shows the frequency distribution of contact numbers for the contiguous U.S. states and the counties of England.

It is evident that very regular patterns, like honeycombs, will have frequency distributions with a pronounced peak at a single value, while more complex patterns will show spreads around a central modal value. The independent random process introduced in Chapter 3 can be used to generate polygonal areas and in the long run produces the expected distribution given in the last column of the table. The *modal*, or most frequently occurring, value is for areas with six neighbors. It is apparent that these administrative areas have lower contact numbers than expected, implying a patterning more regular than random. Note, however, that random expected values cannot be compared directly with the observed case for two reasons. First, the method of defining the random process areas ensures that the minimum contact number must be three, and second, the procedure does not have edge constraints, whereas both the United States and England have edges. Furthermore, as with point pattern analysis, the usefulness of this finding is open to debate, since we know to begin with that the null hypothesis of randomness is unlikely to hold.

Perhaps more useful are measures of fragmentation, or the extent to which the spatial pattern of a set of areas is broken up. Fragmentation indices are used widely in ecology (see, e.g., Turner et al., 2001). This can be particularly relevant when roads are cut through forest or other wilderness areas, changing both the *shape* and the *fragmentation* of habitats considerably, even where total habitat *area* is not much affected. GIS tools for applying many ecological measures of landscape pattern are available in a package called FRAGSTATS (Berry et al., 1998).

7.4. SPATIAL AUTOCORRELATION: INTRODUCING THE JOINS COUNT APPROACH

In this section we develop the idea of spatial autocorrelation, introduced in the context of problems with spatial data in Chapter 2. You will recall that spatial autocorrelation is a technical term for the fact that spatial data from near locations are more likely to be similar than data from distant locations. More correctly, any set of spatial data is likely to have characteristic distances or lengths at which it is correlated with itself, a property known as self-correlation or *autocorrelation*. Furthermore, according to Tobler's (1970) "first law of geography" that "Everything is related to everything else, but near things are more related than distant things", autocorrelation is likely to be most pronounced at short distances. If the world were not spatially autocorrelated in this way, geography would have little point, so autocorrelation is extremely important to the discipline and to GIS analysis. The ubiquity of spatial autocorrelation is the reason why spatial data are special. As a result of autocorrelation, samples from spatial data are not truly random, with consequences for statistics that are a major theme of this book.

As geographers or GI analysts, we are predisposed to seeing spatial patterns in data, and because of autocorrelation, patterns very often appear to be there. One reason for developing analytical approaches to spatial autocorrelation is to provide a more objective basis for deciding whether or not there really is a pattern, and if so, how unusual that pattern is. The by-now-familiar problem is to decide whether or not any observed spatial autocorrelation is significantly different from random. Could the apparent pattern have occurred by chance? Arguably, one of the tests for autocorrelation discussed in the remainder of this chapter should *always* be carried out before we start developing elaborate theories to explain the patterns we think we see in a map.

The autocorrelation concept is applicable to all the types of spatial object we have recognized (point, line, area, and field), but for pedagogic convenience and with one eye on tradition, we introduce the idea in the context of patterns in the attributes of area objects. Our strategy is to develop it using a simple one-dimensional analogy where the attribute is binary, with only two possible outcomes. We then extend this approach to the examination of two-dimensional variation, again using attributes that can take on only one of two values. Finally, we describe measures that allow interval- or ratio-scaled data to be used.

Runs in Serial Data or One-Dimensional Autocorrelation

Imagine taking half a deck of cards, say the red diamonds (♦) and the black clubs (♣). If we shuffle the pack and then draw all 26 cards, noting which suit comes up each time, in order, we might get

♣♦♣♣♦♣♦♣♦♦♦♣♦♣♣♣♦♣♦♦♦♦♣♦♣♣ (7.4)

Is this sequence random? That is, could it have been generated by a random process? If the sequence was

♣♣♣♣♣♣♣♣♣♣♣♣♣♦♦♦♦♦♦♦♦♦♦♦♦♦ (7.5)

or

♦♦♦♦♦♦♦♦♦♦♦♦♦♣♣♣♣♣♣♣♣♣♣♣♣♣ (7.6)

would we consider these likely outcomes of a random process? These both look highly unlikely, and in much the same way. Again, does the sequence

$$♣♦♣♦♣♦♣♦♣♦♣♦♣♦♣♦♣♦♣♦♣♦♣♦♣♦ \tag{7.7}$$

seem a likely outcome of a random process? This also looks unlikely, but for different reasons. Note that all four sequences contain 13 of each suit, so it is impossible to tell them apart using simple distributional statistics. What we are focusing on instead is how each card drawn relates to the one before and the one after—to its *neighbors* in the sequence, in effect.

