when dealing with spatial data.

The spatial weights matrix W is an important component in spatial autocorrelation statistics testing and spatial regression models for the fact that all the parameter estimations and inferences based on spatial autocorrelation hinge upon the specification of the particular W . Commonly, the weights matrix reflects simple contiguity. A binary number is used in each element to represent neighborhood structure. Contiguity can also be defined as a function of the distance, or any other values that reflect the potential interaction between spatial units of observations. In practice, it is often necessary to test several different settings of the spatial weights matrix. In this paper, I assume all the links are homogeneous in their mutual influence, and use a simple binary number to represent neighborhood in W elements. Network autocorrelation is a special kind of spatial autocorrelation. The basic spatial units of observation are links. Unlike points or areas with distance as an intuitive measure of interaction, links do not have a ready-to-use attribute to measure their "closeness". How to derive an appropriate measure reflecting the interaction between links would be an important issue in the next step research.

Reference Cited

1. Anderson, W.P., Gong, G., and Lakshmanan, T.R., 2002. Geographical Variation in the Cost of Air Travel: Analysis of the Domestic Airline Fare Consumer Report. *Transportation Research Record* 1788, 13-18.
2. Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.
3. Bailey, E.E., Graham, D.R., and Kaplan, D.P., 1985. *Deregulating the Airlines*. MIT Press, Cambridge.
4. Black, W., 1992. Network autocorrelation in transport network and flow systems. *Geographical Analysis* 24, 207-222.
5. Borenstein, S., 1989. Hubs and high fares: dominance and market power in the US airline industry. *Rand Journal of Economics* 20, 344-365.
6. Call, G.D., and Keeler, T.E., 1985. Airline deregulation, fares, and market behavior: some empirical evidence, *Analytic Studies in Transport Economics*, ed. A.F. Daughety, 221-247. Cambridge University Press, Cambridge.
7. Cliff, A.D. and J.K. Ord (1981). *Spatial Processes: Models and Applications*. London: Pion.
8. Evans, W.N., and Kessides, I.N., 1993. Localized market power in the US airline industry. *The Review of Economics and Statistics* 75, 66-75.
9. Goodchild, M.F., 1987. Spatial autocorrelation. *Concepts and Techniques in Modern geography*. No. 48.
10. Griffith, D.A. (1987). *Spatial Autocorrelation*. Resource Publications in Geography, AAG.
11. Hurdle, G.J. et al., 1989. Concentration, potential entry, and performance in the airline industry. *Journal of Industrial Economics* 38, 119-139.
12. Moran, P., 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society* 10B, 243-51.
13. Morrison, S., and Winston, C., 1987. Empirical implications and tests of the contestability hypothesis. *Journal of Law and Economics* 30, 53-66.

Endnotes

ⁱ Yield, is defined as the ratio of fare to distance (fare per mile).

ⁱⁱ $\bar{f}_i = \sum_j (f_{ij} \frac{p_{ij}}{\sum_j p_{ij}})$ where f_{ij} (p_{ij}) is the yield (passengers) for market between i and j , and \bar{f}_i is the average yield

for airport i .