Modélisation des dynamiques foncières à partir bases de données géohistoriques de grandes tailles, l'exemple d'UrbanSIMUL.

Ghislain Geniaux

Modélisation des dynamiques foncières à partir bases de données géohistoriques de grandes tailles, l'exemple d'UrbanSIMUL.

Ghislain Geniaux,

UR Ecodéveloppement 767, INRA Avignon

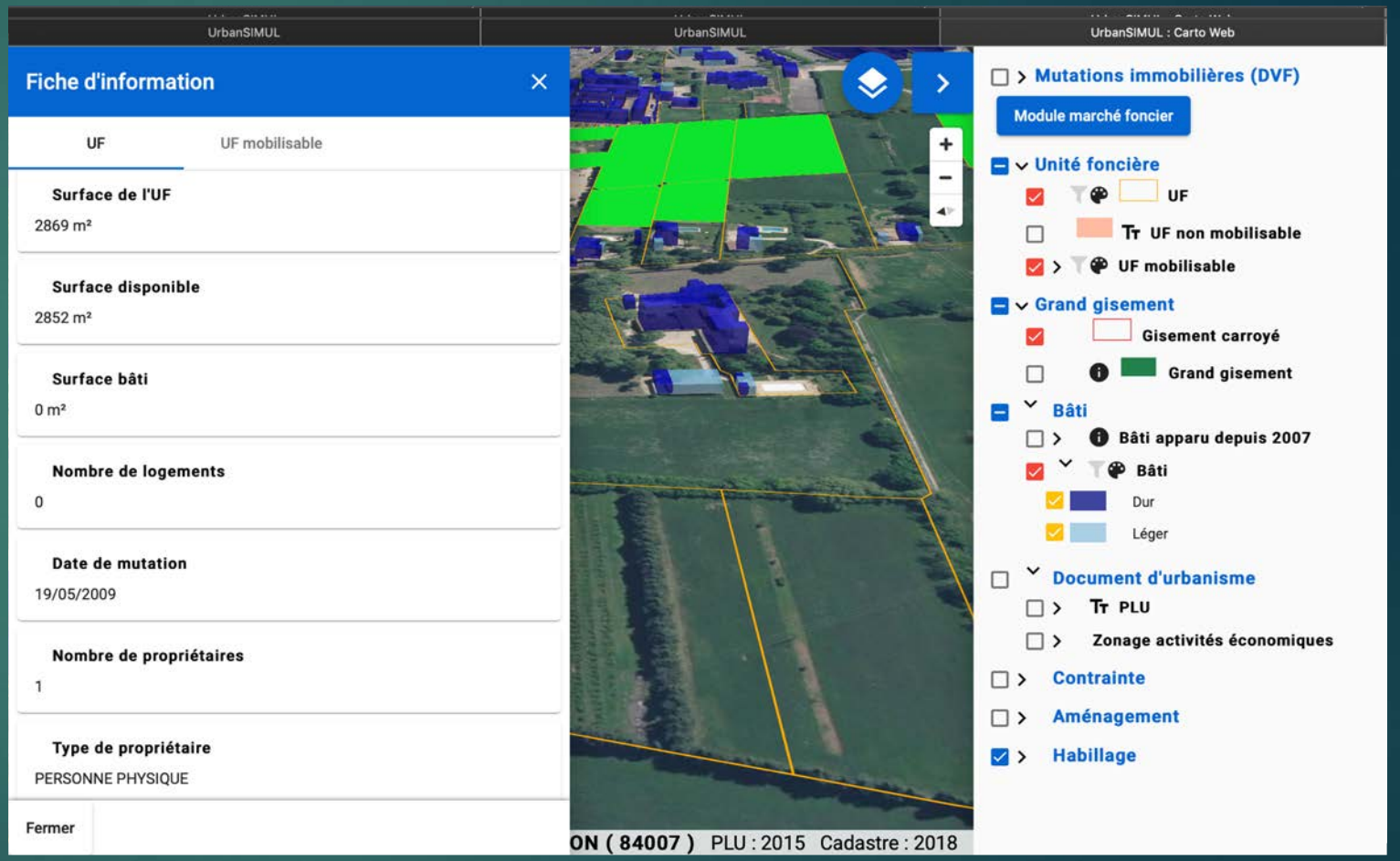
PLAN

1. Panorama du projet UrbanSIMUL
    1. Video
    2. Données
    3. Modules
    4. Enjeux
2. Modélisation et BIG DATA
3. Les modèles du module PRONOSTIC

VIDEO :
https://www.urbansimul.fr/support/presentation

Panorama du projet UrbanSIMUL

- Un SaaS (software as a service) en ligne, avec plus de 500 utilisateurs.
- il inclut des données sur 6 millions de parcelles suivies annuellement et couvre 2 régions avec 1500 municipalités.
- Il s'agit d'un projet gagnant-gagnant entre les chercheurs et les planificateurs/gouvernements locaux.