One way to quantify how unlikely these sequences are is by counting *runs* of either outcome (i.e., unbroken sequences of only clubs or only diamonds). For the examples above we have

$$♣|♦|♣♣|♦|♣|♦|♣|♦♦♦|♣|♦|♣♣♣|♦|♣|♦♦♦♦|♣|♦|♣♣ \tag{7.8}$$

$$♣♣♣♣♣♣♣♣♣♣♣♣♣|♦♦♦♦♦♦♦♦♦♦♦♦♦ \tag{7.9}$$

$$♦♦♦♦♦♦♦♦♦♦♦♦♦|♣♣♣♣♣♣♣♣♣♣♣♣♣ \tag{7.10}$$

$$♣|♦|♣|♦|♣|♦|♣|♦|♣|♦|♣|♦|♣|♦|♣|♦|♣|♦|♣|♦|♣|♦|♣|♦|♣|♦ \tag{7.11}$$

making 17, 2, 2, and 26 runs, respectively. Very low or very high numbers of runs suggest that the sequence is not random. Thus we have the makings of a statistic that summarizes the pattern of similarity between consecutive outcomes in a sequence of card draws.

Get Out the Cards!

It isn't essential, but why not try this experiment yourself? Take the 26 clubs and diamonds from a pack of cards, shuffle them well, and then create a sequence similar to those above. Count the number of runs.

If you wish, repeat the same operation several times (or compare results with your classmates) and develop a histogram of these results. Note that every time you perform the experiment you are creating a *realization* of a random process (see Chapter 3) and your histogram says something about what can be expected from the process.

By determining the sampling distribution of this *runs count statistic* for a random process using probability theory, we can assess the likelihood of each sequence. If there are n_1 ♣'s and n_2 ◆'s in the sequence, the expected number of runs is given by

$$E(\text{no. runs}) = \frac{2n_1n_2}{n_1+n_2} + 1 \tag{7.12}$$

which is binomially distributed (see Appendix A) with an expected standard deviation:

$$E(s_{\text{no. runs}}) = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}} \tag{7.13}$$

In the cases with 13 ♣'s ($n_1 = 13$) and 13 ◆'s ($n_2 = 13$), we have

$$E(\text{no. runs}) = \frac{2 \times 13 \times 13}{13+13} + 1 = 14 \tag{7.14}$$

$$\begin{aligned} E(s_{\text{no. runs}}) &= \sqrt{\frac{2 \times 13 \times 13 \times (2 \times 13 \times 13 - 13 - 13)}{(13+13)^2(13+13-1)}} \\ &= \sqrt{\frac{105{,}456}{16{,}900}} \\ &= \sqrt{6.24} \\ &= 2.4980 \end{aligned} \tag{7.15}$$

For a reasonably large number of trials, we can use these results to calculate z-scores and approximate probabilities, by using the normal distribution as an approximation to the binomial. For the first example above, with 17 runs we have

$$z = \frac{17-14}{2.4980} = 1.202 \tag{7.16}$$

which is a little more than we expect, but not particularly out of the ordinary, and therefore not at all unlikely to have occurred by chance ($p = 0.203$ in a two-tailed test). For the sequence of all ◆'s followed by all ♣'s (or vice versa), we have two runs and a z-score given by

$$\begin{aligned} z &= \frac{2-14}{2.4980} \\ &= -4.8038 \end{aligned} \tag{7.17}$$

This is extremely unlikely to have occurred by chance ($p = 0.0000016$ in a two-tailed test). Similarly, for the alternating pattern with 26 runs we have

$$\begin{aligned} z &= \frac{26 - 14}{2.4980} \\ &= 4.8083 \end{aligned} \tag{7.18}$$

which is equally unlikely.

