Errors in Origin-Destination Surveys Done by Number-Plate Technique

M. M. SLAVIK

ABSTRACT

If the number-plate-matching technique is used to carry out an origin-destination survey, there is, among other dangers, the risk of finding spurious matches and thereby overestimating the amount of through traffic. A spurious match is a pair of identical entries, one recorded upstream and the other downstream, that belong to two different vehicles. Because it is difficult to calculate the number of spurious matches exactly, a sufficiently accurate approximate method has been introduced and used to assess the magnitude of the problem caused by spurious matches. When survey data are arranged and compared in small blocks, the problem is usually nonexistent or negligible. In the case of bigger blocks of data, however, the overestimation of through traffic may be surprisingly large--far too large to be ignored. The four main ways of combating spurious matches are reducing the number of entries per block; moving the upstream and downstream observation posts closer together; recording more symbols or using letters instead of digits, or both; and locating the survey on a route carrying as much through traffic as possible. Of these four, reducing the number of entries per block appears to be the most powerful strategy. If, for practical reasons, large blocks or few symbols must be used, or if the survey must be carried out over a large distance on a route with little through traffic, then the results obtained should be corrected by breaking up the matches found (M) into spurious (S) and genuine (G) ones and by using G instead of M for the evaluation of through traffic. However, when such surveys are planned, situations that call for a drastic correction of results should be identified and avoided.

When information on the movement of vehicles in a road network is needed, origin-destination traffic surveys are carried out. A popular method of surveying is based on matching the number plates of vehicles. Consider the example of a simple network shown in Figure 1, where it is desired to establish the amount of through traffic traveling from A to H. The number-plate-matching method entails the following steps:

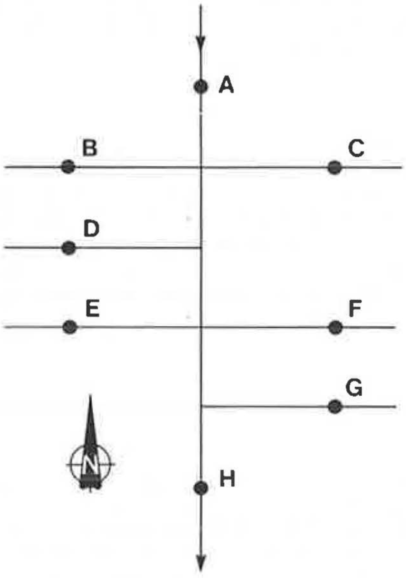

FIGURE 1 Example road network.

- During a certain period of time, say from 06:00 till 18:00, an observer at A records certain predetermined characters, usually the last three digits on the number plates of vehicles traveling south (i.e., toward H) as they go by. The observer also records time intervals by making a mark in his record, say every 15 min, so that the numbers recorded are divided into 15-min time blocks. At the same time, another observer positioned at H also records the last three digits of all southbound vehicles passing his station and marks his record in the same 15-min intervals as the upstream observer.
- At the end of the observation period, two records--the upstream from A and the downstream from H--are available. Each of these records has its entries, the three-digit numbers obtained from the number plates, divided into 15-min blocks (Figure 2).

Next, the minimum and the maximum times that a vehicle may require to travel from A to H are calculated. After this has been done, the two records are ready for matching (i.e., for finding pairs of identical entries).

The actual matching is done manually or, in the case of voluminous records, by means of a computer. The procedure is based on comparing the entries in one upstream block with the entries in a group of downstream blocks. The number of downstream blocks forming a group is determined by the minimum and maximum travel time limits calculated earlier; see the shaded blocks in Figure 2. The upstream entries are taken in sequence, one by one, and compared with

