Comparison Between Geocoded and Road Surrogate Data Sets

Q. Could you please explain more about the PNR customer location data you analyzed?

A. Yes. PNR provided us with two data sets. The first was a complete copy of PNR's road surrogate data set for Kansas, including the longitude and latitude of each estimated customer location. The second was a file reflecting the effects of running the clustering portion of the FCC model using PNR's second data set, which relies partially on road surrogate algorithms and partially on actual customer locations. We were able to use the latter file in running the FCC model, but the detail provided did not allow us to analyze or map the individual locations. The road surrogate algorithm (used in both PNR data sets) spaces customer locations (households and businesses) within each census block uniformly along roads within that block. As explained by the FCC, "the total number of surrogate points is ... divided by the computed road distance to determine spacing between surrogate points. Based on that distance, the surrogate customer locations are uniformly distributed along the road segments". [Inputs Order, ¶ 43]. Since customers typically are located along roads, this procedure is quite logical. The major drawback with this approach is that it assumes customers are spaced uniformly along roads, even though they are not.

Particularly in rural areas, there may be long stretches of roads without any customers. All of the customers may be clustered in a relatively small number of locations within each census block. If customers are clustered along certain portions of the roads, or if they are concentrated along certain roads and not others, the road surrogate process will not accurately represent reality. By uniformly spreading customers along every road, the road surrogate algorithms force the FCC model to

send cable to every part of each census block. In reality, network engineers don't need to send cable to anywhere except where customers are actually located. In urban areas this discrepancy between reality and assumptions may not be tremendously significant, since customers may truly be located on nearly every street, but in rural areas the gap between algorithm and reality may be severe. In some rural areas, there are long stretches of empty roads, yet the road surrogate algorithms will assume they contain customers, and thus force the FCC model to send cable down these roads.

The second data set provided by PNR partially avoids this problem, because it places some customers at their actual locations – where known – and it only relies upon the road surrogate approach for instances where the actual customer locations are not known. This approach is considerably more accurate in urban areas, where many customers can be precisely located. Unfortunately, the improvement occurs in urban areas, where it has the least impact. In rural areas -- where geographic accuracy is most important -- it doesn't offer much improvement, because the actual locations aren't known and the road surrogate algorithms are relied upon instead.

Not only does the second PNR data set not offer great improvements in the rural areas where improvement is most needed, but PNR imposes greater proprietary restrictions on the second data set. In particular, PNR refused to send us a copy of the actual customer location latitude and longitude points, and prevented us from analyzing this data in detail (e.g. comparing it to other data sets). Accordingly, most of our analysis has been concentrated on the pure road surrogate data set. To further examine the impact of actual geocoded customer locations, we relied upon an alternative data vendor, Select Phone. We had previously acquired a copy of their geocoded customer location data, and it was not subject to the types of restrictions imposed by PNR.

Q. Have you compared the PNR road surrogate data with the actual customer locations geocoded by Select Phone?

A. Yes. As mentioned earlier, our firm owns a copy of Select Phone's geocoded data set for Kansas customers, derived from white page listings. This is very similar to PNR's geocoded data set, except that it provides us with the actual latitude and longitude points, and thus it can be analyzed in detail. Map 1 allows one to visually compare the two types of customer location data. We selected Valley Center for this comparison, because this is a wire center where good geocoded data was available for more than 90% of the white page listings, yet it was not an exclusively urban area.

Each blue dot in Map 1 represents an actual customer location, as derived from white page listings in the Select Phone geocode data set; note that the blue dots are not spaced evenly, because actual customer locations tend to cluster in certain areas while other areas remain empty. In contrast, the green dots show the effect of PNR's road surrogate algorithms. Each green dot represents an estimated customer location. Note that the green dots tend to be spread uniformly along the roads. The slight degree of non-uniformity in spacing generally occurs where a road forms the boundary between two different census blocks. Although the spacing is always uniform, different rates of spacing are used in each census block, and some roads include dots from two census blocks (and thus reflect two different spacing intervals).