Some Arithmetic and Some Statistics Revision

If you did the previous experiment, use the number of runs you obtained to find the z-score and the chance that this is a realization from a random process. If you accumulated a histogram of several realizations in the previous experiment, how do the mean and standard deviation for your experimental data compare with the values that theory gives? Again, this is a useful classroom exercise.

The runs statistic and the test against random chance we discuss here is dependent on the number of each outcome in our sequence. This is because drawing cards from a pack is a *nonfree sampling* process. Because there are only 13 of each suit in the pack, as we near the end of the process, the number of possible draws becomes restricted, and eventually for the last card, we know exactly what to expect (this is why some card games of chance actually require skill, the skill of counting cards). Nonfree sampling is analogous to the situation applicable in many geographical contexts, where we know the overall counts of spatial elements in various states and we want to know how likely or unlikely their arrangement is, given the overall counts.

The expressions given above are true even when n_1 or n_2 is very small. For example, suppose that the experiment involved flipping a coin 26 times, and we got the realization

$$\text{HHHHHHHHHHHHHHHHHHHHHHHHHH} \tag{7.19}$$

This is pretty unlikely. But *given that we threw 26 heads*, the number of runs (i.e., 1) is not at all unexpected. We can calculate it from the equations as before. Since $n_1 = 26$ and $n_2 = 0$, we have

$$E(\text{no. runs}) = \frac{2 \times 26 \times 0}{26 + 0} + 1 \\ = 1 \tag{7.20}$$

which is exactly what we got. In fact, this is the only possible number of runs we could get given the number of heads and tails. This means that because only a very limited number of outcomes for the runs count are possible, the nonfree sampling form of the runs statistic is not very useful when one outcome is dominant in the sequence. As we will see in the next section, this observation also applies when we carry the concepts over to two dimensions.

It is also possible to calculate the probability of the sequence of 26 heads above occurring given an a priori probability of throwing a head with each flip of the coin. The expected number of runs is

$$E(\text{no. runs}) = R + 1 \tag{7.21}$$

where R is a binomial distribution. This is based on the logic that the number of runs is at least one plus however many run-ending outcomes occur. A run-ending outcome is an H after a T, or vice versa, and occurs with probability $p = 0.5$ each time a coin is flipped, because whatever the previous outcome, the chance of the next one being different is one in two. This is therefore the probability that any particular flip of the coin will terminate a run, so starting a new one, and increasing the runs count by 1. Thus R is distributed binomially with the number of trials being equal to the number of gaps (or *joins*) between coin flips, 25 in this case, or, more generally, $n - 1$.

Putting all this together, from the binomial distribution the expected value of R will be

$$2(n-1)p(1-p) \tag{7.22}$$

so the expected number of runs is this quantity plus 1. For $n = 26$ and $p = 0.5$, this gives us $E(\text{no. runs}) = 13.5$. The standard deviation for this binomial distribution is

$$\sqrt{(n-1)p(1-p)} \tag{7.23}$$

Again, putting in the values, we get $s = \sqrt{25 \times 0.5 \times 0.5} = 2.5$, from which we can assign a z-score to only a single run, as above, of $(1 - 13.5)/2.5 = -5$. As we might expect, a sequence of 26 heads is highly unlikely given the prior probability of getting heads when we flip a coin.

There are thus slight differences in the runs test statistic depending on the exact question you wish to ask. In geography, we are usually interested in a case analogous to drawing from a deck of cards. We are more interested in asking how likely is the *arrangement*—the *spatial configuration*—of the known outcomes rather than how likely the arrangement is given all possible outcomes. The two perspectives result in different mathematical formulas, but except when the proportions of each outcome are very different, the practical impact on the probabilities calculated is relatively small. Typically, the mathematics for the free-sampling case is simpler, so we focus on it below, even though the nonfree case is often more correct.

Extending Runs Counting to Two Dimensions: The Joins Count

We hope that at this point you can just about see the connection of the runs count test to spatial autocorrelation. A directly analogous procedure in two dimensions can be used to test for spatial autocorrelation. This test is called the *joins count test for spatial autocorrelation*, developed by Cliff and Ord (1973) in their classic book *Spatial Autocorrelation*.

The joins count statistic is applied to a map of areal units where each unit is classified as either black (B) or white (W). These labels are used in developing the statistic, but in an application they could equally be based on a classification of more complex, possibly ratio-scaled data. For example, we might have county unemployment rates, and classify them as above or below the average or the median.