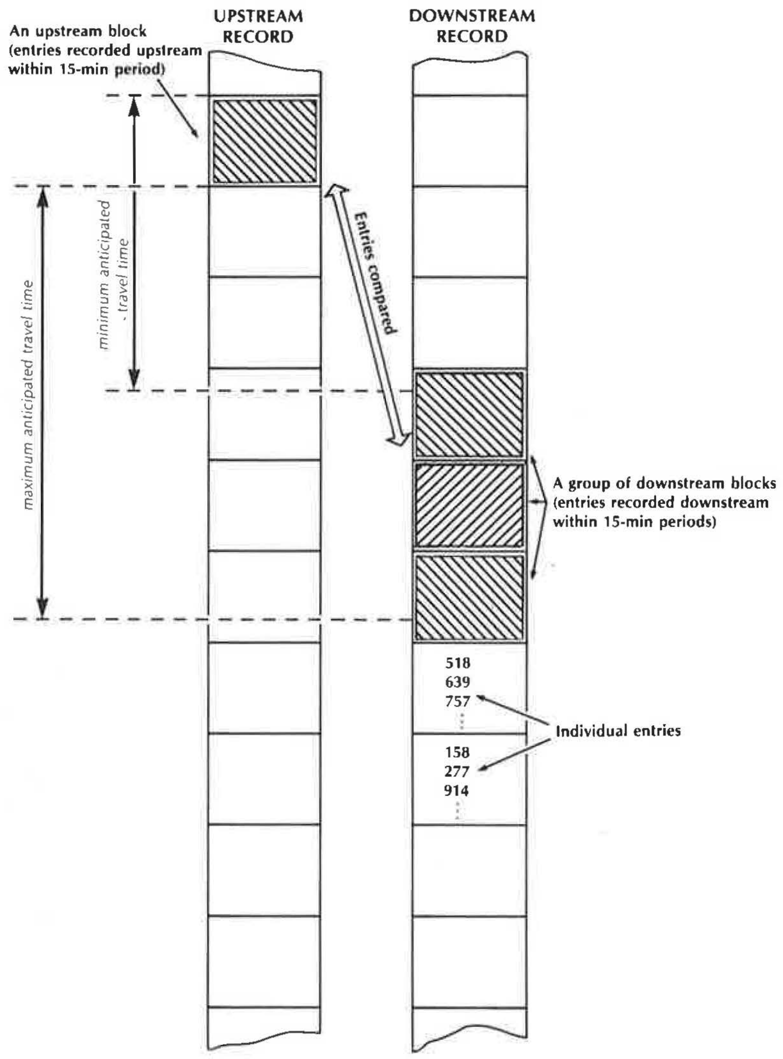
UPSTREAM RECORD
DOWNSTREAM RECORD
An upstream block (entries recorded upstream within 15-min period)
minimum anticipated travel time
maximum anticipated travel time
Entries compared
A group of downstream blocks (entries recorded downstream within 15-min periods)
518
639
757
158
277
914
Individual entries

FIGURE 2 Upstream and downstream records divided into blocks of entries.

the entries in the corresponding group of blocks in the downstream record. Every time a match is found, the respective downstream entry is deleted and no longer available for further matching.

The comparing of entries continues until all the entries in a given upstream block have been exhausted. Then the next upstream block is taken, a new group of downstream blocks is defined in accordance with the travel time limits, and the comparing of entries continues until all the upstream blocks have been exhausted.

The total number of matches (i.e., pairs) found by this procedure should thus represent the through traffic, the number of vehicles traveling from A to H during the period of observation.

This result, however, is subject to error. If, for example, one of the observers misrecords a number, then the amount of through traffic will be erroneously reduced by one vehicle. Also, if a vehicle made a lengthy stop before reaching H, then, in spite of correct recordings by both observers, the match will not be found because the upstream and downstream entries for this vehicle are too far apart. In this case, too, the amount of through traffic will erroneously be reduced by one vehicle.

By the same token, the possibility exists that the amount of through traffic will erroneously be increased as a result of spurious matches.

