The distribution of distance of certain place-name types to Roman roads[1]

Keith Briggs

In 1964,'the Viatores' reported a study of the distribution of the place-names Coldharbour and Caldecote in the south-east midlands, which concluded, 'thus it is clearly demonstrated that in this region the names are not merely of common occurrence along the Roman roads, but are, in fact, to be closely connected to such roads'.[2] This question was studied in greater depth for Coldharbour names by Ogden, who pointed out that these speculations in fact go back to the eighteenth century.[3] Ogden also found a positive correlation—Coldharbours are closer to Roman roads than randomly distributed points. We might reasonably conclude that Coldharbours and Caldecotes are travellers' shelters, and that the Roman roads on which they are situated were still in use at the time the names were given. However, this conclusion became doubtful when Richard Coates gave his definitive treatment of Coldharbour, showing that the name has a mostly post-1600 origin, and no connection to Roman roads is at all likely.[4] It is easy to see how such apparently paradoxical conclusions can be reached—it might be that *all* settlements are closer to Roman roads than randomly distributed points, simply because Roman roads are concentrated in regions better suited to occupation.

Similar claims that other place-name types occur more frequently near to Roman roads have been made:

[1] I thank Trevor Ogden for alerting me to his Coldharbour paper, J. David Young and Jill Bourne for providing data, and Jill Bourne for comments on an earlier draft of this paper.

[2] The Viatores [a group of fieldworkers led by I. D. Margary], *Roman roads in the south-east midlands*, (London, 1964), p. 351.

[3] T. Ogden, 'Coldharbours and Roman Roads', *Durham University Journal*, 59 (1966) 13–14.

[4] Richard Coates, 'Coldharbour—for the last time?', *Nomina*, 8 (1984), pp. 73–78.

- Wilson discusses place-names derived from OE *hearg*, believed to denote pagan religious sites, and for each of his examples he gives a distance to the nearest Roman road.[5]

- Cox states that 'place-names containing the OE element *hām* "a village, a collection of dwellings" are closely related in their distribution to the system of Roman roads and ancient trackways in the Midlands and East Anglia'.[6]

- Gelling states that 'twenty-four [of the twenty-eight place-names in England which derive from Old English *wīchām*] are situated on, or not more than a mile from, a known Roman road. This is not a random distribution'.[7]

- Cole argues that place-names derived from OE *netel* 'nettle' are associated with Roman settlements and roads.[8]

- Cole says of four place-names derived from ON *nata* 'stinging nettle' that 'The pattern of use is similar to that of OE *netel*, namely near Roman forts and old routeways'.[9]

Whilst the conclusions of these papers are very likely correct, I believe that in no case was a rigorous statistical analysis done. This is in fact a delicate question, because to draw a map and look for association by eye is notoriously unreliable. The purpose of the present article is to describe how

---

[5] D. Wilson, 'A note on OE *hearg* and *wēoh* as place-name elements representing different types of pagan worship sites', *Anglo-Saxon Studies on Archaeology and History*, 4 (1985), 179–183.

[6] B. Cox, 'The Significance of the Distribution of English Place-Names in *hām* in the Midlands and East Anglia', *Journal of the English Place-name Society*, 5 (1972–3), 15–73.

[7] M. Gelling, 'English place-names derived from the compound *wīchām*', in *Placename Evidence for the Anglo-Saxon Invasion and Scandinavian Settlements*, edited by Kenneth Cameron (Nottingham, 1977).

[8] A. Cole, 'The use of *Netel* in place-names', *Journal of the English Place-Name Society*, 35 (2002–3), 49–58.

[9] A. Cole, 'The use of ON *nata* in place-names', *Journal of the English Place-Name Society*, 36 (2003–4), 51–53.

proper statistical analysis can be done for this type of problem, which I will do with the omission of mathematical technicalities. The hope is that with modern computing methods, much more precise statistical information will be obtainable. But my main aim here is to give a suitable methodology by which studies of this kind can be done rigorously, rather than draw any new conclusions concerning the distribution of these name-types.

