SDG11.3.1
人口増加率と土地利用率の比率

横浜市立大学大学院データサイエンス研究科　教授
佐藤彰洋

2021年8月1日

資料2

検証作業対象指標（SDG11.3.1）

- 人口増加率　$PGR = \dfrac{\ln Pop_{t+n} - \ln Pop_n}{y}$
- 土地利用率　$LCR = \dfrac{\ln Urb_{t+n} - \ln Urb_n}{y}$
- 比　$LCRPGR = \dfrac{\ln Urb_{t+n} - \ln Urb_n}{\ln Pop_{t+n} - \ln Pop_n}$

・PGRとLCRの値の正負から４つの象限が存在する
・人口増加率と土地利用率の比では第1、第3象限と第2、第4象限が区別できない

メッシュ統計としての
４象限分類図の作成

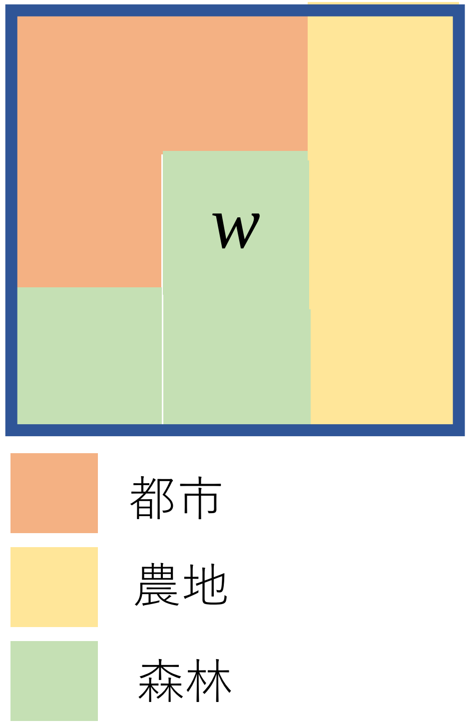
- 3次メッシュ統計を作成して、PGR,LCR,LCRPGRを計算し、４象限分類図の作成試みる。
- ３次メッシュ統計：国勢調査人口Pop_t(2010年)、Pop_{t+n}(2015年)
- ３次メッシュ統計として土地利用比率：JAXA ALOS土地利用土地被覆の分類（都市）の割合　Ubn_t(2006-2011年)、Ubn_{t+n}(2014-2016年)

世界メッシュコードを用いて試作

- 年tにおける都市割合３次メッシュ統計の作成・・・３次メッシュコードwで指示される年tにおける都市と分類される面積を$m_t(w)$とする。
- 3次メッシュwにおける土地利用比率は

$$LCR(w)=(\ln(m_{t+n}(w))-\ln(m_t(w)))/y$$

となる。



世界メッシュコード計算ライブラリ：
https://www.fttsus.org/worldgrids/ja/our_library/
（一般社団法人世界メッシュ研究所）

LCRの誤差評価

- 時点t+nとtにおけるメッシュw内の都市面積をm_{t+n}とm_tとする
- 土地利用比率(LCR)は以下で定義される

$$LCR = \frac{\ln(m_{t+n}) - \ln(m_t)}{y}$$

- ここで、メッシュwの面積をAとし、時点t+nとtにおけるメッシュ内母都市比率をp_tとすると$m_{t+n} = Ap_{t+n}$と$m_t = Ap_t$が成り立つ。
- よって

$$LCR = \frac{\ln(Ap_{t+n}) - \ln(Ap_t)}{y} = \frac{\ln(p_{t+n}) - \ln(p_t)}{y}$$

区間推定値

- 衛星観測画像により識別されたメッシュw内総ピクセルN_tと都市識別ピクセル数n_tにより計算される標本都市比率$\tilde{p}_t = n_t/N_t$により、都市母比率p_tを推計することを考える。
- このとき、十分大きなN_tに対して二項分布B(p_t,N_t)を正規分布N($N_t p_t$,$N_t p_t$(1-p_t))で近似することにより、都市母比率p_tの95%信頼区間は以下で与えられる。

$$\tilde{p}_t - 1.96\sqrt{\frac{\tilde{p}_t(1-\tilde{p}_t)}{N_t}} \leq p_t \leq \tilde{p}_t + 1.96\sqrt{\frac{\tilde{p}_t(1-\tilde{p}_t)}{N_t}}$$

