The Temporal and Spatial Dynamics of the USPS' Service Performance Scores over the Period 2011-2020

Margaret M. Cigno[1], Curtis E. Kidd Telemaque[2], and Soiliou D. Namoro[3]

Abstract

This paper has three objectives. First, it proposes a new score aggregation method that incorporates spatial (geographic) variations in the Postal Service's quarterly service performance data. Second, it combines the proposed composite score with data on market-dominant and competitive mail volume, clerk and mail handlers' labor, and city-carriers' labor to analyze the spatio-temporal dynamics of service-performance scores over the period 2011-2020. Finally, within the limits of the available data, the paper strives to measure the impact of the Covid 19 pandemic on the dynamics and the spatial structure of the Postal Service's service performance. The methods developed in the paper can be applied to any setting in which some score of interest, e.g., a performance score, is measured, possibly over time, for a set of operationally interconnected localities, and a composite score encompassing the spatial average and dispersion of the performance scores is needed along with other metrics to assess the overall performance of the network.

1. Introduction

For each market dominant product[4], the United States Postal Service (USPS) is required by law to measure and report its service performance against an annual target. The Postal Regulatory Commission, henceforth, the Commission, then makes a determination on whether the Postal Service is complying with the statutory requirements regarding service performance. Statutory requirements pertain to an aggregated national number for each product. Commission rules also require USPS to report service performance results below the product level for some categories of mail and on a quarterly basis. Service performance results are reported for each District, along with aggregated Area-level and National-level service performance scores, and volume weights. Variance scores measuring late mails' excess-time relative to standards are also reported.

A possible shortcoming of the reporting of service performance scores is that it does not account for the spatial variations in results at the district or area levels. The statutory requirements regarding the reporting of service performance measures are the level of service, described in terms of speed of delivery and reliability, on an aggregated basis, not on the variation across spatial entities, such as administrative districts. (Order 4697, July 5, 2018). A systematic account of inter-district or inter-area variation in performance scores is absent from the Postal Service's reports and the Commission's compliance reports.

Exploring the spatio-temporal dynamics of service performance provides information on improving overall service performance by identifying high and low performing geographical areas and specific factors effecting performance.

The present paper argues from both economic and policy perspectives that relative performance differences between administrative districts or postal areas are, or should, for a given performance standard, be part of the quality of service measurement. It views the district-specific service performance scores as inter-related dimensions -- spatial dimensions -- of a same ``good'', namely, quality, demanded by the public and supplied by the Postal Service, along with each mail product. For this reason and others that are discussed below, spatial

[1] Director of the Office of Accountability and Compliance, U.S. Postal Regulatory Commission. margaret.cigno@prc.gov
[2] Senior Analyst, U.S. Postal Regulatory Commission. curtis.kidd@prc.gov.
[3] Senior Economist, U.S. Postal Regulatory Commission. soiliou.namoro@prc.gov.
[4] USPS products are categorized as either market dominant or competitive based on specific criteria.

variations in the level of service quality should be deemed relevant to the assessment of service performance. Part of this variance is due to idiosyncratic geographic differences.

To fix the notations, the target will be denoted by τ. If P_j denotes the service performance score pertaining to District j, $P \equiv (P_1, \ldots, P_N)$ the performance score configuration over the set of N districts (for example all the postal districts of the nation), $\eta \equiv (\eta_1, \ldots, \eta_N)$, the corresponding vector of volume weights, with η_j denoting the proportion of (performance measurement) volume assigned to District j's score in total (performance measurement) volume,[5] the aggregate score for the N districts is $\bar{P}_\eta \equiv \sum_{j=1}^{N} \eta_j P_j$.

The Postal Service's assumed goal to get each district's score as close as possible to the target can be expressed as the effort to minimize under some budget constraint, the Euclidean distance,

$$\left|\left|P - (\tau, \ldots, \tau)\right|\right|_\eta^2 \equiv \sum_{j=1}^{N} \eta_j \left(P_j - \tau\right)^2, \qquad (1)$$

Between the performance configuration and the N-dimensional constant target configuration $(\tau, \ldots, \tau)$.[6] The decomposition

$$\sum_{j=1}^{N} \eta_j \left(P_j - \tau\right)^2 = \sum_{j=1}^{N} \eta_j \left(P_j - \bar{P}_\eta\right)^2 + \left(\bar{P}_\eta - \tau\right)^2 \qquad (2)$$

highlights the two inextricably related components of the objective function:[7] the effort to raise the aggregate score, i.e., to minimize $\left(\bar{P}_\eta - \tau\right)^2$, and the effort to minimize the spatial dispersion of the score, i.e., to minimize the variance $\sum_{j=1}^{N} \eta_j \left(P_j - \bar{P}_\eta\right)^2$. This analysis also suggests that assessing the overall service quality of the Postal Service entails looking not only at the individual scores and the aggregate scores, but also at the spatial dispersion of the scores.

The present paper proposes a methodology in which the incorporation of spatial variance in the composite score results from a re-weighting of the reported scores based on the performance measurement volume-weights reported by the Postal Service along with the performance scores. With the new weights, the aggregate (or composite) score automatically incorporates the coefficient of variation across the local (district-level) scores. This new composite score is applied to analyze the dynamics and the spatial structure of the reported scores over the period 2011-2020.

The impact of the Covid-19 pandemic is also considered in the defined framework. The identification of the causal effect of the Covid-19 pandemic on service performance is challenging in three respects: (i)the data pertaining to the Covid period is sparse compared to the previous period and, (ii) the assessment of the causal effect must disentangle the possible effect of the replacement of the old performance measurement system by a new one, and the intrinsic effect of the Covid-19 pandemic, and (iii) factors other than the previous may be mediating the effects.

The relevance of the methods developed in this paper extends to any situation in which there is a need for an overall assessment of a group of entities in which the performances of the members are reported as percentages. As a word of notice, this paper does not address the open-ended quality-relevant issues raised by

[5] The performance measurement volume is the volume determined as part of the service performance measurement system. It does not represent actual mail volume flowing to or departing from a postal district.

[6] The scores are in percentage.

[7] The equality $\sum_{j=1}^{N} \eta_j \left(P_j - \bar{P}_\eta\right)\left(\bar{P}_\eta - \tau\right) = \left(\bar{P}_\eta - \tau\right) \sum_{j=1}^{N} \eta_j \left(P_j - \bar{P}_\eta\right) = 0$ has been applied to obtain relation (2).

the dichotomy existing, in the case of the Postal Service, between market-dominant products and competitive products.