To develop the statistic, we use the simple examples shown in Figure 7.5. As with point patterns, the standard or null hypothesis against which we assess the observed pattern of area attributes is an independent random process that assigns a B or W classification randomly to each spatial object. We could do this by flipping a coin. Even without performing the experiment, you can immediately see that the realizations in Figure 7.5*a* and *c* are unlikely. We would like to see this reflected in our statistic.

The *joins count* is determined by counting the number of occurrences in the map of each of the possible joins between neighboring areal units. The possible joins are BB, WW, and BW/WB and the numbers of each of these types of join, J_{BB}, J_{WW}, and J_{BW} are written alongside each grid. The numbers in the first column consider only the four north–south–east–west neighbors of each grid square, by analogy with the moves that a rook (or castle) can make in chess, the *rook's case*. Counts in the second column include diagonal neighbors that touch at a corner, the *queen's case*.

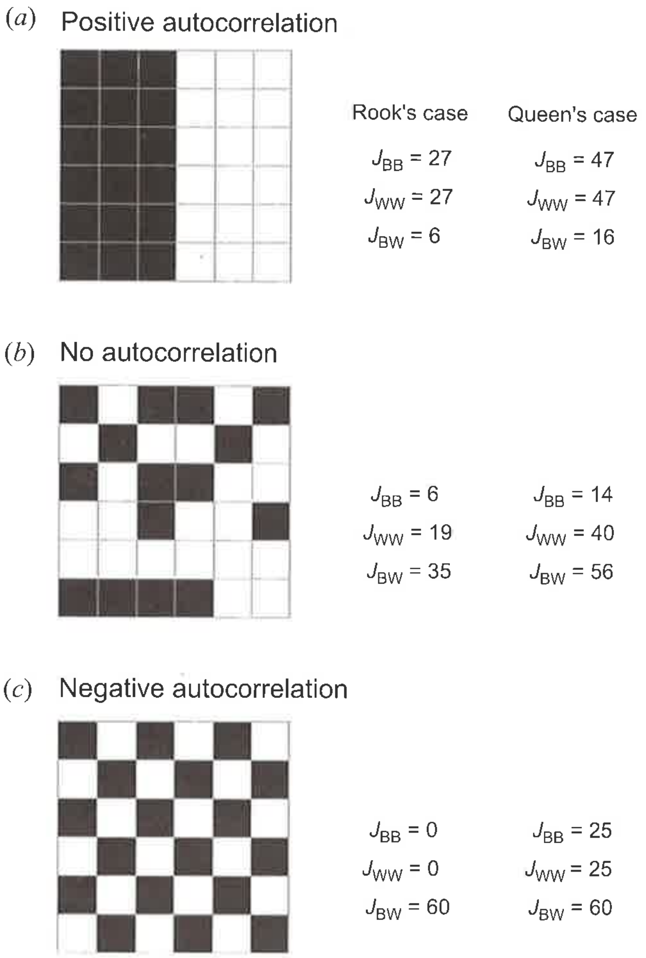

Figure 7.5 Three simple grids used in discussing the joins count statistic.

Your Own Realization

Draw out a 6 × 6 grid, thus creating 36 individual cells. Using a coin toss for each cell, color it black (heads) or white (tails). Your answers will most probably look something like Figure 7.5*b*. Now count the number of joins of each type to get your own values for J_{WW}, J_{BB}, and J_{BW} using the rook's case.

(continues)

(box continued)

When counting, you will find it simpler to work systematically counting all joins for every grid cell in turn. This procedure *double-counts* joins, so you must halve the resulting counts to obtain correct values for J_{WW}, J_{BB}, and J_{BW}. This is usually how computer programs that perform this analysis work.

Having obtained the counts, can you suggest how to check them? What should be the sum of J_{WW}, J_{BB}, and J_{BW}?

Considering only the rook's case counts in the first column of Figure 7.5, we can see that BB and WW joins are the most common in case (a), whereas BW joins are most common in case (c). This reflects the spatial structure of these two examples. Example (a), where cells are usually found next to similarly shaded cells, is referred to as *positive spatial autocorrelation*. When elements next to one another are usually different, as in case (c), *negative spatial autocorrelation* is present. Obviously, the raw joins counts are rough indicators of the overall autocorrelation structure.