Statistical hypothesis testing

We consider problems in which there is some element of randomness. This immediately implies that any conclusion from statistical analysis cannot be absolutely certain. At best we hope to find support in a dataset for or against a hypothesis. This is normally done by stating a null hypothesis, that is, the hypothesis that there is no effect. We would then analyze our data with the hope of finding evidence that the null hypothesis is false, and this is called rejection. If this is not the case, so that the null hypothesis is not rejected, then it is said to be retained (rather than accepted). The latter case means that one of two possibilities is true, but it is not possible to determine which: either the null hypothesis really is true; or, the null hypothesis is false, but there is not enough information in the data to establish this. The consequence of this could be that we try to obtain better data, which might then reject the null hypothesis, or to formulate a different null hypothesis. In all cases, a confidence level has to be assumed, often 95%, so that a rejection actually means ‘this experiment or observation is unusual, in that it shows features that we would expect to see less often than in 1 out of 20 trials if the null hypothesis were true’.

Many varieties of such tests are possible, and these have different strengths. To give an example a little closer to the type of problem which we wish to tackle, suppose as part of a public health study it was desired to know whether 30-year-old men in London were taller than 30-year-old men in Newcastle. Assuming that we cannot measure all 30-year-old men, we have to take a random sample and this already introduces some uncertainty. Next we have to decide exactly what ‘taller than’ means. It could be that we are only concerned with the mean height (the ordinary arithmetic average), or it could be the mode (the most frequently occurring height), or the median (the height which 50% of the men are below), or it could be a question about the full distribution of heights. It is possible that the null hypothesis that the heights are the same is rejected for some of these so-called test statistics, but retained for others. It is also possible for two

different distributions to have the same mean, but the converse (differing means, but the same distribution) is not possible. So it is clear that in analyzing distances to Roman roads, an exact statement of the test we are making is necessary.

I shall choose a stronger test than just using the mean distances, on the grounds that if we see an effect, we want to be as sure as possible that it is real. This will require larger datasets than would be required by tests purely based on the mean, and runs the risk of the test results being inconclusive. But this is better than a false conclusion which is an artefact of a too-small dataset. I shall in all cases use the *cumulative distance distribution*. As an example see Figure 1. I will explain how this is computed later; for now note that this graph shows, for each value on the horizontal axis (which is the square root of a distance measured in kilometres), the fraction of places with a particular name-type which are less than that distance to the nearest Roman road. (The square root does not change any essentials, but merely makes the graph easier to interpret.) In this example, the lower curve U is smooth because it is computed from one million points. The upper curve (V, the villas) is less smooth, but the fact that it lies everywhere above the smooth curve (the reference distribution) is evidence that villas lie closer to Roman roads than randomly distributed points. It is the job of statistical hypothesis testing to tell us whether this evidence is significant or not. It should be clear that these graphs must always go from 0.0 on the vertical axis (since no distance can be less than zero) to 1.0 on the vertical axis at large distances (since no distance can be greater than the maximum distance of any point in England to a Roman road, which is about 42km). Thus, all information on which we shall base conclusions is contained in the shape of this graph, especially in its middle region, where the frequency is close to one half.

Note the critical point that our methods do pairwise comparisons of distributions; that is, in any test there will be a reference distribution (considered known) and a target distribution (the dataset to be tested). But the test itself considers these as equals; which is the reference, and which the target, is arbitrary. The reference distribution might be a set of uniformly distributed random points; the target distribution could be for example the Coldharbours. It makes no sense to make claims about the target distribution without specifying the reference distribution as well.

My tests will be Kolmogorov-Smirnov tests, a standard for this type of problem.[10] The method uses as a test statistic the *maximum vertical distance* between the two cumulative distributions being compared, but most statisticians will simply perform the hypothesis test by using a built-in function in a statistical software package, such as R.[11] This test has proven itself to be one of the best for this type of work. There are other tests, such as Anderson's test. When applied to my data, I found that Anderson's test led to exactly the same conclusions. It is very likely that any reasonable test would give the same results.

The datasets

The first problem concerns what is to be counted as a Roman road. The standard reference is Margary.[12] This includes roads for which there is sufficient archaeological, documentary, topographical, or toponymic evidence, but there are still doubtful cases; this is so especially for missing segments between established roads. These segments may never have been completed. As well as the roads classified and given reference numbers by Margary, there are extra roads established by more recent archaeology, and various additional roads that have been postulated to exist on the basis of indirect evidence. I have chosen to do all tests twice—once with the Margary roads alone (which have M numbers), and once with all roads, the roads additional to Margary's having X (for extra) numbers: see Figure 4. Thus if any hypothesis is rejected with respect to both Roman road datasets, it is highly unlikely to be true. I obtained the road data for England from English Heritage, which maintains a database as part of the National Monument Record. I processed this data into a suitable form to use in my computations. My own tests against OS maps established that this data was accurate to within at least 10 metres. In all calculations, I used the GB National Grid coordinate system.[13] All computations were done with my own programs, thus avoiding the use of any commercial software.