A spurious match is produced when two entries are matched that, in reality, belong to two different vehicles. For example, if a vehicle with registration number HMZ518T passes through A but does not exit through H and, a little later, another vehicle with registration number BYR518T passes through H without having come from A, then a match between the 518 recorded upstream and the 518 recorded downstream will be found. This match is, however, spurious because no vehicle with the 518 registration digits actually traveled from A to H. It should be noted that there are harmless spurious matches: should the same two vehicles travel from A to H, the HMZ518T passing A first but exiting from H second because of having been overtaken by the BYR518T en route, two spurious matches will be formed by the described procedure (the first between HMZ518T upstream and BYR518T downstream, and the other between BYR518T upstream and HMZ518T downstream). These spurious matches are, however, harmless because the amount of through traffic thus determined, two vehicles, is correct.

The following text concentrates only on the harmful spurious matches that falsely increase the amount of through traffic recorded. In the interest of simplicity, the word harmful is omitted and the term "spurious match" is used.

OBJECTIVES

It is assumed that appropriate steps (such as careful planning of the survey, use of experienced staff, and extensive supervision) have been taken to minimize errors due to misrecording. The emphasis in this paper is on those errors in estimation of through traffic that are caused by spurious matches rather than on those caused by human factors. The main objectives are to establish the magnitude of the problem, to introduce a corrective procedure, and to provide some guidance to practitioners.

RATIONALE

First, the following question needs to be answered: If X entries in one upstream block are compared with Y entries in a corresponding group of downstream blocks and M matches are found, how many of these M pairs are likely to be spurious matches (S)?

Then, with M and S known, an estimate of the number of genuine matches (G representing the through traffic) can be obtained by subtracting S from M, G = M - S.

Consider a simple case in which 10 entries (recorded in a given upstream block) are compared with 15 entries (recorded in, say, three consecutive downstream blocks) (Figure 3). Let it be assumed for now that there are no misrecorded entries and that no vehicles traveled from A to H. If there are any matches between the 10 upstream and the 15 downstream entries, they are certainly spurious because there was no through traffic.

Let it be further assumed that the last three digits were recorded from all number plates and that these three-digit numbers follow a uniform random distribution (i.e., any three-digit number is equally likely to occur as any other, upstream or downstream). The task now is to calculate the average number of spurious matches that is likely to occur when the matching procedure described earlier is applied.

To calculate the average number of spurious matches exactly would mean evaluating the probabilities of exactly none, one, two, three, four, five, six, seven, eight, nine, and ten matches, respectively, occurring and using these respective probabilities as weights in the calculation of the weighted average of numbers zero to ten.

Because any of the shaded boxes shown in Figure 3 can contain any of the 1,000 three-digit numbers (from 000 to 999), the number of ways in which these 25 boxes can be filled with numbers is $1{,}000^{25} = 10^{75}$--an incomprehensibly huge number. It is evident from the magnitude of this number that any evaluation based on enumeration of the probabilities that are required as weights is an impossible task.

For this reason, an approximate but simple method for the estimation of the number of spurious matches has been developed.

APPROXIMATE METHOD

The reasoning behind the approximate method can be demonstrated by means of the previous example. In this example, X = 10 upstream entries are compared with Y = 15 downstream entries, and the total number of matches (M) is evaluated.

Imagine that the first upstream entry is being compared with the first downstream entry. The probability of this comparison yielding no match is 1-1/1,000 = 0.999. The probability that the first upstream entry will not match any of the 15 downstream entries is

$$q_1 = 0.999^{15} = 0.985105$$

Consequently, the probability of the first upstream entry being matched with (at least) one of the 15 downstream entries is

$$P_1 = 1 - q_1 = 1 - 0.985105 = 0.014895$$

If the first upstream entry is matched, the matching downstream entry is labeled and the second upstream entry is then compared with the 14 remaining downstream entries. If, however, the first upstream entry is not matched, the second upstream entry is compared with all 15 downstream entries. On average, this number will be between 14 and 15, but closer to 15 because it is more probable that the