We can develop these counts into a statistical test using theoretical results for the expected outcome of the joins count statistics for an independent random process. The expected values of each count are given by

$$\begin{aligned} E(J_{\mathrm{BB}}) &= kp_B^2 \\ E(J_{WW}) &= kp_W^2 \\ E(J_{\mathrm{BW}}) &= 2kp_Bp_W \end{aligned} \tag{7.24}$$

where k is the total number of joins on the map, p_{B} is the probability of an area being coded B, and p_{W} is the probability of an area being coded W. The expected standard deviations are also known. These are given by the more complicated formulas

$$\begin{aligned} E(s_{\mathrm{BB}}) &= \sqrt{kp_{\mathrm{B}}^2 + 2mp_{\mathrm{B}}^3 - (k+2m)p_{\mathrm{B}}^4} \\ E(s_{\mathrm{WW}}) &= \sqrt{kp_{\mathrm{W}}^2 + 2mp_{\mathrm{W}}^3 - (k+2m)p_{\mathrm{W}}^4} \\ E(s_{\mathrm{BW}}) &= \sqrt{2(k+m)p_{\mathrm{B}}p_{\mathrm{W}} - 4(k+2m)p_{\mathrm{B}}^2p_{\mathrm{W}}^2} \end{aligned} \tag{7.25}$$

where k, p_{B}, and p_{W} are as before, and m is given by

$$m = \tfrac{1}{2}\sum_{i=1}^{n} k_i(k_i - 1) \tag{7.26}$$

where k_i is the number of joins to the ith area.

We can determine these values for our examples. We will take $p_{\mathrm{B}} = p_{\mathrm{W}} = 0.5$ in all cases, since we determined example (b) by flipping a coin. Using the N–S–E–W rook's case neighbors, the number of joins in the grid is $k = 60$. The calculation of m is especially tedious. There are three different types of grid square location in these examples. There are four corner squares each with two joins, 16 edge squares each with three joins, and 16 center squares each with four joins. For the calculation of m this gives us

$$\begin{aligned} m &= 0.5\left[\underset{\text{corners}}{(4\times 2\times 1)} + \underset{\text{edges}}{(16\times 3\times 2)} + \underset{\text{center}}{(16\times 4\times 3)}\right] \\ &= 0.5(8 + 96 + 192) \\ &= 148 \end{aligned} \tag{7.27}$$

Slotting these values into the equations above, for our expected values we get

$$\begin{aligned} E(J_{\mathrm{BB}}) &= kp_{\mathrm{B}}^2 &&= 60(0.5^2) &&= 15 \\ E(J_{\mathrm{WW}}) &= kp_{\mathrm{W}}^2 &&= 60(0.5^2) &&= 15 \\ E(J_{\mathrm{BW}}) &= 2kp_{\mathrm{B}}p_{\mathrm{W}} &&= 2(60)(0.5)(0.5) &&= 30 \end{aligned} \tag{7.28}$$

and for the standard deviations we get

$$\begin{aligned} E(s_{\mathrm{BB}}) = E(s_{\mathrm{WW}}) &= \sqrt{kp_{\mathrm{B}}^2 + 2mp_{\mathrm{B}}^3 - (k+2m)p_{\mathrm{B}}^4} \\ &= \sqrt{60(0.5)^2 + 2(148)(0.5)^3 - [60 + 2(148)](0.5)^4} \\ &= \sqrt{29.75} \\ &= 5.454 \end{aligned} \tag{7.29}$$

and

$$
\begin{aligned}
E(s_{\mathrm{BW}}) &= \sqrt{2(k+m)p_{\mathrm{B}}p_{\mathrm{W}} - 4(k+2m)p_{\mathrm{B}}^2 p_{\mathrm{W}}^2} \\
&= \sqrt{2(60+148)(0.5)(0.5) - 4[60+2(148)](0.5)^2(0.5)^2} \\
&= \sqrt{15} \\
&= 3.873
\end{aligned}
\tag{7.30}
$$

We can use these values to construct a table of z-scores for the examples in Figure 7.5 (see Table 7.2). We can see that unusual z-scores (at the $p = 0.05$ level) are recorded for examples (a) and (c) on all three counts. For case (b), all three counts are well within the range we would expect for the independent random process. From these results we can conclude that examples (a) and (c) are not random, but that there is insufficient evidence to reject the null hypothesis of randomness for example (b).