---

[10] A. Stuart, K. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics*. (London, 1999), pp. 25.37–25.43.

[11] Details of this free software package are available at <http://www.r-project.org>.

[12] I. D. Margary. *Roman Roads in Britain*, (London, 1967).

[13] Ordnance Survey, *A Guide to Coordinate Systems in Great Britain*, version 1.9. (Southampton, 2008); accessible from <http://www.ordnancesurvey.co.uk>.

For the reference distributions and the sets of points to be tested against the Roman roads, I have chosen the cases detailed in Table 1. As purely reference data, I use dataset U, a million points uniformly distributed with respect to the National Grid. This indicates complete indifference to the Roman roads and to the topography in general. As a more realistic reference set, I generated random points ignoring Roman roads, but biasing the points towards low-altitude sites, since this better represents the true settlement distribution (L).[14] Set D of the Domesday Book vills is also useful as representing the early settlement density. Set V is a large collection of Roman villa sites, which might be expected to be close to Roman roads. The *-hām* names from Cox[15] are plotted in Figure 5.

An issue with the place-name datasets is the accuracy we can claim in the locations. For Coldharbours and Roman villas there is no problem, because these can be located to about 10 metres. Village locations typically have a maximum uncertainty of about 1km, because of doubt about the right point to take as the centre. This uncertainty can be considered as random error, since it is equally likely to be in any direction, and fortunately statistical analyses such as mine will tend to average out the effects of this error. In any case, when distributions differ, the differences show up at distances larger than 1km, so I believe these potential problems are not serious. There are no precision problems, as I use 6 figure grid references, which are precise to 100 metres. All tests are confined to the area of England and Cornwall; the *-hām* tests are further confined to East Anglia and adjacent parts of central England.

Note that the distance distributions are typically skewed towards zero (that is, small values are more likely), and have a long slowly-decaying tail towards large values of distance. For such distributions, the mean is not a good measure of central tendency (that is, the tendency for data to cluster near a single most frequent value), and I have preferred to use the median (the value below which is 50% of the data) in the following analyses.

---

[14] The data was generated by this heuristic: a point is chosen uniformly and randomly in England. If it is at elevation less than 50m, it is retained. Other points are retained with the following probabilities: 50m–100m, 0.7; 100m–150 m, 0.5; 150m–200m, 0.2. 200–300m, 0.1. Points with elevations above 300m are rejected. This process is repeated until 1 million points have been generated. The resulting distribution is not very different to that of the Domesday vills.

[15] Cox, 'The Significance of the Distribution of English Place-Names in *hām* in the Midlands and East Anglia'.

code	size	description	data source	mean		median	
				M	MX	M	MX
U	10^6	uniform random points		6.10	5.46	4.32	3.74
L	10^6	low-altitude random points		5.85	5.09	4.13	3.45
D	13448	Domesday vills	SN5694	6.13	5.35	3.95	3.35
V	1296	Roman villa sites	J. David Young	3.88	3.26	1.98	1.74
H	353	OE *-hām* names	Cox	3.14	3.14	2.40	2.40
C	222	Coldharbour names	Ogden	4.78	4.28	2.73	2.42
K	68	Kingston names	Jill Bourne	6.00	5.31	3.63	3.32
W	54	OE *wīchām* names	Gelling	3.03	2.33	1.61	1.10
N	24	OE *netel* names	Cole	7.60	6.86	1.94	1.94

Table 1: All datasets used, with mean and median distances in km with respect to both M and MX Roman road sets. Data sources as single surnames refer to papers cited earlier. Full names refer to unpublished data which I obtained by private communication. SN5694 is the Arts and Humanities Data Service Electronic Edition of the Domesday Book.