The paper focuses on SPFCM service performance and is organized as follows: In Section 2 some background on service performance measurement is provided. In Section 3, the proposed score aggregation method is discussed. The resulting composite score and the time evolution in this score is analyzed to stress the periods of relatively large dispersion among the district scores. This analysis is also performed for each area and the contrast between the time-evolution of the within-area and nation-wide dispersion in the scores is stressed. In Section 4, an econometric model for the dynamics of the performance scores, in which the new composite score plays a substantial role, is specified and estimated, and the estimation results are discussed. In Section 5, this model is used to study the causal effect of the Covid-19 pandemic on service performance. Section 6 concludes the paper.

2. Background on Service Performance Measurement and Data

In July 2018, the Postal Regulatory Commission, approved the replacement of the Postal Service's External First-Class (EXFC) service performance measurement system for market dominant products with an internal Service Performance Measurement (SPM) system. Both the legacy and the new system measure delivery performance against delivery service standards. Service standards represent time requirements (in days) for mail piece delivery set by the Postal Service. For example, the service standards pertaining to Single-Piece First-Class Mail (SPFCM), which have changed over time, are 2 days or 3-5 days. The Postal Service sets annual service performance targets for each product and service standard (mail type). These targets represent the percentage of time that the Postal Service will meet or exceed the given service standard. These targets are set by the Postal Service's Executive Leadership Team (ELT), with the Board of Governors approval.

The legacy system, operated by a third party, tracked test pieces injected into the mail delivery system on an end-to-end basis. It determined service performance of letter-shaped mail pieces by measuring the duration from the time a test piece enters the mail stream (via a postal facility ,collection box, post office, or lobby chute) to the time it was delivered to its final destination—typically a home or business address. The recorded duration was compared with the applicable standards to calculate the performance score as the proportion of mail pieces delivered on-time.

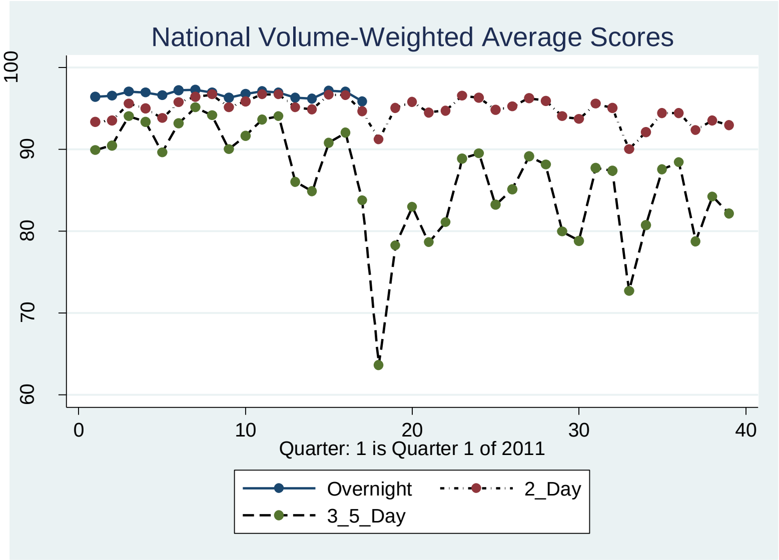
The new system does not track mail pieces end to end. For Single-Piece Letters/Cards that enter the mail stream via a collection receptacle, it combines samples over three stages of the delivery process: First Mile (collection), Processing Operation, and Last Mile (delivery). These performance scores are computed based on the delivery times recorded from these samples.

3. Adjusting the Aggregate Service Performance Scores for Spatial Dispersion

A thorough assessment of service-quality performance should include the national scores, discussion of the individual district scores, and the spatial dispersion among the district scores. Figure 1 displays the time (quarterly) evolution of the current national volume-weighted service performance scores for the 3 service-performance standards.[8] These time trends provide no information on how the scores are spatially distributed.

[8] In 2015 the Postal Service discontinued overnight service for FCSPM.

Figure 1. Nationally Aggregated Performance Scores over the 40 Quarters



Although the overall service-quality performance assessment could use a bi-variate metric (for example, the average and the standard error), a practical reason for not doing so is that score configurations pertaining to two different quarters may not be comparable. One could have a larger average score, but also a larger standard deviation compared to the other.[9]

Building a univariate metric that serves the same purpose should satisfy some minimal requirements. First, the resulting aggregate score should be derived as a weighted average of the initial scores and be comparable to the currently reported aggregate score. Second, the new aggregate score should remain invariant to whether the scores are expressed in decimals or percentage, i.e., the aggregate score must be invariant to a scaling of the score configuration under consideration. These two requirements can be satisfied by appropriately re-weighting the scores using both the score configuration and the score-measurement volumes, as shown in the next section.

3.1. Reweighting the District-Specific Service-Performance Scores

This section concentrates on the national weighted average score. The derived conclusions will apply to all levels of aggregation. The main benefit of reweighting the district-specific scores is to identify areas of poorer performances based on the presumption that an important objective of performance assessment is to invite more scrutiny on localized perturbations that result in below-average performance scores. The underlying rationale is the known fact that localized perturbation in a complex network can induce domino-like sequences

[9] A bi-variate aggregate measure will most likely generate a partial ordering over the set of all possible score configurations. In other words, assuming that a higher average score is better than a lower one and a lower standard deviation better than a higher one, two score configurations may not be comparable using the bivariate metric. For example, the one with higher average may also have a higher dispersion compared to the second. In contrast to this, a one-dimensional composite score would generates a total ordering over the set of score configurations: for every two score configurations, one has a larger composite score than the other or the two have equal composite scores.

of failures and cause major damage to how the overall network functions.[10] The reweighting strategy therefore involves the relative scores, i.e. the ratio of each score to the average score, which are used to scale the score measurement volume-weights up or down depending on whether the district has a below-average score or an above-average score, respectively.

