Factoring Household Travel Surveys

HYUNGJIN KIM, JING LI, STEPHANIE ROODMAN, ASHISH SEN, SIIM SÖÖT, AND ED CHRISTOPHER

Household travel surveys have been conducted recently in a number of metropolitan areas. Metropolitan areas, travel behavior, and data processing capability have changed, requiring current data for analysis and planning; but nonresponse bias persists, often leading to a lack of adequate representation of low-income, low-education, and minority residents. A method to subdivide a study area geographically into microzones to account for the differences in return rates is suggested in conjunction with conventional factoring. In applying this basic step to data collected in suburban Chicago, the representation of families residing in selected areas increased markedly in contrast to factoring by standard techniques, yielding factored data that correspond to county-level data available from standard sources.

Currently there is a surge in collecting and processing travel data in large metropolitan areas. Household travel surveys have been conducted recently in Boston, Baltimore, Dallas, Houston, Cleveland, and Chicago. There are at least three prominent reasons for its occurrence now:

1. Cities continue to decentralize, evolving into urban structures quite unlike those for which data are available. In the past 20 years, for example, the Chicago metropolitan area has had a population increase of only about 4.1 percent, but the amount of land consumed by the urban area has increased by about half (*1*). Nevertheless, many agencies are still using data collected in 1980 or earlier.
2. The 1990 census in general and the work trip information in particular (Census Transportation Planning Package) can be used as a reference for factoring travel surveys conducted now.
3. The advances in computer hardware and software have increased the ability to process, analyze, and display transportation data.

The recent Nationwide Personal Transportation Survey further demonstrates the importance of these data collection efforts by documenting travel growth (*2*). For example, it shows that in the past 21 years (1969–1990), the population of the United States has increased by only 42.2 million inhabitants but the number of licensed drivers has increased by 60.6 million and the number of household vehicles by an astounding 92.7 million. Since much of this growth is in urban areas, the changes in travel demand that these numbers suggest should be examined with fresh data.

H. Kim, J. Li, S. Roodman, A. Sen, and S. Sööt, Urban Transportation Center, MC 357, University of Illinois, Chicago, Ill. 60680. E. Christopher, Chicago Area Transportation Study, 300 West Adams Street, Chicago, Ill. 60606.

The objective of these new travel surveys is that the data reflect the population of the entire study area and their travel patterns. Low response rates among important segments of the population, particularly in minority neighborhoods, can lead to the underrepresentation of these segments in the final data. This in turn leads to their unintended exclusion from the transportation planning process. The purpose of this paper is to suggest a method for improving the factoring process by incorporating the spatial pattern of return rates as an additional step in the factoring process. The aim is to minimize the underrepresentation of population subgroups, regardless of their characteristics.

The survey return rates for suburban Lake County, north of Chicago, used later in the paper to demonstrate the method, significantly correlated with a number of population characteristics. Return rates aggregated by square-mile microzones yield negative correlations (significant at 0.01) with population density, percentage of population with annual incomes under $8,000, percentage of population under age 6, and percentage minority population (Latino, African American, and Asian). College graduation rates were significant at 0.05. Significant positive correlations were recorded with percentage of households with incomes exceeding $50,000 and percentage of households living in single-family homes.

This implies that higher response rates are expected from affluent, low-density suburban communities than those from high-density urban areas with lower incomes. It is reasonable to assume that the lower return rates are found in areas where mobility is limited. The interest in the mobility-disadvantaged makes it necessary to ensure that they are represented adequately in the survey data. It is, however, impractical to try to factor the data for all of the variables described earlier. Therefore, the authors suggest treating the problem directly, using the return rates. Furthermore, this method does not require the respondents to identify what many individuals consider to be confidential information.

Transportation surveying efforts usually include factoring by household demographics, such as household size, number of workers, and automobile availability, which are commonly held to account for variations in travel behavior. Although these demographics are useful for providing weights, they do not always account adequately for minority populations, which may have the same household size and automobile availability but different return rates and, more important, different travel patterns. The method described here addresses this problem directly. It is illustrated in a later section by using the 1989 Lake County portion of the Chicago Area Transportation Study Household Travel Survey.

It should be mentioned that the importance of factoring stems from the use of these data for descriptive purposes.