	U	L	D	V	C	K	W
L	<						
D	<	<					
V	<	<	<				
C	<	<	<	>			
K	<	<	<	>	0.09		
W	<	<	<	0.50	0.13	<	
N	0.16	0.16	0.24	>	0.47	0.47	0.08

Table 2: Kolmogorov-Smirnov test results for datasets with respect to Margary (M) roads. Only the lower triangle is shown, since the table is symmetric. Probabilities are only shown when they indicate retention of the null hypothesis. When the null hypothesis is rejected, the medians are compared, and '<' is inserted if the median of the dataset in the left column is less than the median of the dataset in the top header; otherwise '>' is inserted. Thus, for example, we have that L < U, which means that low-altitude random points are closer to Roman roads than uniformly distributed points, and, by symmetry, it must be the case that U > L. In most cases of rejection, the probabilities are tiny—10^{-6} or less. Note that dataset N (OE *netel*) has only 24 items and as might be expected, most tests are not rejected.

	U	L	D	V	C	K	W
L	<						
D	<	<					
V	<	<	<				
C	<	<	<	>			
K	<	<	<	>	0.09		
W	<	<	<	0.50	0.13	<	
N	0.16	0.16	0.24	>	0.47	0.47	0.08

Table 3: Kolmogorov-Smirnov test results for datasets with respect to Margary (M) and extra (X) roads.

Analysis method and results

My procedure is to perform the following two steps for the Margary-only Roman roads (M) and for the full set of Roman roads (MX):

1. For each dataset I computed its distribution of distance to the nearest Roman road. This is done by computing for each point in the dataset the nearest Roman road (in terms of perpendicular distance). The Roman roads are represented as a very large collection (about 30,000) of short straight-line segments, so this computation involves calculating the distance to each segment, and then finding the minimum over all segments.

2. For each pair of distance distributions computed as above, I consider the null hypothesis to be that the two datasets being compared are both random samples from the same underlying distribution. I then computed the Kolmogorov-Smirnov test statistic and used this to look up a standard table of test statistics. The result is a number between 0 and 1, which can be interpreted as the probability that the two datasets come from the same distribution. If this number is small (say, less than 5%), then the probability is negligible and we reject the null hypothesis. If this number is much greater than 5%, then the null hypothesis is retained and we have a positive result in the sense that the claim that the two datasets are drawn from the same distribution has not been disproven. In this case, further investigation is needed to determine whether this result is genuine. I present the pairwise results as a table, only entering the probability if it implies retention of the null hypothesis. In the case of rejection, I compare the medians and use this to decide which distribution is to be regarded as ‘closer’ to Roman roads.

Conclusion

With respect to the Margary roads (M), all but 9 out the 28 pairwise tests resulted in rejection of the null hypotheses. The results were only slightly different with respect to the Margary and extra roads (MX). This is a satisfying result, as it suggests the results are not too sensitive to the choice of roads. The most interesting results are these:

1. The Coldharbour names are confirmed, as in the analysis of Ogden, as lying closer to Roman roads than uniform random points, but not closer than Roman villas. This conclusion is subject to the caveats mentioned in the introduction.

2. The largest number in Table 2 is 0.5, suggesting that the strongest association of place-name types and Roman roads is between Roman villas and *wīchām* names (though it should be emphasized again that this does not prove a causal association). I plot the cumulative distance distributions in Figure 2. This suggests a hypothesis worthy of further archaeological investigation: the *wīchām*s are an independent random sample of the villa sites, or, in other words, *wīchām*s are survivals of Roman settlements sites into the Anglo-Saxon period, and this is why they received their name.

3. The Kingston names are shown to have a statistically significant bias towards Roman road at distances below 2km (Figure 3). Further work may provide a causative explanation of this.

I hope to have demonstrated that at least one area of name-studies is susceptible to rigorous statistical analysis, and I will have succeeded in the aims of this paper if when statistics are applied in similar fields in the future, attention is given to the issues which I have raised.