Starting from the initial performance measurement volume-weights, $\eta_j, 1 \leq\ j \leq\ N$, new weights, $w_j(\eta, P), 1 \leq\ j \leq\ N$, are described by the mapping

$$\eta_j \mapsto\ w_j(\eta, P, a\}) \equiv \eta_j\ (a + G(P)P_j), \qquad (3)$$

where a is a real number assumed to satisfy $a > 1$ for all $N > 2$, and $G(P)$ is some suitable function of the performance configuration score P. In relation (3), the re-weighting factor, $a + G(P)$ is a two-part factor. The fixed part a represents the maximum factor by with the measurement weight can be scaled up. It is common to all districts and corresponds to the unlikely case where a district has zero performance. For District j the variable part of the scaling factor is the product of the function $G(P)$ by District j ‘s performance. The function $G(P)$ is assumed symmetric in its components and independent of j. In fact, the requirement that the new weights add up to 1, together with the condition that the function $G(P)$ be independent of j , assure that $G(P)$ depends only on the average score $\overline{P_\eta}$. Indeed,

$$\sum_j w_j(\eta, P, a) = \sum_j \eta_j(a + G(P)P_j) \qquad (4)$$

$$= (\ a\ + G(P) \sum_j \eta_j P_j = a + G(P)\overline{P_\eta} = 1, \quad (5)$$

which implies, $G(P) = \frac{1-a}{\overline{P_\eta}}$. (6)

Hence, the reweighting formula takes the form

$$\eta_j \longmapsto\ w_j(\eta, P, a) \equiv \eta_j(a - (a-1)\left(\frac{P_j}{\overline{P_\eta}}\right) = \eta_j\left[1 - (a-1)\left(\frac{P_j}{\overline{P_\eta}} - 1\right)\right]. \qquad (7)$$

The weight $w_j(\eta, P, a)$ will be positive if and only if $P_j < \frac{a}{a-1}\overline{P_\eta}$. The latter inequality determines the set of district scores that will be included in the assessment. Scores larger than the threshold score $\frac{a}{a-1}\overline{P_\eta}$ will not be considered, or will be considered high enough to be left aside of the aggregation. The remaining scores are simply normalized so as to sum to 1.

The mapping (7) can in fact be generalized by introducing a new parameter, $c > 0$, as follows:

$$\eta_j \longmapsto\ w_j(\eta, P, a, c) \equiv \eta_j\left(1 - \left(\frac{a-1}{c}\right)\left[\left(\frac{P_j}{\overline{P_\eta}}\right)^c - \frac{\overline{P_\eta^c}}{(\overline{P_\eta})^c}\right]\right). \qquad (8)$$

where the following additional notation is used for moments of order c: $\overline{P_\eta^c} \equiv \sum_j \eta_j P_j^c$. It can easily be verified that the right-hand side of (8) sums to 1 over the indices j and it is positive if and only if

[10] See for example, Li Daqing, Jiang Yinan, Kang Rui, and Shlomo Havlin (2014). ``Spatial correlation analysis of cascading failures: Congestions and Blackouts,'' Sci Rep. 2014; 4: 5381, Published online 2014 Jun 20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4064325/

$$P_j < \bar{P}_\eta \left[\frac{c}{a-1} + \frac{\overline{P^c_\eta}}{(\bar{P}_\eta)^c}\right]^{1/c} \qquad (9)$$

The case $c = 1$ corresponds to (7). To fix ideas, the rest of the paper concentrates on the case $c = 1$.

The weight $w_j(\eta, P, a)$ in (7) is a one-parameter family of weight vectors, depending on the parameter a, which will be specified by imposing the following two additional requirements: (i) the parameter a is proportional to the total number of districts, i.e., $a = kN$, for some positive integer k, and (ii) $w_j(\eta, P, a) = \eta_j$ for $N = 2$.

The condition (i) has the meaning that the greater the number of districts included in the service performance assessment the more complex the system is to assess and the larger the weights assigned to below-average performance scores will be. .[11] To the extent that the number N of districts approximates the complexity of the service performance network, condition (i) is guided by the fact that localized failures of a complex network are usually not immediately apparent, hence the need to magnify their weight in the aggregate performance in order to help identify them. Condition (ii) is motivated by the conjecture that for a system of only two nodes, score averaging may not even be necessary to assess the system's performance.

The conditions (ii), i.e., $\left[N = 2 \Rightarrow \text{for all } j, \left(w_j(\eta, P, a) = \eta_j\right)\right]$, is equivalent to $[N = 2 \Rightarrow \text{for all } j, 1 - (a-1)\left(\frac{P_j}{\bar{P}_\eta} - 1\right) \equiv 1]$, which implies $a = 1$ for $N = 2$. Together, (i) and (ii) imply $a = \frac{N}{2}$. The resulting weights are

$$w_j = w_j(\eta, P, N/2) = \eta_j \left[\frac{N}{2} - \left(\frac{N}{2} - 1\right)\left(\frac{P_j}{\bar{P}_\eta}\right)\right] = \eta_j \left[1 - \left(\frac{N-2}{2}\right)\left(\frac{P_j}{\bar{P}_\eta} - 1\right)\right]. \qquad (11)$$

Figure 2. Illustration of the Re-Weighting

[11] The derivative of $w_j(\eta, P, a)$ with respect to a is equal to $1 - \frac{P_j}{\bar{P}_\eta}$, and it is positive if $P_j < \bar{P}_\eta$. Hence, the weights assigned to below-average performance scores increase with a.

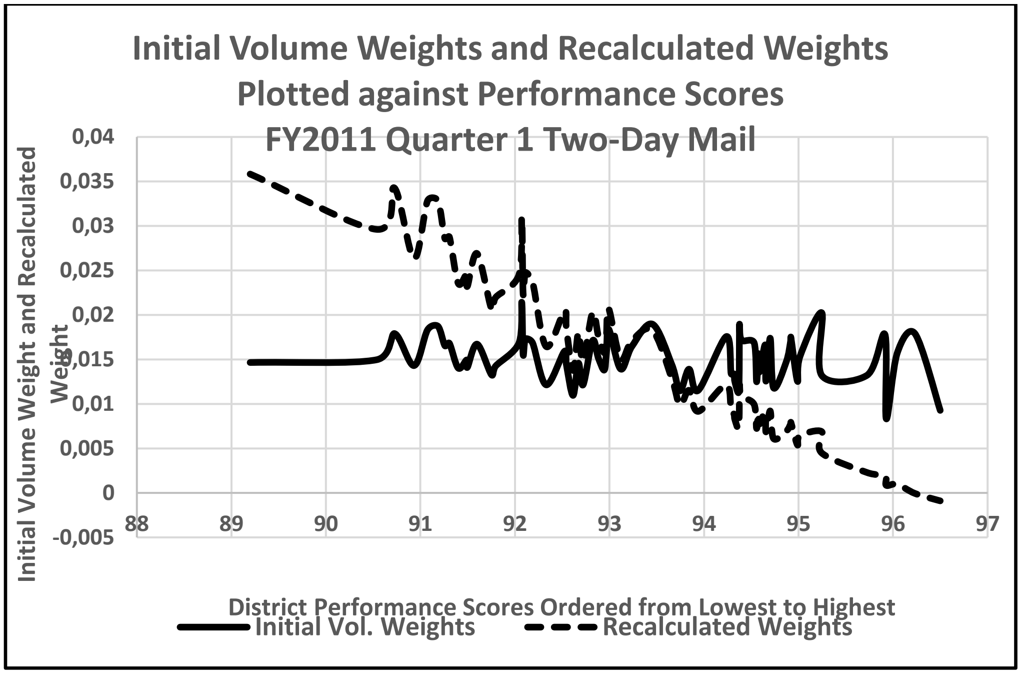


Figure 2 displays the plot of initial volume weights and recalculated weights (the performance measurement weights) against the district performance scores, shown on the axis in the order of the lowest to the highest, for Two-day mail and for FY 2011, Quarter 1. The initial volume weights and the recalculated weight are shown respectively in plain curve and dotted curve. As the figure shows, the re-weighting operation is a point-wise re-scaling and clockwise rotation of the score measurement weight curve. The weighted national average in the considered case is 93.3. The lower a performance score is relative to the national average, the larger the new weight assigned to it will be. Likewise, the larger a performance score is relative to the national average, the lower the new weight assigned to it will be.