誤差伝播公式

- z=f(x)の関係があって、$x = x_0 \pm \varepsilon_x$のとき、 $z_0 = f(x_0)$として、zの誤差伝播は$z = z_0 \pm \left|\frac{df}{dx}\right|_{x=x_0} \varepsilon_x$となる。
- $f(x) = \ln(x)$とすると、$\frac{d}{dx}\ln(x) = \frac{1}{x}$なので、 $\ln(x_0 \pm \varepsilon_x) = \ln(x_0) \pm \frac{\varepsilon_x}{x_0}$である。
- よって、標本比率を使って計算されるLCRの95%信頼区間は、p_{t+n}とp_tの95%信頼区間幅の半分をそれぞれ$\delta\alpha$と$\delta\beta$とおくと以下で与えられる。

$$\widetilde{LCR} = \frac{\ln(p_{t+n} \pm \delta\alpha) - \ln(p_t \pm \delta\beta)}{y} = \frac{\ln(p_{t+n}) - \ln(p_t)}{y} \pm \frac{1}{y}\left(\frac{\delta\alpha}{p_{t+n}} + \frac{\delta\beta}{p_t}\right)$$

LCRの95%信頼区間

$$\widetilde{LCR} = LRC \pm \frac{1}{y}\left(\frac{\delta\alpha}{p_{t+n}} + \frac{\delta\beta}{p_t}\right)$$

であり、$\delta\alpha = 1.96\sqrt{\frac{p_{t+n}(1-p_{t+n})}{N_{t+n}}}$、 $\delta\beta = 1.96\sqrt{\frac{p_t(1-p_t)}{N_t}}$であるので、

$$\widetilde{LCR} - \frac{1.96}{y}\left(\sqrt{\frac{1-\widetilde{p_{t+n}}}{n_{t+n}}} + \sqrt{\frac{1-\widetilde{p}_t}{n_t}}\right) \le LRC$$

$$\le \widetilde{LCR} + \frac{1.96}{y}\left(\sqrt{\frac{1-\widetilde{p_{t+n}}}{n_{t+n}}} + \sqrt{\frac{1-\widetilde{p}_t}{n_t}}\right)$$

となる。■

分類誤りへの対処方法

- SDG15.4.2で用いた分類誤りの混合行列（コンフュージョンマトリックス）による補正方法を用いる
- 都市(1)/非都市(0)の２値分類に対する２×２混合行列から条件付き確率$R(a|b)$を算出することで、都市/非都市２値分類問題における都市被覆比率の分類誤り補正を行う
- ３次メッシュごとに都市/非都市の分類誤りを補正した土地利用比率の９５％信頼区間を算出する
- ３次メッシュレベルでのLCRの評価行うことを初期的には目指す

$R(a|b)$の定義

$$R(0|0)=\frac{TN}{FP+TN}, R(1|0)=\frac{FP}{FP+TN}$$
$$R(0|1)=\frac{FN}{TP+FN}, R(1|1)=\frac{TP}{TP+FN}$$

$R(a|b)$のコンフュージョンマトリックスからの推計は標本比率なので95%信頼区間がある

- 真のメッシュw_iの都市母比率をp_iとし、$R(a|b)$を実際(validated)はb={都市,非都市}であるが、データ分類(classified)上はa={都市,非都市}である条件付き確率として定義する。

		validated b	
	a / b	都市 (X=1)	非都市 (X=0)
classified a	都市 (X=1)	TP	FP
	非都市 (X=0)	FN	TN

条件付き確率： $$R(a|b)=\frac{p_{AB}(a,b)}{p_B(b)}$$

$$p_{AB}(a,b)=R(a|b)p_B(b)$$

$$p_A(a)=\sum_b p_{AB}(a,b)=\sum_b R(a|b)p_B(b)$$

$$p_A(1)=\sum_{b=0,1} R(1|b)p_B(b)=R(1|1)p_B(1)+R(1|0)p_B(0)$$

$$p_A(0)=\sum_{b=0,1} R(0|b)p_B(b)=R(0|0)p_B(0)+R(0|1)p_B(1)$$

ここで、$p_B(1)=p_i, p_B(0)=1-p_i$である。

都市比率の誤差評価

$$Z_i = \sum_{t=1}^{N_i} X_i(t)$$

とすると、Z_iは二項分布$B\left(N_i, R(1|1)p_i + R(1|0)(1-p_i)\right)$に従う。
更に、N_iが十分大きいと、Z_iは
平均: $N_i\left(R(1|1)p_i + R(1|0)\ (1-p_i)\right)$、
分散:$N_i\left(R(1|1)p_i + R(1|0)(1-p_i)\right)(R(0|0)(1-p_i) + R(0|1)p_i)$
の正規分布により近似できる。よって、都市比率推定値