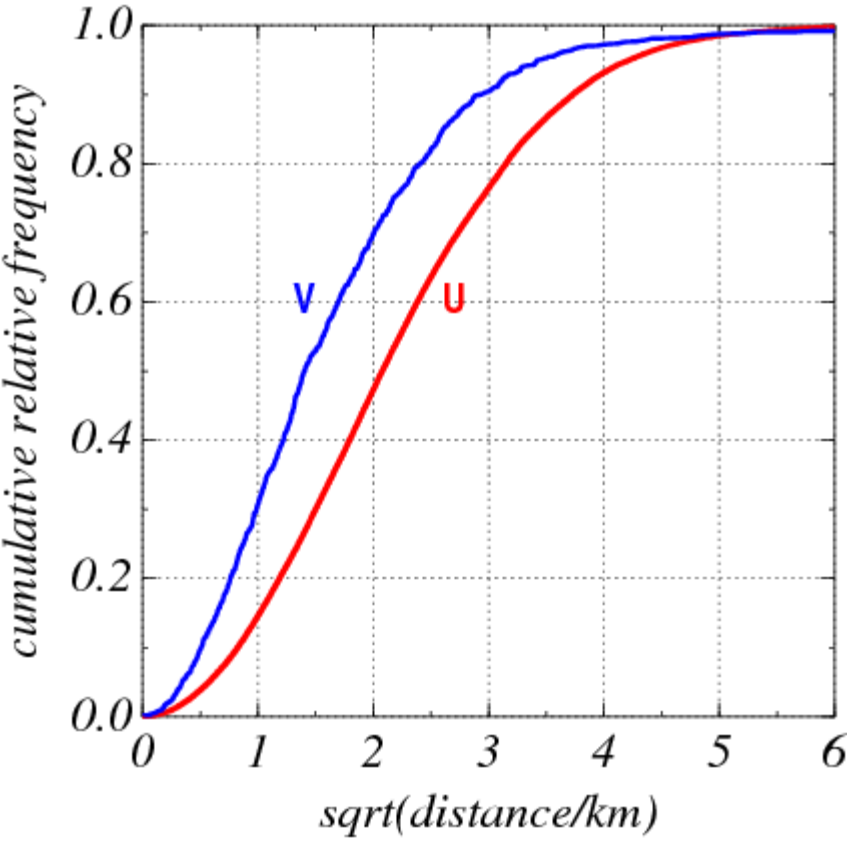

Figure 1: The observed cumulative distance distributions with respect to Margary roads (M) for datasets U (uniform random); and V (Roman villas). The horizontal axis is the square root (*sqrt*) of the distance in kilometres. As an example of reading this graph, the U curve cuts 0.8 at 3.14 on the horizontal axis. This means that 80% of random points are less than 9.86km (3.14 squared) from a Roman road. Similarly, 80% of Roman villas are less than 5.79km (2.41 squared) from a Roman road. The villas are strongly biased towards Roman roads.

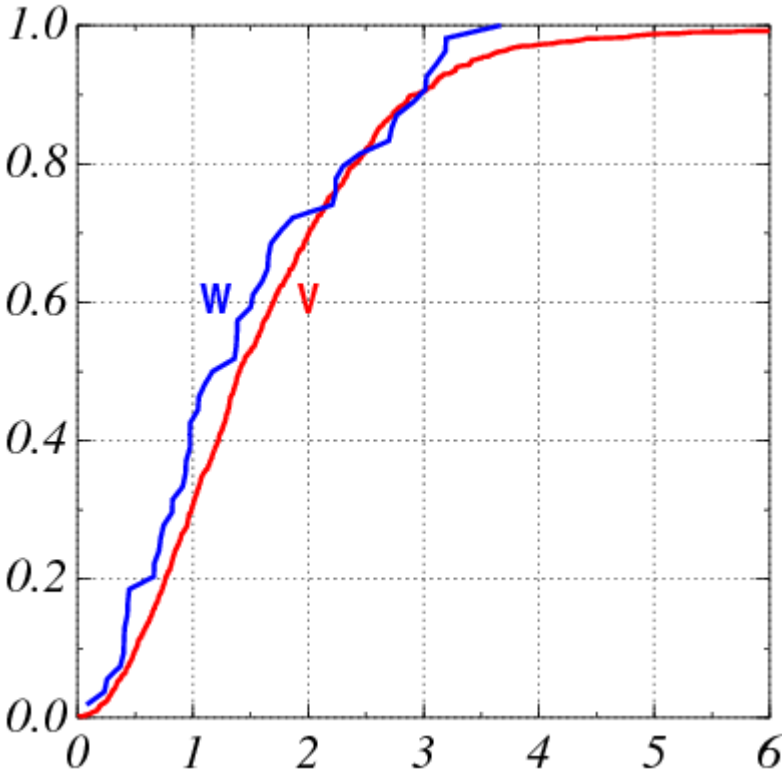

Figure 2: The observed cumulative distance distributions with respect to Margary roads for Roman villas (V) and *wīchām* names (W). The axis labels are as in Figure 1. The distributions are not significantly different; the null hypothesis is retained. The claim that the W points are a random subset of the V points is not disproven.