How Probable Is Your Example?

Repeat the analysis above using the counts that you obtained for the chess board of grid cells. You can, of course, use the expected values that we have calculated.

A number of points are worthy of note here:

1. Care is required in interpreting the z-scores. As in case (a), a large *negative* z-score on J_{BW} indicates *positive* autocorrelation since it indicates that there are fewer BW joins than expected. The converse is true. A large *positive* z-score on J_{BW} is indicative of *negative* autocorrelation, as in case (c).

Table 7.2 *Z*-Scores for the Three Patterns in Figure 7.5 Using the Rook's Case

	Example		
Join type	(a)	(b)	(c)
BB	2.200	−1.650	−2.750
WW	2.200	0.733	−2.750
BW	−6.197	1.291	7.746

2. It is possible for the three tests to appear to contradict one another.
3. The scores for J_{BB} and J_{WW} in case (*c*) are problematic. This is because with $J_{BB} = J_{WW} = 0$, the distribution of the counts clearly cannot be truly normal, given that there is no lower bound on the normal curve. This is a minor point in such an extreme example. In fact, the predicted results are only truly normally distributed for large n, with neither p_B nor p_W close to 0. This is analogous to the difficulties encountered in using the runs test with n_1 or n_2 close to zero.
4. Generally, the choice of neighborhoods should not affect the scores and the decision about the overall structure, but this is not always the case. For example, if we count diagonal joins, we get the counts in the second column of Figure 7.5. The results these produce for example (*c*) are particularly interesting. In this case, $k = 110, p_B = p_W = 0.5$ as before, and the calculation of m is given by

$$\begin{aligned} m &= 0.5\left[\underset{\text{corners}}{(4\times 3\times 2)} + \underset{\text{edges}}{(16\times 5\times 4)} + \underset{\text{center}}{(16\times 8\times 7)}\right] \\ &= 0.5(24+320+896) \\ &= 620 \end{aligned} \tag{7.31}$$

This gives us

$$\begin{aligned} E(J_{BB}) &= E(J_{WW}) = 110(0.5)^2 = 27.5 \\ E(J_{BW}) &= 2(110)(0.5)(0.5) = 55 \end{aligned} \tag{7.32}$$

and

$$\begin{aligned} E(s_{BB}) = E(s_{WW}) &= \sqrt{110(0.5)^2 + 2(620)(0.5)^3 - [110+2(620)](0.5)^4} \\ &= \sqrt{98.125} \\ &= 9.906 \end{aligned} \tag{7.33}$$

$$\begin{aligned} E(s_{BW}) &= \sqrt{2(110+620)(0.5)(0.5) - 4[110+2(620)](0.5)^2(0.5)^2} \\ &= \sqrt{27.5} \\ &= 5.244 \end{aligned} \tag{7.34}$$

Table 7.3 *Z*-Scores for Example (*c*) in Figure 7.5 Using the Queen's Case

Join type	*z-score*
BB	−0.252
WW	−0.252
BW	0.953

This gives us the *z*-scores for case (*c*) shown in Table 7.3. Despite the clearly nonrandom appearance of the grid, *none* of these indicate any departure from the independent random process! This demonstrates that the joins count statistic must be treated with some caution.

5. In this example we have assumed that p_B and p_W are known in advance. Normally, they are estimated from the data, and properly speaking, much more complex expressions should be used for the expected join counts and their standard deviations. These are given in Cliff and Ord (1973) and are equivalent to the difference between the card-drawing and coin-flipping examples described in the discussion of the simple runs test.
6. Finally, to make the calculations relatively simple, in this example we have used a regular grid. As the worked example that follows shows, this isn't actually a restriction on the test itself, which easily generalizes to irregular, planar-enforced lattices of areas.

7.5. FULLY WORKED EXAMPLE: THE 2000 U.S. PRESIDENTIAL ELECTION

We can perform a similar calculation on the U.S. presidential election results from November 2000. These are mapped in Figure 7.6. Visually, the state electoral outcomes seem to have a geography, but is this pattern significantly different from random? Only after we have shown that it is can we be justified in hypothesizing about its possible causes. The data are binary (Bush–Gore), so that a joins count test can be applied.