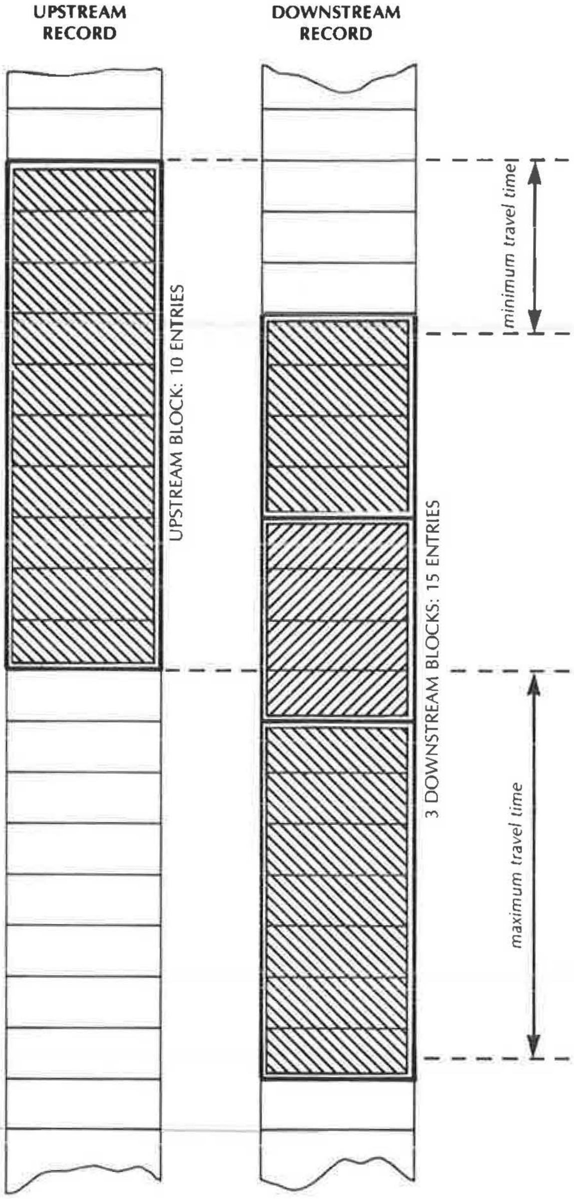

FIGURE 3 Example of matching procedure.

first upstream entry will not be matched, and the second one will thus be compared with 15 downstream entries. It could be said that the frequency with which the first upstream entry is matched is 0.014895 and that the average number of remaining downstream entries with which the second upstream entry will be compared is

$$15 - 0.014895 = 14.985105$$

The probability of the second upstream entry not being matched with any of the 14.985104 remaining downstream entries is

$$q_2 = 0.999^{14.985105} = 0.985119$$

whereas the probability of its being matched is thus

$$p_2 = 1 - q_2 = 0.014881$$

Similarly, the average number of downstream entries with which the third upstream entry will be compared is

$$15 - 0.014895 - 0.014881 = 14.970224$$

which means that

$$q_3 = 0.999^{14.970224} = 0.985134$$

and

$$p_3 = 1 - q_3 = 1 - 0.985139 = 0.014866$$

This procedure is continued until the last probability (p_{10} in the example) has been calculated. The approximate average number of spurious matches (S) is then the sum of all the ps:

$$\begin{aligned} S &= p_1 + p_2 + p_3 + \ldots + p_{10} \\ &= 0.014895 + 0.014881 + 0.014866 + \ldots \\ &= 0.148 \end{aligned}$$

This means that when 10 upstream entries (three-digit numbers) are compared with 15 downstream entries, all the entries being random numbers, then about 0.15 spurious matches are likely to be found, on average. In general,

$$S = \sum_{i=1}^{X} p_i \qquad (1)$$

where

S = average number of spurious matches;
X = number of entries in a given upstream block;
$p_i = 1 - q_i$;
$q_i = [(N - 1)/N]^w$;
N = number of permutations obtainable from the symbols recorded (e.g., N = 1,000 if three digits are recorded);
$w = \sum_{i=1}^{X-1} p_{i-1}$; and
Y = number of entries in the corresponding downstream blocks and $p_o = 0$.