When the data are used for model development it may not be necessary to factor, and in some circumstances it may be ill-advised (see the paper by Thakuriah et al. in this Record).

BACKGROUND: BASICS OF FACTORING

This paper focuses on mail surveys, but the problems identified and the method suggested also apply to telephone surveying and personal interviewing. Telephone surveying has recently become popular, but research continues on ways to improve mail surveys as an effective way to collect data (*3,4*). Cost-effectiveness and steady return rates are the principle advantages of mail surveys (*5*).

Work has also focused on methods of factoring survey data. FHWA has provided guidelines on ways to factor surveys (*6*), and alternative procedures have been evaluated (*7*). Applications have been many: the Metropolitan Transportation Commission study (MTC) (*8*) and the more recent study in Phoenix (*9*) both illustrate the effective use of zone-specific demographic data in factoring.

The factoring method proposed here consists of three steps. First the zonal structure is selected, then demographic data are obtained and tabulated, and finally the weights are computed. A fourth step, establishing reasonableness, is advisable but not pertinent.

Factoring Zones

The selection of zones is a classic problem. It is a trade-off between using small zones to ensure internal homogeneity and using large zones so that data handling costs are minimized. Zonal homogeneity is important and can be measured by variables such as housing costs, distribution of jobs, proximity to amenities, and municipal zoning ordinances. Ideally, transportation planning zones would follow these neighborhood differences and the variation inherent in the corresponding travel behavior.

Most agencies use zonal geography to account at least partially for these patterns. The studies of MTC (San Francisco Bay area) and Phoenix used 34 and 8 superdistricts, respectively. It is unlikely, however, that factoring-zone boundaries follow the socioeconomically defined neighborhoods. Some demographic communities are split or become minority sections within factoring zones. As a consequence, residents may be grouped with individuals unlike themselves.

The authors propose the following two-stage procedure to address this problem. First, a two-tier zonal system is identified that includes basic factoring zones or districts and a system of subzones, here called microzones. In the Lake County application, the basic factoring zones are townships in the township and range system, and the microzones are 1-mi^2 neighborhoods. Since the typical township is 6 mi^2, there are 36 microzones in a basic factoring zone (Figure 1). The microzones do not need to be this small, but they should be easily identifiable zones that are internally homogeneous. In the Chicago area the township and range geography is widely used for planning purposes because major arterials often constitute the boundaries and therefore they are easily identified. Moreover, these arterials frequently are the delimiters of socioeconomic differences between adjacent neighborhoods.

Second, the microzone return rates are computed and mapped. If the basic factoring zone has relatively high and uniform microzone return rates, it remains unaltered. For example, in the southern quarter of Lake County almost all the microzones had adequate return rates (more than 20 percent) or fewer than 10 survey instruments mailed and therefore the return rates were not considered. Near the northwestern and northeastern portions of the county are areas with low return rates. Here microzones with particularly low return rates are combined into a factoring subzone, whether they are contiguous or not. The rest of the township becomes another factoring subzone. If further discrepancies in return rates remain, additional factoring subzones may be created and the original zone may be split into any number of subzones. In this way the original number of factoring zones increases, but the increase may well be moderate.

Providing special treatment to microzone areas within the basic factoring zone with particularly low response rates creates a means of reasonable representation for groups that might otherwise be underrepresented, even with the demographic factoring common to most studies.

Advances in address matching and geographic information systems make this procedure feasible. A requisite is the ability to build different geographically defined zones (microzones and basic factoring zones). Local convention may well dictate whether microzones are square miles, census tracts, or other zones.

Present computer technology allows the use of a large number of analysis and factoring zones, but there are still several practical reasons to keep the number moderate. First, communicating information in reports would be difficult if tables had hundreds of zones. Not only is the amount of information overwhelming, but as the number of zones increases, the likelihood decreases that readers understand where these zones are located. Second, for longitudinal studies it would be useful to maintain consistency in zonal geography. Therefore, the increase in the number of zones should be controlled.

Factoring on Basis of Demographics

In the second step, the data are factored using traditional demographic characteristics that account for much of the variation in trip-making behavior. In the authors' survey of large metropolitan transportation planning agencies across the country, 11 of the 23 organizations surveyed used some method of factoring in their survey work. Of these, the authors received information on eight, and six crossed specific variables and used direct extrapolation in the matrix (two were unsure of the factoring method) using census information as their standard. The most commonly used variables—household size and number of vehicles per household—were used by half of the organizations, and they are used here. Other variables such as number of workers in the household would serve well as alternatives.