3.2. The Mean-Variance Composite Score

The aggregate score corresponding to the reweighting, denoted by $MV(P)$, is

$$MV(P) = \sum_{j=1}^{N} w_j P_j = \bar{P}_\eta \left[1 - \left(\frac{N-2}{2}\right) C_v^2\right] \qquad (12)$$

where C_v^2 denotes the squared coefficient of variation of the scores computed using the measurement volume-weights, i.e., $C_v^2 = \frac{Var(P)}{\bar{P_\eta}^2}$, and $Var(P) = \overline{P_\eta^2} - \left(\bar{P}_\eta\right)^2 \equiv \sum_j \eta_j P_j^2 - \left(\sum_j \eta_j P_j\right)^2$. (13)

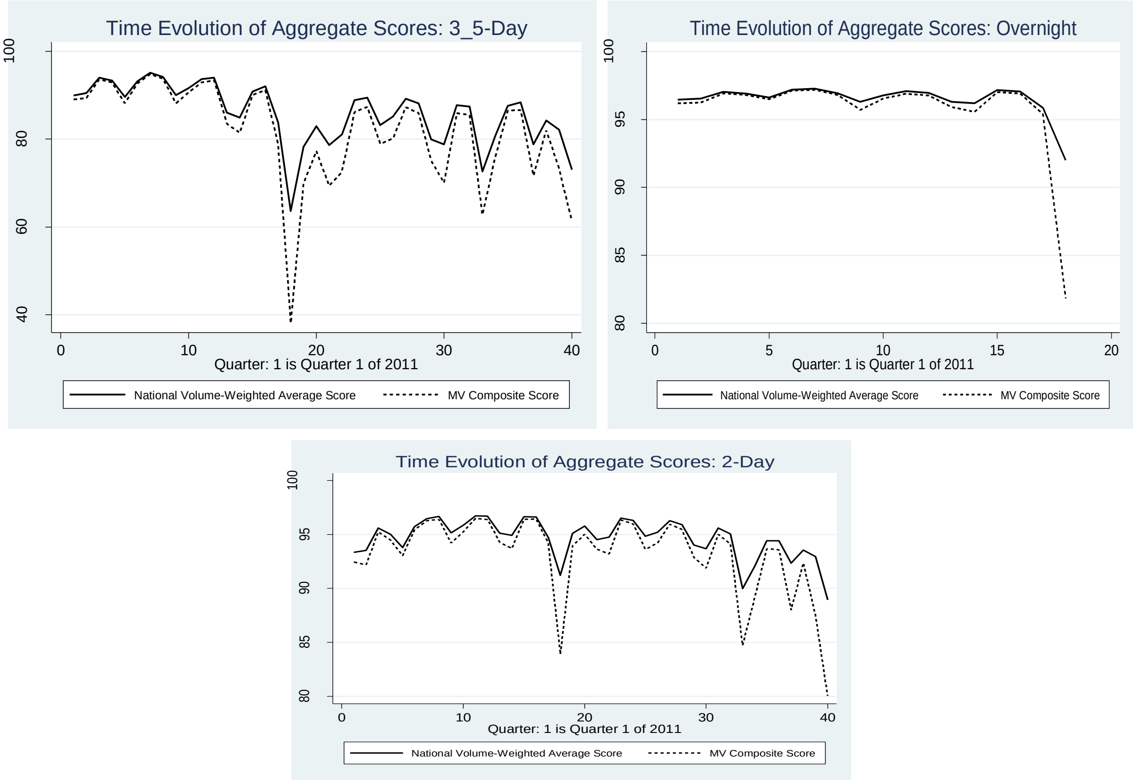
The aggregate score $MV(P)$ will be referred to as the Mean-Variance composite Score, in short, the MV score, for it incorporates both the mean and the coefficient of variation. It represents a downscaled version of the reported national aggregate score. It discounts more or less the reported aggregate score when the coefficient of variation increases or decreases.[12]

[12] The ranking of service performance score configurations based on the MV score falls into the general setting of two-moment decision models. The consistency of these models is discussed in " Jack Meyer (1987) “Two-Moment Decision Models and Expected Utility Maximization,” *The American Economic Review,* Jun., Vol. 77, No. 3 (Jun., 1987), pp. 421- 430.,

The *MV* score can be interpreted as measuring how well the operator has managed its resources to increase the aggregate national score, while reducing the dispersion among the district performance scores. The higher the *MV* score is the more it reflects a better allocation of performance-relevant resources among the geographically dispersed nodes of the network. This interpretation should be made, however, conditional on uncontrollable factors, such as geographic characteristics and weather conditions. Alternatively, the *MV* score can be interpreted as measuring how geographically integrated the performance scores are, given the level of relevant resources that are spent. In this interpretation, a higher *MV* score is evidence of a greater geographic integration of management efforts devoted to service performance.

The magnitude of the scale factor, $1-\left(\frac{N-2}{2}\right)C_v^2$, can be interpreted as a penalization.

Figure 3. Quarterly Evolution of the Reported and the MV Scores



The 3 graphs in Figure 3 suggest that the gap between the quarterly reported national volume-weighted average score and the MV composite score is the largest for 3-5-day mail, followed by 2-day mail. Overnight mail appears to have performed best in that respect before its discontinuation.

A quantitative comparison can be made by first noting the equality:

$$\frac{MV(P)-\overline{P_\eta}}{\overline{P_\eta}} = -\left(\frac{N-2}{2}\right)C_v^2 \qquad (14)$$

and Haim Levy (1989) "Two-Moment Decision Models and Expected Utility Maximization: Comment," *The American Economic Review* , Jun., 1989, Vol. 79, No. 3 (Jun., 1989), pp. 597- 600.