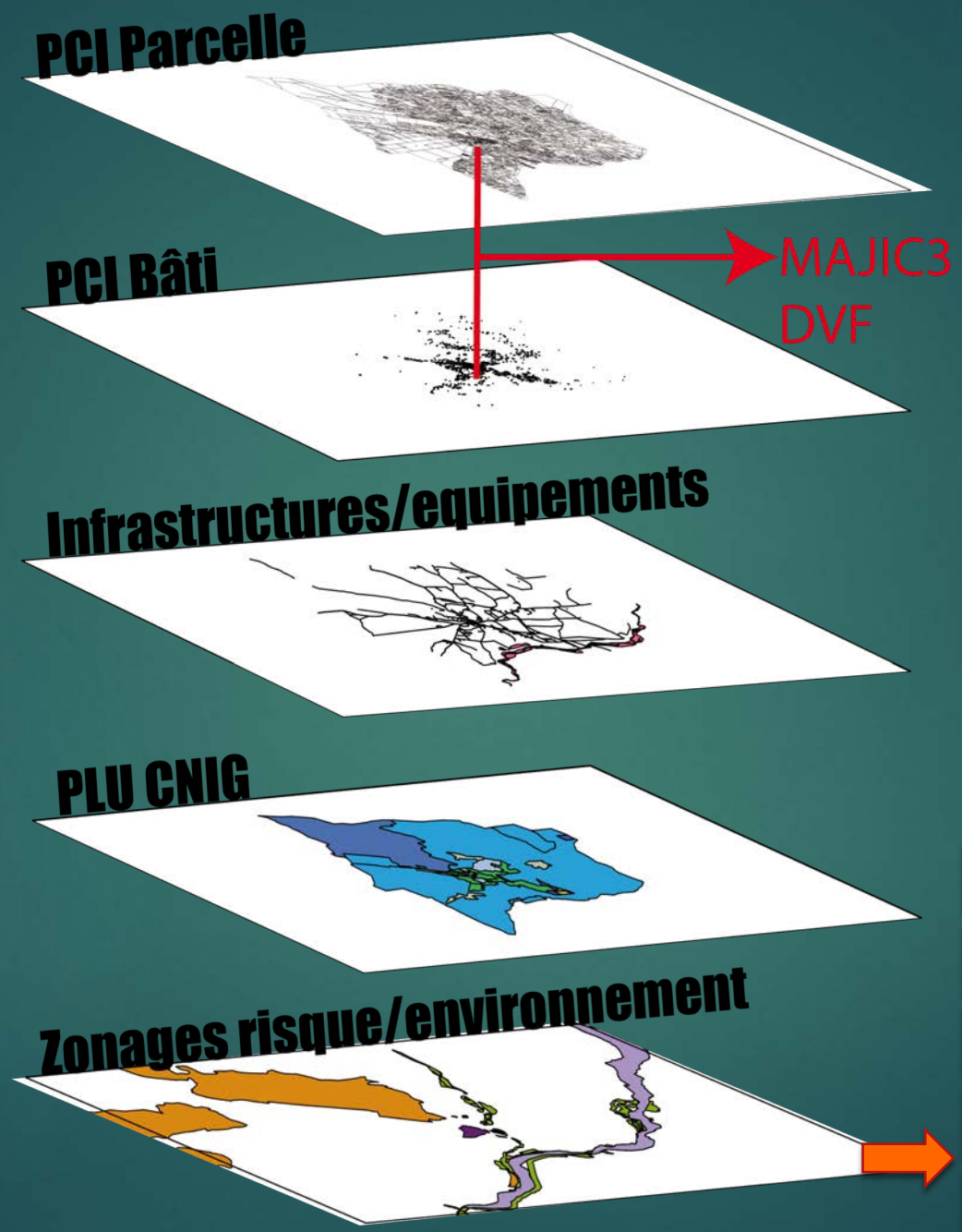


DONNEES

Base de données

GéoHistorique

2007-2017



1 • Très forte contrainte

2 • Forte contrainte

3 • Contrainte moyenne

4 • Faible contrainte