For those who prefer computer language, these formulas can be expressed as follows:

```
100  S = 0;
110  for J = 1 to X;
120  Q = [(N - 1)/N] ↑ (Y - S);
130  P = 1 - Q;
140  S = S + P;
150  next J; and
160  print "SPURIOUS MATCHES S ="; S.
```

Note that X, Y, and N are given.

ACCURACY OF THE APPROXIMATION

To see how well the correct results can be approximated when Equation 1 is used, Figure 4 was prepared.

Three block sizes were selected: small blocks with 10 entries upstream facing 20 entries in the downstream blocks, medium blocks with 50 upstream entries facing 80 downstream ones, and large blocks with 70 upstream entries facing as many as 300 entries downstream.

The exact number of spurious matches was calculated with the aid of simulation. In each case the random number generator produced X uniformly distributed random numbers, the upstream entries, and further Y numbers, the downstream entries. Then the number of pairs was established and recorded. This procedure was repeated 10,000 times in each case. Each simulated number of spurious matches (S) is thus the average of 10,000 individual values of S obtained by means of the simulations. The errors shown in Figure 4 were calculated using the simulated results as a reference.

Figure 4 reveals three interesting facts:

1. The error tends to diminish with the increased number of permutations obtainable from the symbols recorded (N).
2. To a lesser degree, the error also tends to diminish with increased block size.
3. For a number of permutations (N) as small as 63, the error is still less than 5 percent irrespective of the block size.

The typical combinations of symbols used in practice and the number of permutations (N) are given in Table 1. Here D denotes digits from 0 to 9 and L represents a letter of the English alphabet. (It is assumed that only 22 of the 26 letters are eligible to be used on vehicle registration plates.)

In all of these cases the number of permutations (N) is greater than 63 and the error caused by approximation will thus be less than 5 percent, which is negligible. This means that the approximate method is a sufficiently accurate instrument.

SPURIOUS AND GENUINE MATCHES

In practice the upstream and downstream entries are not all random numbers: some vehicles do, indeed, travel from A to H, and their upstream and downstream entries thus form genuine matches. Only the remaining entries can then produce spurious matches (S).

If the number of remaining upstream and downstream entries is X and Y, respectively, and the number of genuine matches is G, then

$$X = U - G \qquad (2)$$

and

$$Y = D - G \qquad (3)$$

where

U = total number of all entries in one given upstream block and
D = total number of all entries in the corresponding group of downstream blocks.