Using the logarithmic approximation $\frac{MV(P)-\overline{P_\eta}}{\overline{P_\eta}} \approx \ln\left(\frac{MV(P)}{\overline{P_\eta}}\right)$, the average rate of penalization over the time period considered, expressed in percentage, can be defined for the service standard *S*, π_S, as

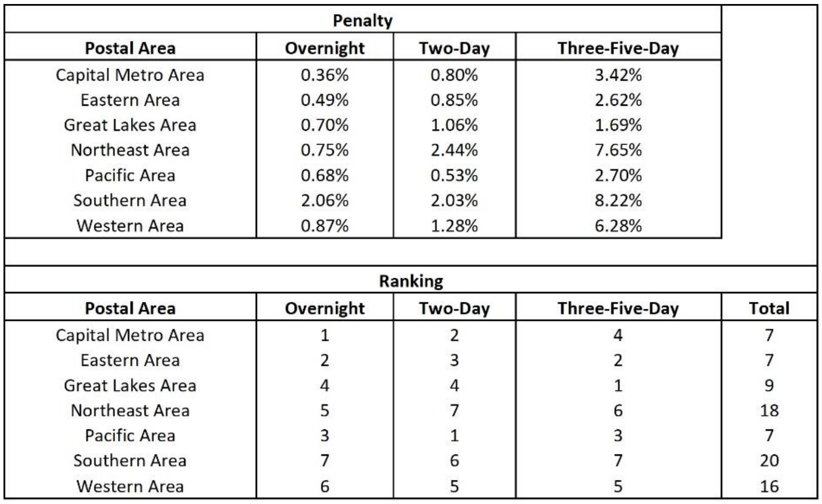
$$\pi_S = -100\frac{\sum_t\left(\frac{MV(P)-\overline{P_\eta}}{\overline{P_\eta}}\right)_t}{T} = 100\left(\frac{N-2}{2}\right)\frac{\sum_t C_v^2(t)}{T} \approx -100\frac{\sum_t \ln\left(\frac{MV(P)}{\overline{P_\eta}}\right)_t}{T} \qquad (15)$$

where T denotes the total number of quarters.[13] A larger π_S means a larger penalization applied to the aggregate national score to reduce it. Hence, a lower π_S value is evidence of a smaller spatial dispersion of the performance scores. The calculated penalization magnitudes are:

$\pi_{overnight} = 0.7\%, \pi_{2-day} = 1.5\%$, and $\pi_{3-5-day} = 4.0\%$.

The calculations can, of course, also be conducted at the postal area level, which allows comparisons to the 7 postal areas. Using the total rank as a summary of the ranking on the 3 service standards, the top performing group includes the Capital Metro Area, the Eastern Area, and the Pacific Area, which have, each, a total rank equal to 7. The remaining areas are in the following order of performance: Great Lakes Area, Western Area, Northeast Area, and Southern Area.

Table 1. Area-Level Penalty



	Penalty		
Postal Area	Overnight	Two-Day	Three-Five-Day
Capital Metro Area	0.36%	0.80%	3.42%
Eastern Area	0.49%	0.85%	2.62%
Great Lakes Area	0.70%	1.06%	1.69%
Northeast Area	0.75%	2.44%	7.65%
Pacific Area	0.68%	0.53%	2.70%
Southern Area	2.06%	2.03%	8.22%
Western Area	0.87%	1.28%	6.28%

		Ranking		
Postal Area	Overnight	Two-Day	Three-Five-Day	Total
Capital Metro Area	1	2	4	7
Eastern Area	2	3	2	7
Great Lakes Area	4	4	1	9
Northeast Area	5	7	6	18
Pacific Area	3	1	3	7
Southern Area	7	6	7	20
Western Area	6	5	5	16

4. An Econometric Model for the Dynamics of Service Performance

The departure point of the present section is the interpretation of the *MV* score as the degree of geographic or spatial integration of the performance scores, given the performance relevant resources committed. However, the specific way in which the interpretation is used requires consideration of time. The first quarter of Fiscal year 2011 is interpreted as time zero. At each subsequent quarter, say t, the history of the *MV* score up to time t describes the pattern of integration between the performance scores over that period [0, t].

The network nature of the Postal Service suggests that the district-level scores are statistically interrelated, both temporally and spatially. The modelling assumption that will be maintained throughout the section is that once their interrelation is controlled for by the history of the *MV* scores, the scores behave statistically as if, at each

[13] The subscript t is applied to a parenthesis to indicate that all variables in the parentheses pertain to the quarter t. Also, $C_v^2(t)$ denotes the squared coefficient variation corresponding to quarter t.

given time, they are spatially independent from each other. Stated alternatively and more precisely, the assumption means that at each given time, given the history of the *MV* scores and a set of control variables to be listed, any group of scores is independent of the history (up to that time) of the remaining scores (non-members of the group).

4.1. The Model

The assumption described in the introduction to this section is now made formally precise: If X is a (column) vector of conditioning (or control) variables, the assumptions that the history of the *MV* scores summarizes the spatial interplay between the district-level scores is expressed by the following 3 relations in which, α, $\lambda_1, \dots, \lambda_p$, β, γ, δ, are parameters and ϵ_{tj} and ϑ_t are error terms:

$$(\ P_L \ )_t \underline{\perp\!\!\!\perp} \ (P_K)_t, (P_K)_{t-1}, \dots, (P_K)_0 \mid MV(P)_t, \ MV(P)_{t-1,}, MV(P)_{t-2}, \dots \ , \ MV(P)_{0,} \ , \ X_t \text{ , for all } L, K \tag{16}$$

where $(P_L)_t$ is a group scores taken at time *t,* and $(P_K)_t, (P_K)_{t-1}, \dots, (P_K)_0$ the time history of the remaining scores (non-members of the first group), and $MV(P)_t, MV(P)_{t-1,}, MV(P)_{t-2}, \dots \ , \ MV(P)_{0,}$ is the history of the *MV scores*.

$$P_{tj} = \alpha + \beta MV_t(P) + X'_{tj}\gamma + \epsilon_{tj} \tag{17}$$

$$MV_t(P) = \delta + \lambda_1 MV_{t-1}(P) + \cdots + \lambda_p MV_{t-p}(P) + \ \vartheta_t \tag{18}$$

The equations (16) and (17) describe the dynamics of the score configuration. Relation (16) states that given the history of the *MV* scores, the scores group $(P_L)_t$ does not depend on the history of the remaining scores non-members of (P_L) .

Equation (17) states that at each time t, District $j's$ score, P_{tj}, is a linear function of the time-t MV score and the control variables.[14] Equation (18) states that the *MV* score's dynamics is linear and Markovian, i.e., it is a linear autoregressive process or a given order denoted by p.