Another point that can be observed by comparing the blue and green dots is that the PNR road surrogate algorithm places many green dots at the far edges of the wire center–along roads at or near the boundary. Because the road surrogate method distributes uniformly along all roads in the wire center, and because roads are often used to estimate wire center boundaries, many surrogate customer locations are placed along the wire center boundaries. To the extent these customers are actually located

further inside the wire center, the amount of cable required to connect them at their surrogate location will exceed the amount of cable needed to serve them at their true location.

To more fully appreciate the potential magnitude of this phenomenon look closely at the Northwest corner of the Colwich wire center, as shown on Map 2. The roads follow Section lines and hence each road is one mile long between intersections. Note that there are blue dots spread to the farthest edge of the wire center, whereas the Select Phone data only identifies actual customer locations well inward from this point. In order to connect the farthest green dot in the Northwest corner to the nearest blue dot requires approximately four or five miles of additional cable or 20,000 to 25,000 feet. This additional cable is required by the road surrogate algorithm, but would not be required if one were to rely upon the Select Phone data set. In other words, connecting this one green dot requires more than 2% of the 1.03 million distribution route feet estimated by the FCC default model for the entire Colwich wire center. This vividly demonstrates that a single of spurious location generated by the road surrogate algorithm (or a handful of such spurious locations) can potentially impact the total amount of cable deployed by the FCC model.

This potential problem is not limited to the edges of the wire center. The road surrogate methodology also forces deployment of excess cable in every situation where it places surrogate locations at the far edges of a customer cluster or serving area, even though the customer in question is actually located much closer to the middle of the customer cluster.

Q. Have you analyzed the implications of different customer location data sets in terms of differences in the resulting forward looking network configurations and cable quantities?

A. Yes. One example is contained in Map 3, which provides two views of the same portion of the Valley Center wire center. The upper view is based upon the road surrogate customer locations (green dots) and the lower view is based upon Select Phone's white page listing data (blue dots). These maps also show the amount of distribution cable needed to connect these dots to the Serving Area Interface. The distribution cable is depicted by the narrow purple lines. (Feeder cable is depicted with wider, lighter purple lines, and roads are depicted with thin grey lines.) A careful visual comparison of the two networks shows that it takes more cable to connect the surrogate locations than it does to connect the Select Phone locations.

Maps 4 and 5 allow the same type of comparison for the entire Valley Center wire center (at a smaller scale). Overall, we found that connecting the surrogate locations to the SAIs (depicted as numbered blue triangles) takes hundreds of thousands of additional route feet of distribution cable compared to the corresponding amount needed to connect the Select Phone locations to the same SAIs.

Q. Do similar problems occur in other wire centers?

A. Yes. As I explain later in my testimony, over the past several months we examined 14 wire centers in detail; they all showed indications of potential problems due to the uniform spacing of the surrogate locations. Unfortunately, in most cases the Select Phone data only places customers accurately within the town center. Outside of town, the actual locations are not known, and thus we can't rely upon the Select Phone data to measure the extent of the problem.

In some cases, however, a simple visual inspection confirms that there are problems with the road surrogate data. Maps 6 and 7 provide an example of this in the Desoto wire center. Map 6 shows the PNR road surrogate locations, represented again by the green dots, along with the corresponding cable routes (purple lines). Map 7 shows the analogous information using the Select Phone locations. As is typical of many Kansas wire centers, a large number of actual locations could not be accurately geocoded, and thus the Select Phone data understates the amount of cable that actually be needed to serve this area. Nevertheless, a careful review of these two maps can be quite informative.

In reviewing Map 6, notice that the PNR surrogate algorithms place green dots along mile-long stretches of roads which may not actually contain any customers. The pattern of uniform spacing generated by the road surrogate algorithms is particularly suspicious along country roads which are filled with closely spaced green dots. For example, look closely at the spacing of green dots in the vicinity of SAI's number 3 and 6, just above the legend on Map 6. Compare this to the same area on Map 7, where the Select Phone geocoded customer addresses are depicted by blue dots. Very few customer locations are geocoded in this particular area.