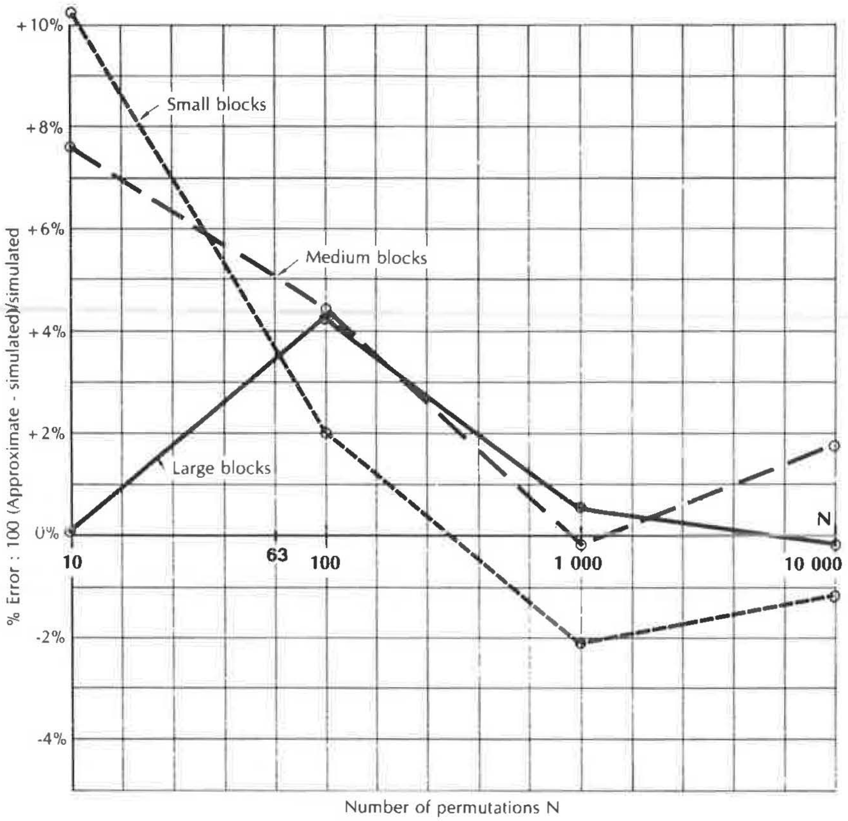

FIGURE 4 Errors of approximate method.

TABLE 1 Typical Combinations of Symbols Recorded

Combination	Example	Permutations (N)
DD	48	100
LD	G4	220
LL	MG	484
DDD	775	1,000
LDD	H58	2,200
LLD	FM7	4,840
LLL	HMZ	10,648
DDDD	3229	10,000
LDDD	K322	22,000
LLDD	BF45	48,400
LLLD	BYR8	106,480

The number of all matches found by the matching procedure when comparing U upstream with D downstream entries is M. This number comprises both the genuine and the spurious matches:

$$M = G + S$$

and

$$S = M - G. \quad (4)$$

Because the number of genuine matches (G) is not yet known, it is not possible to calculate the number of spurious matches (S) directly from Equation 4. However, it may be calculated as follows by means of an iterative procedure based on known U, D, and M:

1. First assume that S = 0; consequently, G = M.
2. Use G to calculate X and Y from Equations 2 and 3, respectively.
3. Use X and Y, as calculated, to compute a new estimate of S using Equation 1. Round this S off to the nearest integer and record it.
4. Check whether the latest S still differs from that calculated in the previous step. If so, calculate G = M - S and repeat Steps 2 and 3. When S does not differ any more terminate the iteration.

The last value of S is the estimate of the number of spurious matches. The estimated number of genuine matches (G) is G = M - S.

5. Perform the first four steps for each upstream block of entries and record G in each case.
6. The total number of genuine matches (representing the total through traffic) is the total of the individual genuine matches (G) found in Step 5.

Table 2 gives the iterative procedure with an example taken from an origin-destination survey carried out on a section of the road between Johannesburg and Durban, South Africa, in October 1983. In this survey the last three digits on the number plates were recorded, thus N = 1,000. In a particular case an upstream block had U = 129 entries. These, compared with D = 278 downstream entries, yielded 53 matches. The breaking up of the total of 53 matches into 20 spurious and 33 genuine ones was achieved in four steps (Table 2).

TABLE 2 Example of Iterative Procedure

Step	X	Y	Spurious (S)	Genuine (G)
(0)	(129)	(278)	0	53
1	76 (= 129 - 53)	225 (= 278 - 53)	15[a]	38 (= 53 - 15)
2	91 (= 129 - 38)	240 (= 278 - 38)	19[a]	34 (= 53 - 19)
3	95	244	20	33
4	96	245	20	33

[a] Calculated using Equation 1.

In this example, 53 of the 129 upstream entries appeared to be through traffic, some 41 percent. However, this result is incorrect because of the 20 spurious matches that are masquerading as through traffic; the genuine through traffic is likely to be only about 33/129 = 26 percent.

MAGNITUDE OF THE PROBLEM

The preceding example suggests that the number of spurious matches may be considerable and that the problem should not be ignored.

To demonstrate the degree to which spurious matches may jeopardize the correct evaluation of through traffic, Figure 5 was prepared.