0 • Aucune contrainte

DONNEES REMARQUABLES

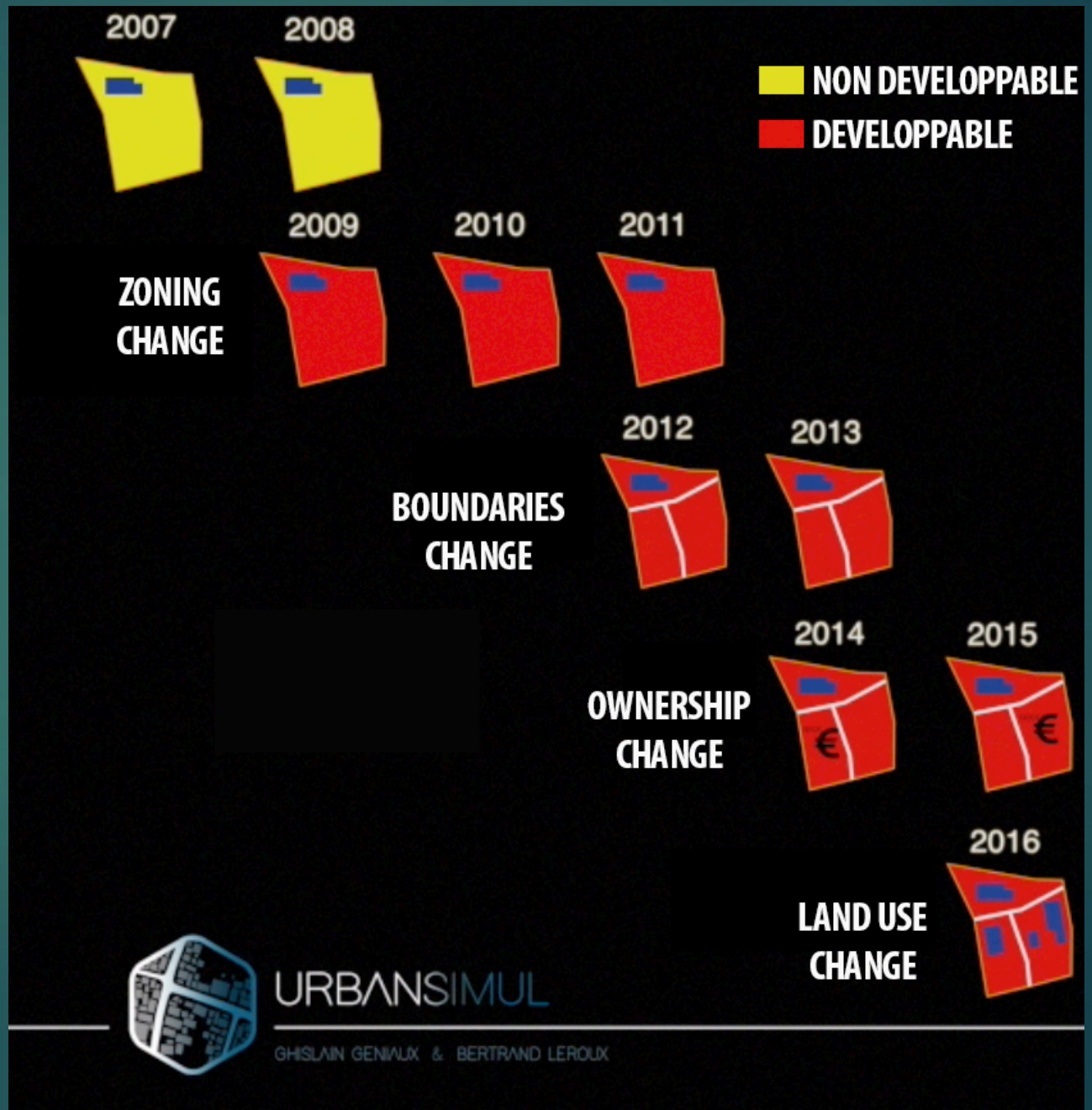
- Données exhaustives sur :
  - Les parcelles et les bâtis
  - Les prix,

et surtout sur :

