Ⅳ－184　データ領域の特性の予測誤差への影響について

北海道大学工学部　正員　○山形耕一

1. はじめに

交通計画において予測は，計画手段の配列の結果としての予測年次の交通諸量を出力し，意志決定のための情報を提供する機能をもつが，現象の観測，モデル設定，予測のプロセスを通じて諸々の原因のため予測誤差がもたらされる。本研究では，これらの予測誤差を発生させる要因について分析し，このうち従来議論されることの少なかったデータ領域の拡がり，データ数および予測領域との関係が予測精度に及ぼす影響を考察する。ここで，データを多次元空間における点の集合と理解し，点の分布する領域をデータ領域と呼ぶことにする。

2. 交通モデルにおける予測誤差

交通予測における誤差の混入は，データ収集，モデル設定，予測の3段階に分けて考えることができよう。データ収集段階としては，交通現象に影響する要因が多岐に亙り，その全てを調査し切れないことや価値観，効用の如く計測の困難・不能な要因を含むことはモデル設定を制約する。また実査上および標本抽出による調査誤差は，モデルのパラメータ誤差を通じて予測誤差をもたらす。これらに加えて，現実に起こった事象しかデータとして採取できない制約がある。個々の要因のデータ領域における値の分布域（値域と呼ぶ）は調査時点以前に現実に起こった事象における値の範囲に限られる。また，要因同志は互に相関をもつため，データは空間全体に散在するのではなく，相関関係に従って局在する傾向がある。この限られたデータ領域を対象に現象把握を行わなければならないため，データ領域外への予測性の制約，モデルの不安定性，さらには値域の小さい要因の影響の分析の困難さ等の問題が生じる。

モデル設定段階では上記のデータ収集上の制約に基づく問題点に加えて，複雑な現象を操作可能なモデルで記述する困難さを伴う。この困難さの一因としては，交通現象が混合的な性質を持っていることが考えられる。交通は個人の，さらに個々の時点での交通行動の集合であり，個人による価値観の差異や個々の時点の行動環境の違いによる価値意識構造の差異が内在している。このため，交通行動に対する要因の作用性が異る現象をマクロ的に記述することが要求されているのである。従って，モデルは平均的な現象を記述することとなり，モデルで説明し切れない残差を生じる。すなわち，モデルにはデータの誤差や残差等の撹乱項が含まれるが，撹乱項を含むデータから関係式を安定的に抽出するために必要なデータ数は明らかにされているとは言い難い。従来のサンプル理論は集団における要素の数標識の推定に関する理論であり，関係式の安定的な把握のためのサンプル理論が今後開発されることが必要であろう。

予測段階で誤差をもたらす原因としては，先ず予測のフレーム設定の問題がある。予測モデルでは，予測年次における説明変数の将来値が必要であるが，これを定めるためのフレーム設定や説明変数値の将来値のもつ予測誤差は被説明変数の予測に大きな誤差をもたらす。第二には，データ領域と予測時点で予測対象値の存在する領域（予測領域と呼ぶ）の食い違いによる予測誤差がある。モデルはデータ領域外においては検証されていない場合が多く，予測領域においてモデルの構造やパラメータの値が保持されているかは保証されない。また，モデルの有効性が予測領域においても仮定できる場合でも，予測領域がデータ領域から遠く離れる程大きくなるし，説明変数の相関関係の変化もまた予測誤差をもたらす。第三には，交通システムを取り巻く社会経済システム全体の構造変化であり，モデルに顕在的に取り込まれていない価値観等の要因の変化や予測時点においてその作用性が顕在化した要因の影響等が挙げられる。モデルに取り込まれなかった要因を含めて予測の空間を考えると，モデルに取り込まれていない要因の値域や相関関係の変化は予測領域の空間における位置を変化させている。さらに，予測モデルの説明変数は，説明変数と相関をもつところのモデルに取り込まれていない要因の影響をも代表

して表わしている。それゆえ，このような要因と説明変数の相関性の変化は，説明変数の作用性を変え予測誤差をもたらすと考えられる。したがって，予測領域の変化や相関性の変化に対するモデルの有効性の検証やデータ領域の拡大を意識モデル等を用いて行なうことは極めて重要と考えられる。

3．重回帰モデルにおけるデータ領域の予測誤差への影響

重回帰モデルは，発生交通量予測等の交通計画のみならず多くの分野で用いられているモデルであり，これを対象にデータの領域およびデータ数に関する検討を行う。対象とする現象が構造式

$Y(x)=\beta_0+\beta_1X_1+\cdots\cdots+\beta_pX_p+\varepsilon$ (1)　εは撹乱項で独立に$N(0,\sigma^2)$に従うものとする。

で記述されるものとする。説明変数$X_1,\cdots,X_p$，被説明変数Yに関するn組のデータ$(y_\nu, x_{\nu1},\cdots\cdots,x_{\nu p})$ $\nu=1,\cdots,n$から定められた(1)の回帰モデルを