Again, the small, medium, and large blocks of entries were considered, together with two values of N; that is, N = 1,000 and N = 10,000. The meaning of the six curves (three of which are identical, see the top of the graph) can best be explained with an example.

Suppose that a survey based on recording three digits (N = 1,000) was carried out using medium-sized blocks of entries. Suppose, further, that according to matches found, the apparent through traffic was 20 percent. Then, as shown in Figure 5, the genuine through traffic is only 0.7 of the apparent traffic (i.e., 14 instead of 20 percent of the upstream traffic).

The figures on the vertical axis can be viewed as correction factors: in this example a factor of 0.7 should be applied to the apparent through traffic in order to obtain the real one.

Several things are immediately evident from Figure 5:

1. If the survey is done using small blocks of entries, no correction of the apparent through traffic is necessary, provided that N is at least 1,000.

2. If N is as large as 10,000, then, again, no correction is necessary except where large blocks were used and the apparent through traffic is less than 80 percent of upstream traffic.

3. In all other cases a correction of the apparent through traffic is necessary, and when the percentage of through traffic is small, such correction is essential.

4. In some cases the correction required may be so drastic that the situation should be avoided. For instance, when a survey based on large blocks and N = 1,000 indicates an apparent through traffic of less than 25 percent of the upstream traffic, the correction factor to be applied is zero: all the matches found are likely to be spurious and no meaningful statement can be made about the real through traffic.

Note that the curves in Figure 5 are not quite smooth. This is because the numbers of matches were rounded off to integers before division.

DISCUSSION

Several courses of action are available to combat spurious matches.

One option is to arrange for a small block size by marking the records, both upstream and downstream, in short time intervals, say, every 5 min instead of every 15 min, or by moving the observation stations as close to each other as possible, which will shorten the eligible travel time and thus reduce the number of downstream blocks to be compared with a given upstream block. When small-sized blocks (of about 10 entries each) have been achieved, the danger of spurious matches can be neglected.

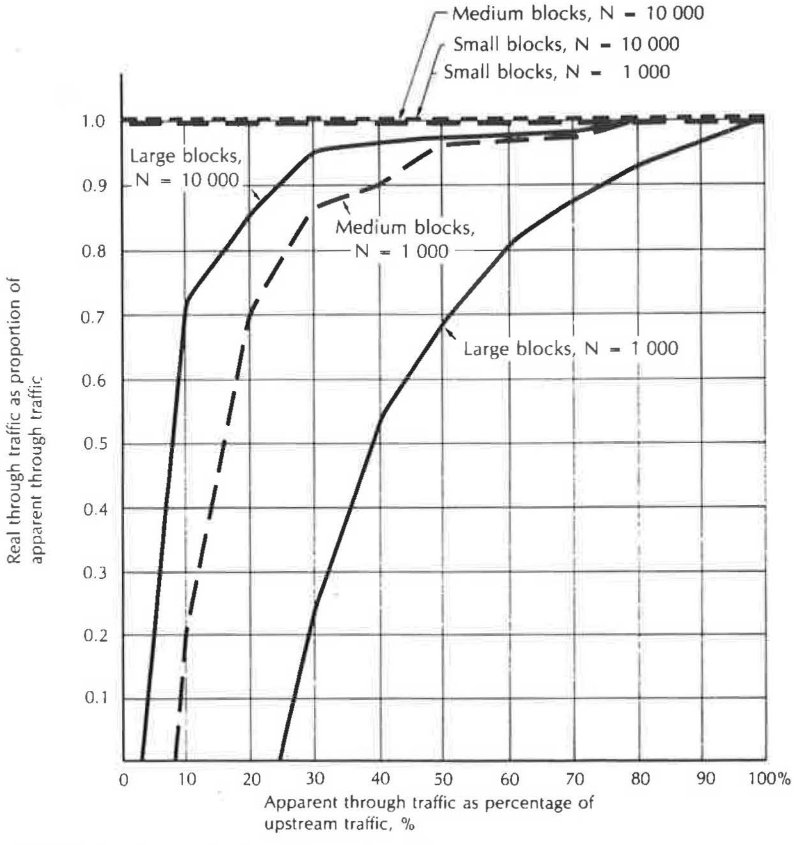

FIGURE 5 Apparent and real through traffic.