Combining (17) and (18), one obtains

$$P_{tj} = \alpha_0 + \alpha_1 MV_{t-1}(P) + \cdots + \alpha_p MV_{t-p}(P) + X'_{tj}\gamma + u_{tj}, \tag{19}$$

where α_0, α_1 are functions of the previous parameters and u_{tj} is a linear combination of ϵ_{tj} and ϑ_t.

Model (19) has a panel-data structure and the error u_{tj} will be assumed to be the sum of a district-specific effect, ξ_j, and an idiosyncratic error ν_{jt}: $u_{tj} = \xi_j + \nu_{jt}$.

The assumptions (16)-(18) have the following implications:

(i) The prediction of a district-level performance score for time *t, i.e.,* P_{tj}, only depends on the history of the *MV* score and the control variables to be listed.

(ii) The dependence in (i) on the history of the *MV* scores and the control variables is linear.

[14] It is important to note here that the *MV* score is likely endogenous in (17) since it is built from all the scores, including P_{tj}.

(iii) The P_{tj} may still be dependent over time.

These implications are summarized in model (19), the estimation of which will assume that the order of the autoregressive process, p, is equal to 2.

4.2. Variable Description and Model Estimation

The estimation of the model (19) hinges upon the availability of control variables that are both time and district dependent. Actual mail volumes, not the volumes included in the performance score measurement, are the main control variables and they all depend on time only. Fortunately, the measurement of the score configuration is conditioned on measurement volumes that are estimated by the Postal Service. Hence, the modelling of the distribution of the score configuration must take these volumes as given, even though they will likely be uncorrelated with the scores.[15] Their inclusion assures that at least one explanatory variable is both district and time dependent. As a byproduct, the claim that these measurement volumes have no effect on the scores becomes a statistically testable assumption.

The volume variables are the actual market-dominant FCSPM total volume and the competitive volumes: Express, Priority, Return, and International. The choice of actual mail volumes as explanatory variables is motivated by the assumption that higher volume may put more pressure on the delivery network and may, therefore, reduce service quality. Volumes, however, are handled by labor and the Postal Service is a labor-intensive network. Yearly work hours are also controlled for--specifically, yearly total hours for clerks and mail handlers are included, as well as work hours for City Delivery Carriers and Vehicle Service Drivers.

Table 2 displays the summary statistics of the variables involved in the model. The estimation is performed separately for each service standard. The specified model is a fixed-effect linear panel data model and it is estimated using the STATA command *xtreg* with robust standard error.

In the estimation, each of 67 districts is observed over the number of quarters for which data are available. This number is only 18 for overnight mail and 40 for 2-day and 3-5-day mail.[16] Measurement volume is the only variable that is both time and district dependent. All the other covariates are time-dependent only, some of which, namely labor variables, are only observed annually. The summary statistics are shown in Table 2. The estimation results are displayed in Table 3. To ease the interpretation of the results, the marginal effects per chosen units of change in the variables are summarized in Table 4, where the volume effects (of actual volumes) are measured per one-million pieces change.

[15] This lack of correlation can be viewed as a positive feature of the measurement system.

[16] There are a few missing data albeit very small in number due to small modification of area compositions in some quarters.

Table 2. Summary Statistics[17]

Variable	Obs	Unit	Mean	Std.	Min	Max
Overnight Performance Score	1,206	Percent	96.44	2.10	68.29	100.00
Overnight Mail Measurement Volume	1,206	Pieces	7947.32	2561.13	38	23637
2-Day Performance Score	2,660	Percent	94.61	2.78	67.00	99.19
2-Day Mail Measurement Volume	2,660	Pieces	13300000.00	29000000.00	2333	176000000
3-5--Day Performance Score	2,677	Percent	85.82	7.63	44.29	96.54
3-5--Day Mail Measurement Volume	2,677	Pieces	4959157.00	11300000.00	3151	90800000
Total First-Class Mail Volume	2,679	Thousands	15600000.00	1766865.00	12000000.00	19900000.00
Volume of Express	2,679	Thousands	9216.16	4224.62	5533.00	29763.00
Volume of Priority	2,679	Thousands	279198.50	150674.30	184188.00	1022959.00
Volume of Select	2,679	Thousands	602005.10	499188.40	74461.00	2796085.00
Volume of Return	2,679	Thousands	17642.26	10697.78	8715.00	69154.00
Volume of International	2,679	Thousands	70562.67	33278.02	37567.00	207404.00
City Delivery Carriers and Vehicle Service Drivers	2,679	Million Hours	14.46	2.25	12.30	18.30
Clerks and Mail Handlers	2,679	Million Hours	200.60	10.36	189.10	221.00

Table 3. Estimation Results

Variable	Overnight			2-Day			3-5-Day		
	Coef.	P>t		Coef.	P>t		Coef.	P>t	
Measurement Volume	0.0004271	0.00	***	0.0000000	0.00	***	-0.0000001	0.00	***
Total First-Class Mail Volume	0.0000045	0.00	***	0.0000006	0.00	***	0.0000033	0.00	***
Express	-0.0004704	0.01	**	0.0007279	0.00	***	0.0028752	0.00	***
Priority	0.0000168	0.17		-0.0000376	0.00	***	-0.0000983	0.00	***
Select	0.0000019	0.65		0.0000069	0.00	***	0.0000085	0.00	***
Return	-0.0000132	0.98		-0.0000277	0.33		-0.0001852	0.00	***
International	0.0000261	0.25		0.0000012	0.63		0.0000257	0.00	***
City Delivery Carriers	6.7897530	0.00	***	0.1797834	0.04	**	3.0769660	0.00	***
Clerks and Mail Handlers	-0.2437459	0.00	***	-0.0595082	0.00	***	-0.2102238	0.00	***
Quarter2	6.4271090	0.00	***	1.0717130	0.00	***	4.7555150	0.00	***
Quarter3	10.9260800	0.00	***	2.5307440	0.00	***	12.5128300	0.00	***
Quarter4	12.2306400	0.00	***	1.7311750	0.00	***	10.4984500	0.00	***
Lag 1 of MV	1.1664500	0.08	*	0.1683766	0.00	***	0.2964039	0.00	***
Lag 2 of MV	1.5027160	0.02	**	-0.0935849	0.00	***	-0.1115757	0.00	***
Constant	-284.7015000	**0.05**		86.8835700	0.00	***	6.7576230	0.09	**
Sample Size	1072			2528			2543		
R-sq	**Within**	0.4117		**Within**	0.3674		**Within**	0.6166	
	Between	0.0000		**Between**	0.0036		**Between**	0.0721	
	Overall	0.3395		**Overall**	0.3089		**Overall**	0.5595	
F	F(14,66)=23.61			F(14,66)=93.55			F(14,66)=212.58		
P>F	0.0000			0.0000			0.0000		