$\hat{Y}(x)=\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_pX_p$ (2)　S^{kj}; 偏差平方和積和行列S_{ij}の逆行列のkj要素

$\hat{\beta}_k=\sum_{j=1}^{p}S^{kj}S_{jy}$ (3)　$S_{jy}=\sum_{\nu=1}^{n}(x_{\nu j}-\bar{x}_j)(x_{jy}-\bar{y})$，$S_{jy}=\sum_{\nu=1}^{n}(x_{\nu j}-\bar{x}_j)(y_\nu-\bar{y})$

とする。推定パラメータ$\hat{\beta}_k$はβ_kの不偏推定量であり，その分散$V(\hat{\beta}_k)$は，

$V\{\hat{\beta}_k\}=S^{kk}\sigma^2=\sigma^2/S_{kk}(1-R_k^2)$ (4)

となる。ここでR_kはX_kをそれ以外の説明変数$X_1,\cdots,X_{k-1},X_{k+1}\cdots X_p$で説明した場合の重相関係数であり，説明変数間に相関がある場合のモデルの不安定性が$(1-R_k^2)$に表されている。パラメータ$\hat{\beta}_k$の推定誤差はS_{kk}が大きいときに小さくなる。S_{kk}は明らかに$X_{\nu k}$の各々の値が平均$\bar{x}_k$から離れているとき大きくなるので，$\hat{\beta}_k$の安定性を増すためには，$x_{\nu k}$の分布する領域を大きくとることおよび$\bar{x}_k$から離れたデータを収集することが有効となる。従って，Yに対するX_kの影響の線形性が経験的に知られ残差の分析を要しないときには，$x_{\nu k}$のデータを上述の如く，計画的に配置して収集することが有効な方法となる。また，S_{kk}はデータ数nが増すと大きくなる。データが無作為に抽出されX_kが確率変数とみなせる場合には，X_kの分布が正規分布$N(\mu_k,\sigma_k^2)$で近似できるとすれば，S_{kk}の期待値が$n-1$となることから，$\hat{\beta}_k$の分散は$1/(n-1)$に比例して小さくなることが期待される。

説明変数の将来値$X_0=(x_{01},\cdots,x_{0p})$に対する構造式上の値$Y(X_0)$と予測値$\hat{Y}(X_0)$の差の分散は，

$V\{Y(X_0)-\hat{Y}(X_0)\}=\{1+1/n+D_0^2/(n-1)\}\sigma^2$，　$D_0^2=\sum_{k=1}^{p}\sum_{k'=1}^{p}(x_{0k}-\bar{x}_k)(x_{0k'}-\bar{x}_{k'})S^{kk'}/(n-1)$ (5)

となる。ここに，D_0^2はP次元空間におけるデータの重心とX_0の間のマハラノビスの距離であり，D_0^2とデータ数nが予測精度に関与している。すなわち，データの分布する領域と予測対象となるX_0の分布する領域との食い違いによる予測誤差は両空間の間のマハラノビスの距離を用いて表現される。D_0^2はデータの分布する領域の外に出ると急激に大きくなるし，また，説明変数間の相関関係を反映しており，X_0がこの相関関係を満たしていないときには著しく大きくなるためには，予測が外挿となるときにはD_0^2の増大に対応してデータ数を増加する必要がある。$\hat{Y}(X_0)\pm d$が確率$1-\alpha$をもって予測対象値$Y(X_0)$を覆うようにして予測精度の一定化を図るためには，

$1+1/n+D_0^2/(n-1)\leqq(d/t_{\frac{\alpha}{2},n-p-1})^2$ (6)　t;自由度$n-p-1$のt分布の$\frac{\alpha}{2}$点，　$\sigma^2=S_{yy}(1-R^2)/(n-p-1)$

なる関係を用いることが考えられる。必要データ数nを事前に定めるためには，モデルの安定性および予測精度の場合共，未知の分散σ^2の大きさを仮定する必要があるが，$S_{yy}(1-R^2)/\sigma^2$が自由度$n-p-1$のχ^2分布することを用いて，分析で期待する重相関係数の大きさをもとに想定する方法が考えられる。

4．おわりに

本研究では，データ領域すなわち限られた次元およびデータの値域の観測に基づき予測を行うことの予測誤差への影響について考察した。そして，予測誤差がデータ領域の拡がり，データ数および予測領域とのマハラノビスの距離との関連において理解されることが分かった。関係式を抽出するためのサンプリングの考え方，意識モデルの併用による予測性の向上等が今後の課題として残されている。

参考文献　山形・相馬"交通需要予測の予測誤差について"第1回土木計画学発表会，奥野・芳賀・久米・吉沢"多変量解析法"日科技連