Because relatively few spurious matches occur when many matches are found, an additional improvement may be obtained by situating the upstream and the downstream observation stations on a route carrying as much through traffic as possible.

Another option is to increase N, the number of permutations that can be obtained, by varying the symbols recorded. Although recording three digits (N = 1,000) appears to be a popular practice, four-digit or three-letter entries (with N = 10,000) would greatly alleviate the problem, as has been seen in Figure 5. Recording four-symbol combinations of digits and letters would practically obviate the problem even in the case of large blocks of entries. Unfortunately, recording a larger number of symbols means a greater risk of human error resulting in misrecording and in the consequent loss of genuine matches. It would appear that, in recording three digits, a pragmatic balance has been struck between the danger of finding many spurious matches and the risk of misrecording.

The last option is to correct for spurious matches. This is done by resolving the matches found into genuine and spurious ones, using Equations 1-4 as described earlier. It must, however, be kept in mind that this method, in addition to being approximate, is stochastic--it estimates the average number of spurious matches (S) that is likely to occur in the long term rather than the exact value of S in each individual case.

It is interesting to see the relative influence of the options mentioned. This is given in Table 3 with an example based on a large block (U = 70, D = 300), three-digit recording (N = 1,000), and 43 percent apparent through traffic (M = 30 pairs). According to the correction procedure described earlier, the expected number of spurious matches (S) is 13 (see Line 1 in the table).

Now the variables U, D, N, and M are adjusted, one by one, by 20 percent, and S is recalculated in each case (see Lines 2, 3, 4, and 5). Line 6 reflects the case in which all four variables are simultaneously adjusted by 20 percent.

A comparison of the results reveals that reducing the size of the blocks, particularly of the upstream ones, appears to be the most powerful single strategy for combating spurious matches--a 20 percent reduction in upstream block size yields a 36 percent reduction in spurious matches.

TABLE 3 Impact of Change in the Individual Variables U, D, N, and M on the Number of Spurious Matches (S)

Line No.	U	D	N	M	S	Change in S (%)
1	70	300	1,000	30	13	
2	56	300	1,000	30	8	-38
3	70	240	1,000	30	9	-31
4	70	300	1,200	30	10	-23
5	70	300	1,000	36	10	-23
6	56	240	1,200	36	4	-69

CONCLUSIONS

When an origin-destination survey is performed by means of the number-plate-matching technique, there is the risk of overestimating through traffic as a result of finding spurious matches. The number of spurious matches that is likely to be found depends on four main factors:

1. The number of entries per block,
2. The travel time limits that are determined by the reasonably shortest and longest time required to travel from the upstream observation station to the downstream one,
3. The number of permutations (N) that can be obtained by varying the symbols recorded, and
4. The total number of matches found.

Virtually no spurious matches will be encountered in the case of small blocks (about 10 upstream entries per block facing about 20 entries in the downstream blocks) provided that N is at least 1,000. There is no danger of spurious matches occurring, even in the case of medium-sized blocks (about 50 upstream entries per block facing some 80 entries downstream) provided that N is as large as 10,000.

In all other cases the danger of finding spurious matches and consequently overestimating the amount of through traffic is too great to be ignored. The options to combat spurious matches are reducing the block size, shortening the distance between the upstream and the downstream observation stations, increasing the number of permutations (N) by recording more symbols or using letters instead of digits, and locating the survey on a route that has a large proportion of through traffic.

When these options are impractical or impossible, the amount of through traffic established by the matching procedure should be corrected in accordance with the procedure explained. In this case the origin-destination surveys should be carefully planned to obviate the need for drastic corrections by factors close to zero.