Computing Weights

The third and final step is the computation of the weights for factoring. This is discussed later.

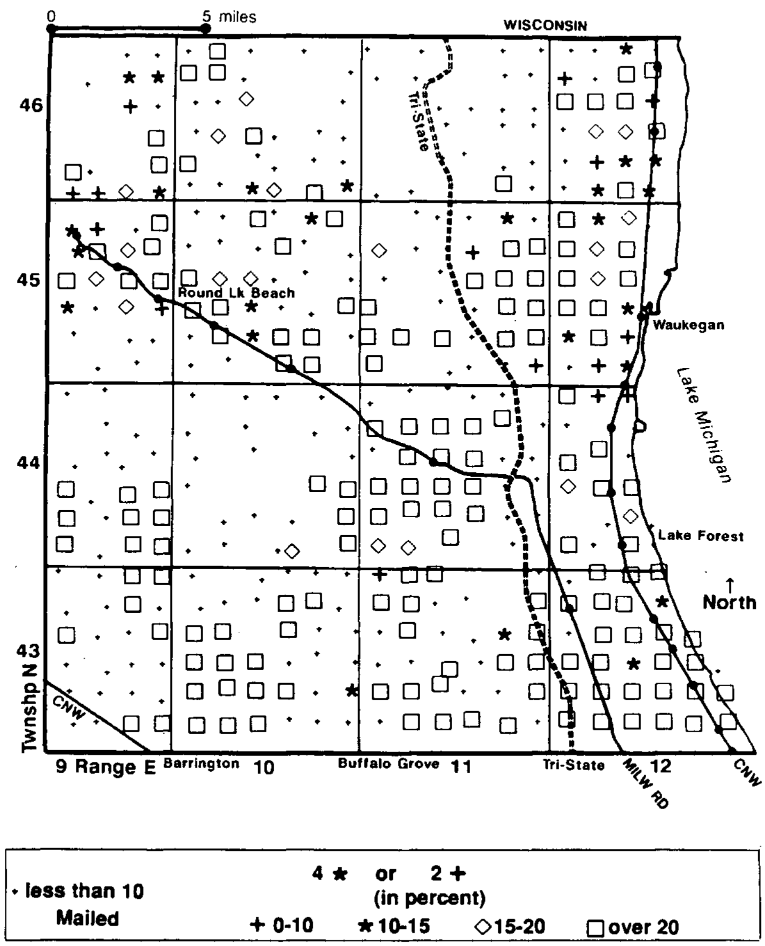

FIGURE 1 **Lake County return rate by square-mile zones.**

FACTORING: PROCEDURE

In this section we will describe the data and the computational steps to determine the factoring weights. A sample application illustrating the procedure is provided later.

Input Data

For the purpose of factoring there are four key pieces of information in a mail-out/mail-back survey:

M_c = number of questionnaires mailed to Zone c ($c \in b$, and $b \in a$),

Q_c = number of questionnaires returned from Zone c ($c \in b$, and $b \in a$),

S_{ad} = number of survey households in Zone a that belong to Demographic Category d, and

H_{ad} = number of households reported by census in Zone a that belong to Demographic Category d,

where

c = microzone (e.g., 1-mi^2 zone),
b = combination of microzones or factoring subzone, and
a = large zone or basic factoring zone (e.g., a township).

Step 1: Identifying Different Factoring Areas Using Return Rates

The intent of the first step is to identify areas with approximately the same return rates. These are either basic factoring zones a or, if they are not satisfactory, a combination of the microzones (sum of c zones).

Figure 1 illustrates a sample case in which there are initially 16 basic factoring zones (townships). The return rate, R_c, of each microzone c,

$$R_c = Q_c/M_c \tag{1}$$

is used to form subzones (b's) with approximately the same return rates. Along the southern tier the townships are found to be homogeneous and therefore remain unaltered. The two townships in the northeastern corner of Lake County along Lake Michigan (12-45 and 12-46) have many microzones with very low return rates. These microzones, regardless of contiguity, are aggregated as a separate factoring zone, yielding two factoring zones in the township. Although in this paper the aggregated area is called a subzone, it is a factoring zone just like the unaltered township.