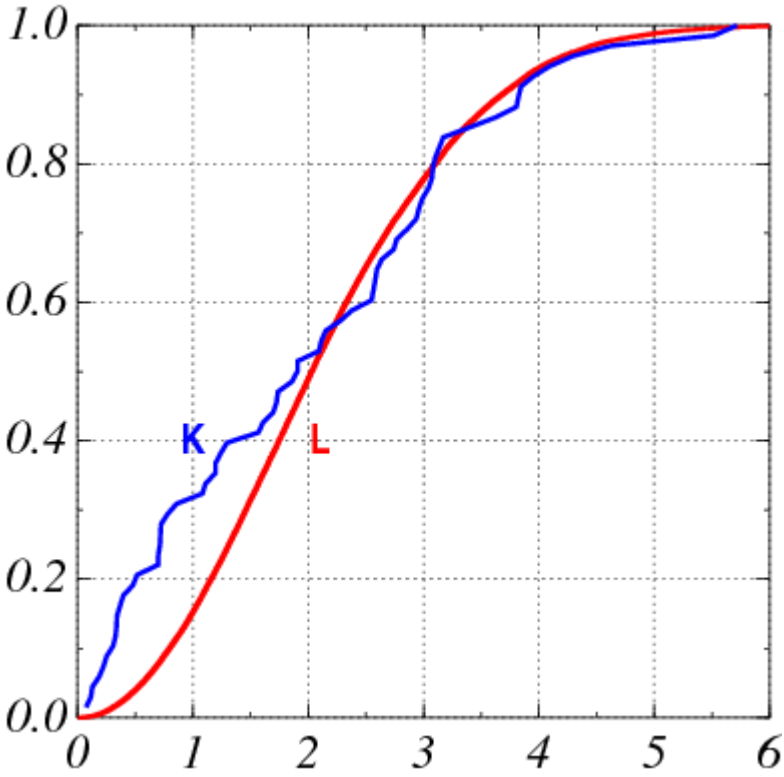

Figure 3: The observed cumulative distance distributions with respect to Margary roads for low-altitude random points (L) and Kingston names (K). The axis labels are as in Figure 1. The distributions are significantly different (the null hypothesis is rejected), and this is reflected in the slightly lower median distance for Kingston names as shown in Table 1.