The Select Phone data shown on Map 7 may be misleading, since many of the Desoto addresses couldn't be geocoded. However, the type of close spacing of customers shown on Map 6 seemed suspicious to us. This type of close spacing of customers more typically occurs in urban areas, where the street grid is closer together. When driving on long stretches of roads in rural areas, where the intersections between roads are a mile or more apart, one doesn't often see a home or business spaced every few hundred feet.

In visually inspecting this portion of the Desoto road surrogate data, we strongly suspected that the customers were not, in fact, where the PNR algorithm places them. To check out this hypothesis, we obtained a very detailed U. S. Geological Survey map of this particular area. Map 8 shows a blow up of the relevant portion of this USGS map. For convenience, we have superimposed the PNR surrogate locations (once again depicted as green dots).

As it turns out, virtually all of the customers in this area are actually located in a dense cluster of military housing units, which are the small black rectangles shown in the lower right corner of the USGS map. The PNR algorithm spread these household units along all of the roads in the census block, rather than concentrating them in the relevant location. The actual buildings are tightly clustered in the lower right corner of Map 8. In fact, the USGS map does not show any structures located along the adjacent mile-long stretches of country roads–contrary to the PNR road surrogate data.

The Valley Center, Desoto and Colwich examples just discussed all confirm that actual geocoded customer locations tend to be more tightly clustered than the road surrogate locations. In turn, this suggests that the cable quantities and monthly loop costs generated by the FCC model are too high.

Q. Have you looked at the results of the two PNR data sets as a whole?

A. Yes. During the course of our work in this proceeding, we compared the monthly costs and cable quantities generated by the FCC model using each PNR data set. This analysis was performed using a slightly different set of FCC model inputs, and thus the data is not directly comparable to that shown elsewhere in my schedules. Nevertheless, the general pattern we found is still relevant. As shown, on Schedule 9, the PNR data set which includes some actual customer locations resulted in lower cost estimates in

139 of 167 SWBT wire centers. Similarly, as shown on Schedule 10, the PNR data set which includes some geocoded locations resulted in lower cable quantities in the great majority of the SWBT wire centers–134 of 164. Overall, the total quantity of cable deployed by the FCC model declined by approximately 6% when using the PNR data set that includes some actual customer locations. Not surprisingly, we found that this reduction occurs almost entirely within the distribution category.

In considering these results, it is important to remember that the difference is due to differences in the two PNR data sets, yet they both rely almost exclusively on the road surrogate alogrithm in rural areas. Clearly, if actual customer locations were also known for rural areas–rather than just within the town centers–the reductions in cable quantities and monthly cost would be even more dramatic. The difference between the two data sets is limited almost exclusively to urban and suburban areas which have a high proportion of addressable road segments and thus where PNR was able to geocode actual locations. In lower density rural areas the potential impact of geocoding accuracy is greater, but this impact isn't realized because both data sets are essentially the same.

The paradox is that low density, high cost areas are the focal point of the USF analysis, and the area where accurate locations are most important, but these are the areas where neither PNR data set contains actual customer locations. Stated differently, since the PNR data set with some actual locations results in a noticeable reduction in cable quantities, one can anticipate that a data set consisting entirely of actual locations (including all rural areas) would reduce cable quantities by a even larger margin–perhaps by as much as 20% or more statewide, with even larger reductions in some wire centers.

Consider again the example of Valley Center, which is a wire center where PNR's actual geocoding process was probably quite successful. The second PNR data set results in 21.1% less cable being deployed by the FCC model in this suburban wire center, when compared to the results using PNR's pure road surrogate data set. If a complete set of actual customer locations were available for even lower density wire centers, even greater discrepancies in cable quantities might arise when comparing the road surrogate results to the actual location results. The problem is that no such complete set of actual location data exists at the present time. Thus, the challenge is to determine how best to run the FCC model in the absence of an ideal customer location data set.