Step 2: Preparing Demographic Data—Household Size and Vehicle Availability

After defining subzones on the basis of return rates in Step 1, the demographic data for factoring need to be prepared. As stated, household size and vehicle availability are used. Since low numbers of households in each cell of the cross tabulation in Figure 2 would result in high variances, the authors combined cells to ensure minimum sample sizes in each factoring cell. For example, many single-person households with more than one vehicle would not be expected, so these are combined.

In Lake County the six aggregates of Cells I through VI were selected as shown in Figure 2. For example, Category I includes all zero-vehicle households and those one-vehicle households with three or more members. By tallying the data from DuPage and Cook counties (the two most populated counties in metropolitan Chicago) by the six categories in each township, it is found that in order of frequency they are V, IV, VI, I, II, and III. Although on average Category III was the least populated, there was at least one township in Lake County in which it recorded the largest number of households.

Step 3: Computing Factoring Weights

For factoring zones that are not subdivided, such as those in the southern portion of Lake County (Figure 1), factoring is simple and direct. Each record is weighted by

$$W_{ad} = H_{ad}/S_{ad} \tag{2}$$

And if the cross tabulation in Figure 3 can be completed for each microzone (c's), and thereby for the aggregates of these microzones (b's) then Equation 2 may be used.

In cases in which there are several factoring subzones within a township and the cross tabulation data are not available, the calculations deriving the weights are described here.

Determine the weight, W_{bd}, such that (see Figure 3)

$$\sum_{b=1}^{n} W_{bd}S_{bd} = H_{+d}$$

and

$$\sum_{d=1}^{m} W_{bd}S_{bd} = H_{b+} \tag{3}$$

The example in Figure 1 has only two subzones in each basic factoring zone (township).

Household Size

Vehicles Available	1	2	3	4+
0	I			
1	II	III		
2		IV	V	
3+	VI			

FIGURE 2 Cross tabulation of data for each factoring area.

Household size and Vehicles available (d)

Subzone (b)*	I	II	III	IV	V	VI	H_{b+}
1							
2							
.							
n							
H_{+d}							

FIGURE 3 Demographic cross tabulation for each large zone.

Since detailed census data are not yet available, the overall weights, W_b and W_d, for each subzone and demographic category can be used in Equation 3. Although there are two subzones in each basic factoring zone in Figure 1, the notation of Equation 3 illustrates n subzones.

The weights

$$W_b = H_{b+}/S_{b+} \tag{4}$$

are applied to S_{bd} to satisfy the second part of Equation 3. However, the first part of Equation 3 is not satisfied. In other words, W_b scales up the number of households in each Zone b to match with census data, but it might not match with the census demographic categories. Therefore, adjusting weights are needed. The adjusting weight is

$$W_d = H_{+d} \Big/ \sum_{b=1}^{n} W_b S_{bd} \tag{5}$$

It is verified that $W_b W_d S_{bd}$ satisfies the first part of Equation 3 but not necessarily the second. To satisfy the second part, continue this adjusting process on each occasion satisfying one of the two parts of Equation 3. After r pairs of such steps, compute the general form of Equations 3 and 4 as

$$W_b^{r+1} = \frac{H_{b+}}{\Sigma_{d=1}^{m} \left\{ \prod_{p=1}^{r} W_b^p W_d^p \right\} S_{bd}} \tag{6}$$

and

$$W_d^{r+1} = \frac{H_{+d}}{\Sigma_{b=1}^{n} \left\{ \prod_{p=1}^{r} W_b^p W_d^p \right\} W_b^{r+1} S_{bd}} \tag{7}$$

This is the Deming-Stephan or Furness (DSF) procedure. It is easily generalized to more dimensions—for example, if there are more demographic categories or more zones.

Final Factor

The final factor, W_{bd}, can be calculated using W_b and W_d obtained from the DSF procedure (Equations 6 and 7). The final factor is

$$\left\{ \prod_{p=1}^{r} W_b^p W_d^p \right\} \rightarrow W_{bd} \tag{8}$$

In the example presented later, W_{bd} converges in five iterations.