$$\widehat{p_i} = \frac{\sum_{t=1}^{N_i} X_i(t)}{N_i}$$

とすると、$\widehat{p_i}$は
平均:$R(1|1)p_i + R(1|0)\ (1-p_i)$、
分散:$\left(R(1|1)p_i + R(1|0)\ (1-p_i)\right)(R(0|0)(1-p_i) + R(0|1)p_i)/Ni$
の正規分布で近似できる。

分類誤りが指数推定値に及ぼす影響

- 指数推定値の平均値

$$\mathrm{E}[\widehat{p_i}] = R(1|1)p_i + R(1|0)(1 - p_i)$$

・都市/非都市に対する判定誤りがあると指数推定値は真の母指数p_iよりも減少または増加する
・都市/非都市の分類判定率が判明している場合は、推定される指数推定値を次の式で補正するのが良いかもしれない

$$p_i = \frac{\mathrm{E}[\widehat{p_i}] - R(1|0)}{R(1|1) - R(1|0)}$$

分類精度の検証

- $R(a|b)$を実際(validated)はb={都市,非都市}であるが、データ分類(classified)上はa={都市,非都市}である条件付き確率として定義する。

		validated b	
	a / b	都市 (X=1)	非都市 (X=0)
classified a	都市 (X=1)	TP	FP
	非都市 (X=0)	FN	TN

- $R(a|b)$のコンフュージョンマトリックスからの推計は標本比率なので95%信頼区間がある

$$R(0|0) = \frac{TN}{FP+TN}, R(1|0) = \frac{FP}{FP+TN}$$
$$R(0|1) = \frac{FN}{TP+FN}, R(1|1) = \frac{TP}{TP+FN}$$

$$\sigma^2 = \hat{q}(1-\hat{q})/N$$

95%信頼区間：$\hat{q} \pm 1.96\sigma$
99%信頼区間：$\hat{q} \pm 2.58\sigma$

（$\hat{q}$を標本比率、Nを標本サイズとする）

$R(a|b)$の95%信頼区間

コンフュージョンマトリックス		**validated** b	
	a / b	都市 (X=1)	非都市 (X=0)
classified a	都市 (X=1)	TP	FP
	非都市 (X=0)	FN	TN

$$\frac{TN}{FP+TN}-1.96\sqrt{\frac{TN}{FP+TN}\left(1-\frac{TN}{FP+TN}\right)/(FP+TN)}\le R(0|0)\le\frac{TN}{FP+TN}+1.96\sqrt{\frac{TN}{FP+TN}\left(1-\frac{TN}{FP+TN}\right)/(FP+TN)}$$

$$\boxed{\frac{FP}{FP+TN}-1.96\sqrt{\frac{FP}{FP+TN}\left(1-\frac{FP}{FP+TN}\right)/(FP+TN)}\le R(1|0)\le\frac{FP}{FP+TN}+1.96\sqrt{\frac{FP}{FP+TN}\left(1-\frac{FP}{FP+TN}\right)/(FP+TN)}}$$

$$\frac{FN}{TP+FN}-1.96\sqrt{\frac{FN}{TP+FN}\left(1-\frac{FN}{TP+FN}\right)/(TP+FN)}\le R(0|1)\le\frac{FN}{TP+FN}+1.96\sqrt{\frac{FN}{TP+FN}\left(1-\frac{FN}{TP+FN}\right)/(TP+FN)}$$

$$\boxed{\frac{TP}{TP+FN}-1.96\sqrt{\frac{TP}{TP+FN}\left(1-\frac{TP}{TP+FN}\right)/(TP+FN)}\le R(1|1)\le\frac{TP}{TP+FN}+1.96\sqrt{\frac{TP}{TP+FN}\left(1-\frac{TP}{TP+FN}\right)/(TP+FN)}}$$