[17] The volume data are collected from the Postal Service's Revenue and Revenue, Pieces & Weight (RPW) quarterly reports, Financials - What we do - About.usps.com. The labor data are collected from the USPS Annual Tables, TFP (Total Factor Productivity). USPS Reports | Postal Regulatory Commission (prc.gov)

Table 4: Marginal Effects

Variable	Marginal Effect					
	Overnight		2-Day		3-5-Day	
	Significance	Effect	Significance	Effect	Significance	Effect
Measurement Volume	***	0.000	***	0.000	***	0.000
Total First-Class Mail Volume	***	0.004/Million	***	0.001/Million	***	0.003/Million
Express	**	-0.470/Million	***	0.728/Million	***	2.875/Million
Priority	No	0/000/Million	***	-0.038/Million	***	-0.098/Million
Select	No	0.000/Million	***	0.007/Million	***	0.009/Million
Return	No	0.000/Million	No	-0.028/Million	***	-0.185/Million
International	No	0.000/Million	No	0.001/Million	***	0.026/Million
City Delivery Carriers	***	6.790/Million	**	0.180/Million	***	3.077/Million
Clerks and Mail Handlers	***	-0.244/Million	***	-0.060/Million	***	-0.210/Million
Quarter2	***	6.427	***	1.072	***	4.756
Quarter3	***	10.926	***	2.531	***	12.513
Quarter4	***	12.231	***	1.731	***	10.498
Lag 1 of MV	*	1.166	***	0.168	***	0.296
Lag 2 of MV	**	1.503	***	-0.094	***	-0.112
Std. of the Performance Score		2.10		2.78		7.63

Statistical significance is indicated in Tables 3 and 4 with 3 stars for 1% level, 2 stars for 5% level, and 1 star for 10% level. As expected, from the within and between R-squares displayed in Table 3, it can be concluded that most of the explained variation in the scores is the variation over time. Table 4 shows that the lags in the *MV* scores are all significant, albeit the significance of the first lag is only at 10% for Overnight mail. They are strongly significant (1%) for 2-day and 3-5-day mail, which validates the assumptions (16)-(18) underlying the model. Tables 3 and 4 show that measurement volumes are all significant and their effects are, as expected, negligible compared to the standard deviation of the scores corresponding to the service standards, displayed in the last row of Table 4. These standard deviations are calculated using in each case the entire sample, i.e., the sample observed over the entire time period of the analysis.

Volume effects are seen in Table 4 to be negligible when they are compared to score standard deviations. With the exception of Express volume, which has a negative effect on Performance, all other competitive mail volumes have zero effect on overnight performance scores. The signs of the volume effects are consistently the same for 2-day and 3-5-day mail. Priority and Return volumes have a negative effect on service performance while Express, Select and International volumes display positive effects.

Table 4 display an interesting contrast between the effects of City-carrier labor and clerk and mail handler labor, both measured in millions of hours.[18] A one-million-hour increase in City-carrier labor has positive effect on performance scores while the same change in clerk and mail handler labor has a negative (though somewhat small) effect on performance scores.

[18] Recall that labor variables are measured annually in the data set.

Seasonal effects are all significant at 1% level and indicate that the first quarter (October, November, and December) is the most burdensome quarter by its effect on service performance in comparison to the other quarters. These effects are stronger for overnight and 3-5-day mails than for 2-day mails.

5. Gauging the Impact of the C-ovid-19 Pandemics on Service Performance

5.1. Change in the Service Performance Measurement System

Given that the measurement system for service performance was completely replaced by a new one from the first month of the last quarter of 2018, a first test to be conducted is whether there has been a significant change in service performance before and after the change in the performance measurement system. This test is carried out here by including in the models pertaining to 2-day and 3-5-day mail, a time dummy for post 2018 quarters. The estimation results are shown in Table 5. The change of the service performance measurement system has a negative effect significant at the 1% level. The magnitude of the effect is -3.048 percentage point for 2-day mail and -2.50 percentage points for 3-5-day mail. Whether this effect is causal or it is just a correlation is open to discussion.

5.2. The Covid-19 Effect

The World Health Organization (WHO) was informed of a cluster of cases of pneumonia of unknown cause detected in Wuhan City, Hubei Province of China on 31 December 2019.[19] By mid-April, all fifty U.S. states had confirmed cases, which, arguably, can be taken as the beginning of the general awareness in the U.S. of the enormous risks (health, business, and others) created by the virus. Consequently, the present paper considers post- April 2019 to be the covid-19 period to be investigated. The statistical question that will be investigated here is whether, given the change that occurred in the performance measurement system, there has been any additional significant break in the service performance scores after the 3d quarter of 2019 until the end of 2020. This test will also be performed by introducing a second time dummy for the 4th quarter of 2019 until the 4th quarter of 2020.

Table 6 shows the estimation results when the replacement of the service measurement system is not controlled for in the test for a Covid-19 effect. The dummy for Covid-19 is seen to have no significant in effect on 2-day service performance in contrast to 3-day service performance on which it has a negative effect equal to -1.62 (to be compared to 7.63 standard deviation of 3-5-day performance scores).

[19] See https://web.archive.org/web/20200202151307/https://www.who.int/westernpacific/emergencies/novel-coronavirus.

Table 5. Test for the Effect of Service Measurement System Replacement.

	2-Day			3-5-Day		
Variable	**Coef.**	**P>t**		**Coef.**	**P>t**	
Measurement Volume	0.0000000	0.22		0.0000000	0.04	**
Total Fist-Class Mail Volume	0.0000005	0.00	***	0.0000033	0.00	***
Express	0.0008323	0.00	***	0.0029913	0.00	***
Priority	-0.0000303	0.00	***	-0.0000906	0.00	***
Select	0.0000062	0.00	***	0.0000074	0.00	***
Return	-0.0001553	0.00	***	-0.0003099	0.00	***
International	-0.0000003	0.91		0.0000241	0.00	***
City Delivery Carriers	0.5302491	0.00	***	3.4496680	0.00	***
Clerks and Mail Handlers	-0.0742214	0.00	***	-0.2238962	0.00	***
Quarter2	1.2427660	0.00	***	4.9426650	0.00	***
Quarter3	2.6667690	0.00	***	12.6708800	0.00	***
Quarter4	1.9414300	0.00	***	10.6852900	0.00	***
System Change	-3.0478630	0.00	***	-2.4993580	0.00	***
Lag 1 of MV	0.1390870	0.00	***	0.2928262	0.00	***
Lag 2 of MV	-0.1075448	0.00	***	-0.1125535	0.00	***
Constant	89.3203200	0.00	***	5.3435700	0.19	
Sample Size	2528			2543		
R-sq	**Within**	0.3809		**Within**	0.6180	
	Between	0.0101		**Between**	0.0998	
	Overall	0.3189		**Overall**	0.5628	
F	F(15,66)=102.26			F(15,66)=256.64		
P>F	0.0000			0.0000		

Table 6: Table 6. Test for the Effect of Service Covid-19, Unconditional on Measurement System Replacement.