APPLICATION

The proposed method of factoring was applied using the Chicago Area Transportation Study Household Travel Survey. The data were collected from Lake County, Illinois, a county of 173,996 households in 1990. It is a rapidly growing county in northern suburban Chicago, bordering Wisconsin, Lake Michigan, and Cook County (Chicago). Demographically, the county has a large range of characteristics, with wealthy neighborhoods encompassing large estates as well as sizable low-income communities.

The three principal minority groups—African Americans, Latinos, and Asians and Pacific Islanders—constitute about 16 percent of the population, with Latinos surpassing African Americans in the 1980s as the largest group. In sum, the county is a low-density suburban area with great contrasts.

Household Survey

Survey instruments were mailed to 9,143 households on the basis of square-mile zones of residence, or the microzones. Since there were approximately 500 such microzones in the county, and this county constituted only one of seven counties in the entire study, it was not practical to make the microzone the basic factoring zone. Instead 16 townships, 36 mi^2 each, constituted the basic factoring zones.

The average return rate in the county was 27.2 percent, but the range per microzone was more than 50 percentage points. The overall pattern by microzone was considered, and zones with fewer than 15 percent, approximately half of the county average, were flagged. If there were enough microzones in a township, they were combined into a separate factoring subzone in each township. Six townships had sufficient numbers of microzones with low return rates that they were combined into factoring subzones. For the 10 townships not subdivided, the response rate ranged from 25 to 37 percent. In sum there were 22 factoring zones.

Results of Proposed Method

The proposed method was used, and a summary of the results for the six townships that were subdivided is presented in Table 1. It indicates, for example, that Township 9-46 (in the very northwestern corner of the county) was one of six in which two subzones were defined. Considering the subzone with a low return rate, if the survey returns were factored only by the six demographic categories of Figure 2, then this portion of Township 9-46 (the subzone) would be factored to 1,285 households. Using the suggested method, this area of low return rate is factored to 2,454 households, or an increase of 1,169 households. In this instance factoring by only demographic characteristics would have vastly underrepresented the travel generated by the subzone with low return rates. In essence, what has been achieved is a geographic redistribution of survey weights within the township without changing the township total.

TABLE 1 Effect of Proposed Method

	Number of Households in Selected Low Return-Rate Subzones		
Township[a]	Proposed Method	Conventional Method	Township
9-46	2,454	1,285	4,389
12-46	4,447	2,766	12,325
9-45	2,070	860	6,709
12-44	1,533	283	7,840
11-45	1,106	253	12,440
12-45	12,018	6,552	29,097

[a]See Figure 1 for Cartesian reference system used to identify townships.

There is little doubt from Table 1 that the proposed method has an effect on the factoring results. For the conventional method it is likely that reasonableness checks would alter the results and bring them closer to the actual. Still, there is such a sizable discrepancy between the two results in Table 1 that many reasonableness checks and adjustments would probably be necessary for the conventional method. Each adjustment may also bring undesirable effects.

REASONABLENESS CHECKS

This survey has three components: the household, person, and trip files. A systematic check of the factored data would include an examination of all three files. Some individuals in a household may not have been included, and certainly not all trips were included for all respondents. Files that do not yield adequate results can be given new weights to compensate for missing data. Ideally, the weights computed for the household file are adequate for the person and trip files, and additional weights need not be computed.

The data structure of the Nationwide Personal Transportation Survey includes a half-dozen separate files, each with its own factors. Although this approach achieves the desired effects on a file-by-file basis, the data that overlap between files do not match, highlighting the apparent inconsistencies in the data. For example, the number of vehicles is reported in several files but the household and vehicles files yield different totals. This is unavoidable when each file carries its own weight.

Conversely, if only the household file is factored there is the risk that the data in the person and the trip files are not properly adjusted. It is essential that great care be taken to factor the household data properly so that the person and trip files need not be factored separately.

After the household file was factored and these factors were applied to the person and trip files, checks of reasonableness were performed on a selected variables from each of the two files. Total population was selected to check the person file, and commuter rail use and miles traveled were used for the trip file. Since the factoring procedure was based on household size and number of workers per household, it was not necessary to check the household files.

A tally of the number of persons yields a population total that is 98 percent of the 1990 census figure for Lake County, indicating that the person file is adjusted adequately. A comparison of commuter rail users, a test of the trip file, is slightly more difficult since ridership data are reported by station whereas survey data are aggregated by township. Station data within a township can be added, but their service areas typically do not follow township boundaries. Table 2 summarizes the commuter rail ridership data for selected townships, where service area definition appeared obvious.