四則演算（＋－×÷）に対する誤差伝播公式

- 加算誤差の伝播

$$(a \pm \delta a) + (b \pm \delta b) = (a + b) \pm (\delta a + \delta b)$$

- 減算誤差の伝播

$$(a \pm \delta a) - (b \pm \delta b) = (a - b) \pm (\delta a + \delta b)$$

- 乗算誤差の伝播

$$(a \pm \delta a)(b \pm \delta b) = ab\left(1 \pm \left(\frac{\delta a}{a} + \frac{\delta b}{b}\right)\right)$$

- 除算誤差の伝播

$$\frac{a \pm \delta a}{b \pm \delta b} = \frac{a}{b}\left(1 \pm \left(\frac{\delta a}{a} + \frac{\delta b}{b}\right)\right)$$

母比率の補正の誤差伝播

$$p_i = \frac{\mathrm{E}[\hat{g}_i] - R(1|0)}{R(1|1) - R(1|0)}$$

の誤差伝播について考える。

四則演算に対する誤差の伝播公式より

$$a = \mathrm{E}[\hat{p}_i], b = R(1|0), c = R(1|1)$$

$\delta a, \delta b, \delta c$をそれぞれ$\mathrm{E}[\hat{p}_i], R(1|0), R(1|1)$ の95%信頼区間幅の半分とすると、母比率p_iの信頼区間は以下で与えられる。

$$\frac{a-b}{c-b}\left(1 - \left(\frac{\delta a+\delta b}{a-b} + \frac{\delta c+\delta b}{c-b}\right)\right) \le p_i \le \frac{a-b}{c-b}\left(1 + \left(\frac{\delta a+\delta b}{a-b} + \frac{\delta c+\delta b}{c-b}\right)\right)$$

LCRの95%信頼区間

$$\widetilde{LCR} = LRC \pm \frac{1}{y}\left(\frac{\delta\alpha}{p_{t+n}} + \frac{\delta\beta}{p_t}\right)$$

であり、

$$\widetilde{LCR} = \frac{\ln\left(\frac{a_{t+n} - b_{t+n}}{c_{t+n} - b_{t+n}}\right) - \ln\left(\frac{a_t - b_t}{c_t - b_t}\right)}{y}$$

として、

$\delta\alpha = \frac{a_{t+n}-b_{t+n}}{c_{t+n}-b_{t+n}}\left(\frac{\delta a_{t+n}+\delta b_{t+n}}{a_{t+n}-b_{t+n}} + \frac{\delta c_{t+n}+\delta b_{t+n}}{c_{t+n}-b_{t+n}}\right)$、 $\delta\beta = \frac{a_t-b_t}{c_t-b_t}\left(\frac{\delta a_t+\delta b_t}{a_t-b_t} + \frac{\delta c_t+\delta b_t}{c_t-b_t}\right)$、$p_{t+n} = \frac{a_{t+n}-b_{t+n}}{c_{t+n}-b_{t+n}}$、 $p_t = \frac{a_t-b_t}{c_t-b_t}$

であるので、

$$\widetilde{LCR} - \frac{1}{y}\left(\frac{\delta a_{t+n} + \delta b_{t+n}}{a_{t+n} - b_{t+n}} + \frac{\delta c_{t+n} + \delta b_{t+n}}{c_{t+n} - b_{t+n}} + \frac{\delta a_t + \delta b_t}{a_t - b_t} + \frac{\delta c_t + \delta b_t}{c_t - b_t}\right) \leq LRC$$
$$\leq \widetilde{LCR} + \frac{1}{y}\left(\frac{\delta a_{t+n} + \delta b_{t+n}}{a_{t+n} - b_{t+n}} + \frac{\delta c_{t+n} + \delta b_{t+n}}{c_{t+n} - b_{t+n}} + \frac{\delta a_t + \delta b_t}{a_t - b_t} + \frac{\delta c_t + \delta b_t}{c_t - b_t}\right)$$

となる。■

作業の流れ（案）

3次メッシュレベルでいくつかの抜き取り、土地利用比率の誤差評価方式の開発

→

3次メッシュレベルでPGRとLCRの誤差評価を人口集中７都市で拡大実施

→

４象限分類図３次メッシュレベルで人口集中７大都市で検証

→

日本全体でのPGRとLCRを４分類で集計して調表作業を行う

3次メッシュレベルでのLCR誤差評価方式（９５％信頼区間）

→

３次メッシュレベルでの分類誤り補正の検証＋土地利用土地分類図の突合分析（混合行列の精度向上）

→

３次メッシュレベルでの分類誤りを考慮したLCRの算出