	2-Day			3-5-Day		
Variable	**Coef.**	**P>t**		**Coef.**	**P>t**	
Measurement Volume	0.0000000	0.00	***	-0.0000001	0.00	***
Total Fist-Class Mail Volume	0.0000006	0.00	***	0.0000030	0.00	***
Express	0.0006963	0.00	***	0.0031869	0.00	***
Priority	-0.0000375	0.00	***	-0.0000988	0.00	***
Select	0.0000070	0.00	***	0.0000084	0.00	***
Return	-0.0000135	0.73		-0.0003311	0.00	***
International	0.0000006	0.82		0.0000303	0.00	***
City Delivery Carriers	0.1500445	0.11		3.3796080	0.00	***
Clerks and Mail Handlers	-0.0588858	0.00	***	-0.2172191	0.00	***
Quarter2	1.0877610	0.00	***	4.6293330	0.00	***
Quarter3	2.5774500	0.00	***	12.1247300	0.00	***
Quarter4	1.8013710	0.00	***	9.8930810	0.00	***
Covid-19	0.1672463	0.69		-1.6194820	0.00	***
Lag 1 of MV	0.1656716	0.00	***	0.3023431	0.00	***
Lag 2 of MV	-0.0903941	0.00	***	-0.1194328	0.00	***
Constant	86.4700500	0.00	***	10.1476500	0.01	**
Sample Size	2528			2543		
R-sq	**Within**	0.3675		**Within**	0.6172	
	Between	0.0604		**Between**	0.0748	
	Overall	0.3089		**Overall**	0.5603	
F	F(15,66)=87.31			F(15,66)=205.09		
P>F	0.0000			0.0000		

When the replacement of the service measurement system is controlled for, its effect remains negative and significant at 1% level across the two service standards. In Table 7, the Covid-19 dummy is as in Table 6, with a slight change in its magnitude for 3-5-day mail, from -1.62 to -1.40.

Under the maintained assumptions of the model and after controlling for the replacement of the service measurement system by a new one, it can be concluded from the above results that the data provide some evidence for a Covid-19 effect on 3-5 service performance scores, while providing no evidence on a Covid-19 effect on 2-day service performance scores. Here also, whether these effects are causal or they express simple correlation is open to discussion.

Table 7. Test for the Effect of Service Covid, Conditional on Measurement System Replacement.

	2-Day			3-5-Day		
Variable	**Coef.**	**P>t**		**Coef.**	**P>t**	
Measurement Volume	0.0000000	0.22		0.0000000	0.04	**
Total Fist-Class Mail Volume	0.0000007	0.00	***	0.0000030	0.00	***
Express	0.0007331	0.00	***	0.0032535	0.00	***
Priority	-0.0000298	0.00	***	-0.0000914	0.00	***
Select	0.0000062	0.00	***	0.0000074	0.00	***
Return	-0.0001129	0.01	**	-0.0004289	0.00	***
International	-0.0000021	0.42		0.0000281	0.00	***
City Delivery Carriers	0.4433400	0.00	***	3.6902060	0.00	***
Clerks and Mail Handlers	-0.0725928	0.00	***	-0.2291832	0.00	***
Quarter2	1.2989430	0.00	***	4.8238980	0.00	***
Quarter3	2.8209480	0.00	***	12.3281100	0.00	***
Quarter4	2.1733030	0.00	***	10.1538400	0.00	***
System Change	-3.1263920	0.00	***	-2.3645310	0.00	***
Covid-19	0.5395451	0.20		-1.3947430	0.01	**
Lag 1 of MV	0.1296059	0.00	***	0.2981342	0.00	***
Lag 2 of MV	-0.0976109	0.00	***	-0.1192675	0.00	***
Constant	88.0490600	0.00	***	8.3394340	0.03	**
Sample Size	2528			2543		
R-sq	**Within**	0.3814		**Within**	0.6184	
	Between	0.0101		**Between**	0.1033	
	Overall	0.3193		**Overall**	0.5632	
F	F(16,66)=96.24			F(16,66)=244.07		
P>F	0.0000			0.0000		

Factors that may have contributed to mask a Covid-19 effect on 2-day the fact that competitive volumes have soared in comparison to pre-Covid years, even though market-dominant mail volume has decreased over the same period. As established previously, volume has in general had non-material effects on service performance in contrast to labor. In Table 7, city-carrier labor and clerk and mail handler labor still have similar effects as in Table 3.

6. Conclusion

This paper has introduced a methodology for reweighting the quarterly district-specific service performance score so as to give the underperforming district, more representativeness and, hence, more visibility in the aggregate score. The outcome of the methodology is a composite score, called the Mean-Variance score (the *MV* score). The *MV* score is the reported aggregate score discounted by a factor which is larger, the larger the coefficient of variation among the scores is. It therefore put a penalty on the spatial discrepancy between the performance scores.

The *MV* score is used next to control for the spatial integration between the performance scores in an econometric model, which seeks to explain the statistical variation in performance scores by the variations in a set of covariates including actual market dominant and competitive mail volumes, city-carrier labor and clerk and mail handler labor, and seasonal (quarter) dummies. The considered model is a linear pane data model and it is estimated as a fixed effect model with robust standard error. Measurement volumes are controlled for since they condition all performance measurements and have the desirable property of varying with both district and quarter.

The results suggest that FCSPM mail volume, while statistically significant, has almost no effect on service performance. The same can be said about competitive volumes, although the corresponding effect are greater in comparison with FCSPM volume. Labor variables have stronger effects on service performance. The *MV* score, which appears in its first 2 lagged form has significant effects, providing some empirical support for the assumptions that motivate the inclusion of these lags among the covariates.

The results may also suggest that the replacement of the old service measurement system by a new one has had a significant negative effect on the performance scores although whether these effects are causal or simple correlation is unclear. Finally, under the maintained assumptions of the model, the data provide evidence for a Covid-19 effect on 3-5-day service performance scores, while they provide no evidence for a similar effect on 2-day service performance scores. These results may be driven by the fact that the quarters falling into the Covid periods aren't that many yet. Additional insight may be gained in the future as the data accumulates.