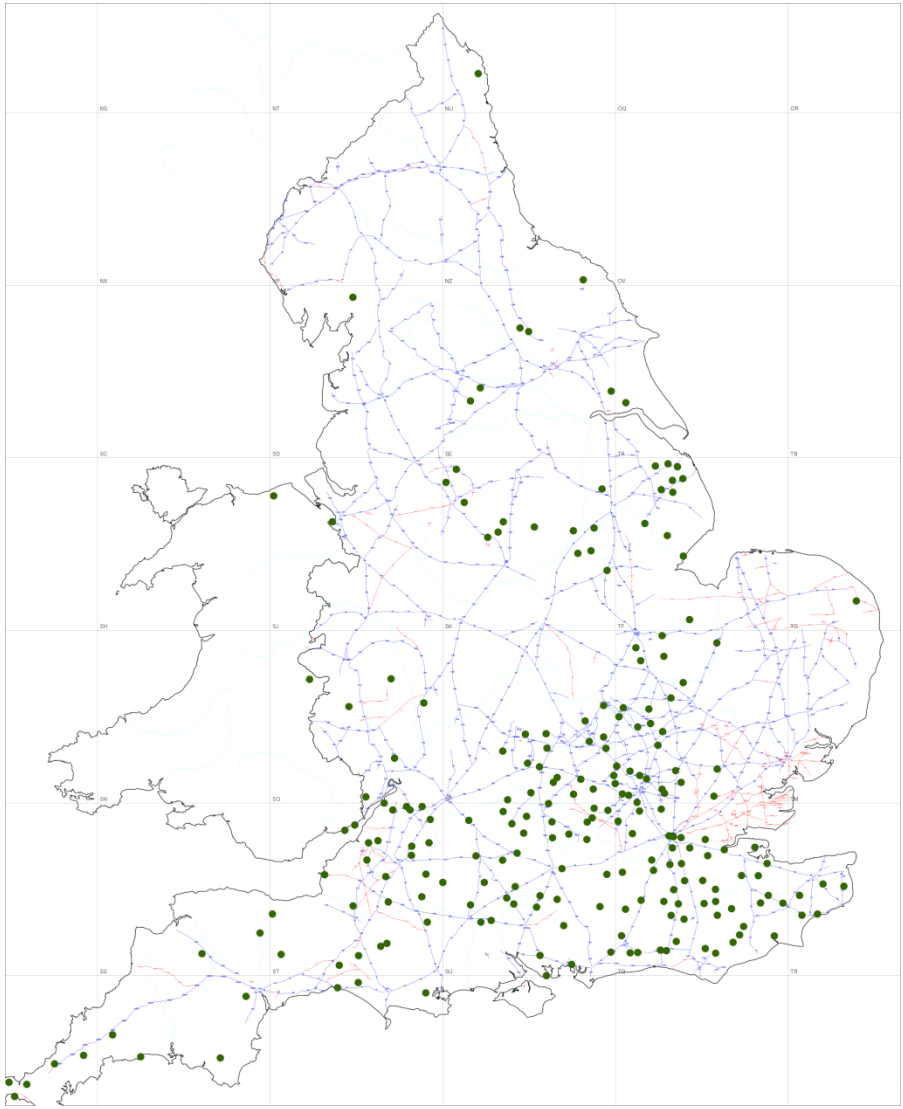

Figure 4: All Roman roads used in this study, and the Coldharbours.

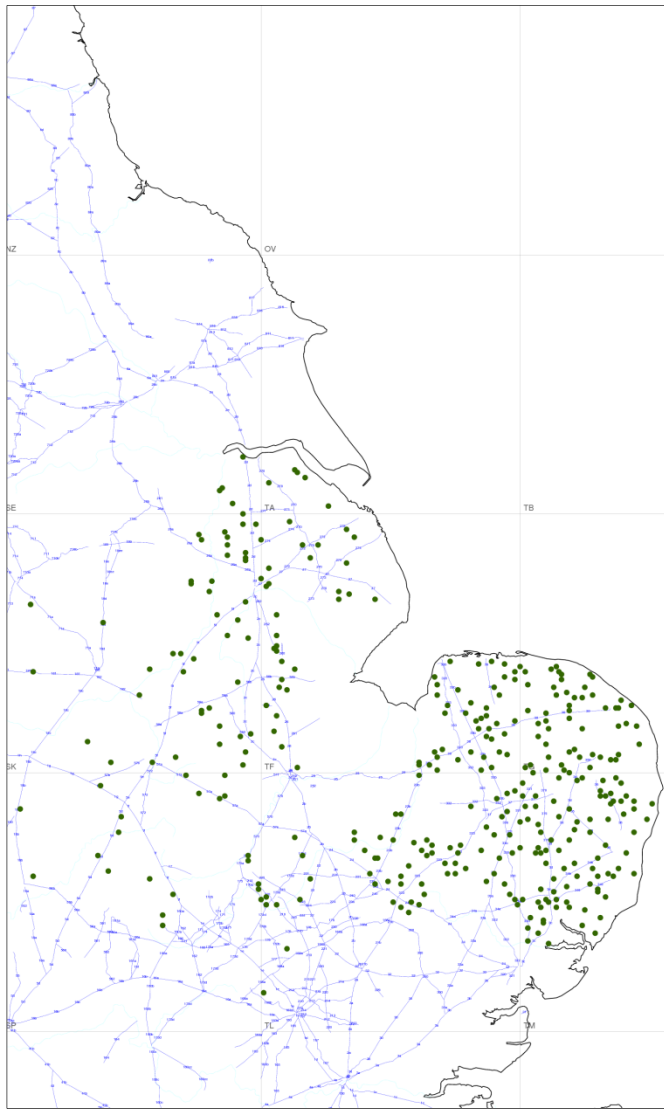

Figure 5: *-hām* place-names in Eastern England, and Margary roads.