Since the survey data overstate the number of actual commuter rail users, as one would expect from a survey that has higher returns from high-income neighborhoods, the greater emphasis on low-return-rate zones has not overadjusted for this group. In sum, the survey data match the actual ridership data well enough that recomputation of the weights does not appear to be necessary.

The survey trip file also reports trip origins and destinations, coded by ½-mi^2 zones. Summing the air-line distances provides an estimate of 9 million mi of travel for Lake County. Converting these miles to route miles of actual travel adds about 20 percent, yielding an estimate of 11 million mi.

Finally, since commercial traffic is not included in the survey, the estimate of a million such miles needs to be added, resulting in a total estimate of 12 million vehicle miles traveled. This compares very favorably with the estimate of 11.9 million mi from the State of Illinois Roadway file for Lake County (1989) and the CATS assignment model estimate of 12.1 million mi, suggesting that no further adjustment of the trip file is necessary.

If the data do not match the expectation, they need to be adjusted accordingly, using the DSF procedure described earlier. Clearly, the problems associated with readjusting the weights can best be diminished by starting with data that are well factored. Adjustments of the weights are occasionally necessary, but this process cannot create data that do not exist in the survey.

CONCLUSIONS

Considerable research has been conducted on alleviating and controlling for bias in travel surveys. Many agencies use a variation of the factoring method based on demographic characteristics. Although these methods have considerable merit, they do not necessarily address a major deficiency in many travel surveys—namely, that key population subgroups characterized by low incomes and low education achievement rates tend to remain underrepresented. When income information is not solicited it cannot be used to adjust the results, nor should it be used when the information is solicited but the responses are unreliable. It is for these cases that this method is useful.

Since it is likely that the underrepresented population is mobility-disadvantaged, it is particularly important that they be included in a transportation data base. In suburban Chicago the proposed method increased—in contrast to conventional factoring—the representation in low-return-rate areas by more than 10,000 households, yet it did not overrepresent this group.

Two advantages of the method are its simplicity and cost. It is logical and easily applied, especially given the ability to manipulate geographic information by computer.

TABLE 2 Commuter Rail Use: Survey Estimates and Metra Station Data Aggregated by Township

Township	Metra Data	CATS Survey	Difference
12-46	197	211	14
9-45	1,168	1,068	80
11-44	2,331	2,141	190
10-45	1,339	1,203[a]	136
Total	17,604	18,039[b]	435

[a]Includes adjacent townships without Metra stations.
[b]Assumes that 5,000 Lake County Metra commuter rail users crossed county line to closest station.

REFERENCES

1. Gridlock Without Growth? *Planning in Northern Illinois*, Vol. 31, No. 1, Summer 1991.
2. Hu, P., and J. Young. *Summary of Travel Trends, 1990 National Personal Transportation Survey*. FHWA, U.S. Department of Transportation, March 1992.
3. Ayidiya, S., and M. McClendon. Response Effects in Mail Surveys. *Public Opinion Quarterly*, Vol. 54, No. 2, 1990, pp. 229–247.
4. Bishop, G., H. Hippler, N. Schwartz, and F. Strack. A Comparison of Response Effects in Self-Administered Surveys. In *Telephone Survey Methodology* (R. Graves, ed.), Wiley, New York, 1988, pp. 321–340.
5. Harvey, L. Factors Affecting Response Rates of Mailed Questionnaires: A Comprehensive Review. *Journal of the Market Research Society*, Vol. 29, No. 3, 1987, pp. 341–353.
6. *Urban Origin-Destination Surveys*. FHWA, U.S. Department of Transportation, 1973, Ch. 7.
7. Ollman, L. T., S. M. Howe, K. W. Kloeber, and G. S. Cohen. Marginal Weighting of Transportation Survey Data. In *Transportation Research Record 677*, TRB, National Research Council, Washington, D.C., 1979, pp. 73–76.
8. *Sample Weighting and Trip Expansion*. Working Paper 4A, 1981 MTC Travel Survey. Metropolitan Transportation Commission, Berkeley, Calif., 1983.
9. *1988/1989 Household Travel Survey*. Behavior Research Center, Inc., Phoenix, Ariz., 1989.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.