  - Les propriétaires
  - La règlementation en matière d'urbanisme

→ Offre foncière communale et portefeuille foncier/immobilier

DES DECISIONS DANS L'ESPACE ET LE TEMPS



Enjeux empiriques

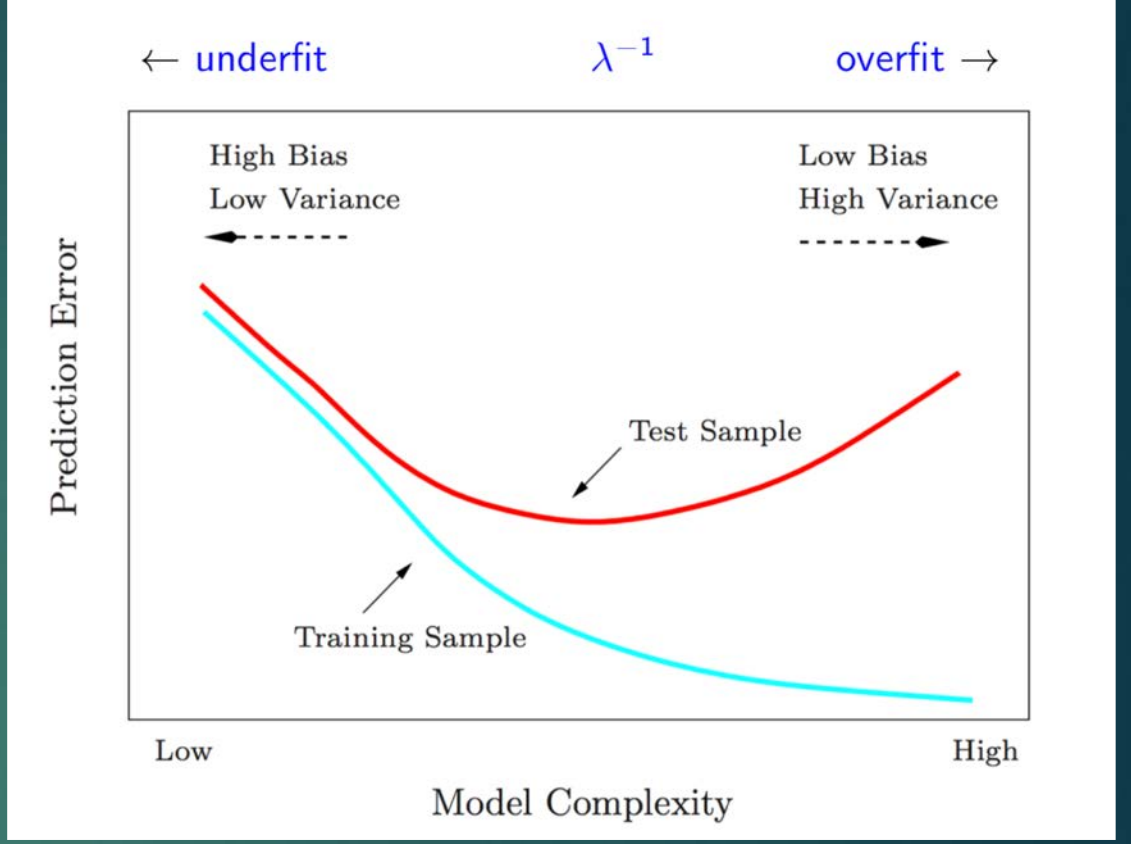
- L'évaluation des politiques publiques, notamment foncière
  - Nécessaire mais complexe
  - Bonne connaissance des déterminants des prix et des usages du sol à une échelle meso (Irwin et al 2009,Esco Artificialisation des sols INRA-Ifstar 2017)
  - besoin d'analyse et de données à une échelle plus fine spatialement (Irwin 2010)
  - Données de grande dimension, avec de moins en moins d'a priori sur leurs rôles pour expliquer les phénomènes économiques

Modélisation et BIG DATA

- Les conséquences notables sur la pratique de l'économétrie des *Big Data*
  - Le mode de recueil et de traitement de l'information
  - Méthodes statistiques à privilégier
  - Conception/validation des modèles

H.Varian (2014) Big data ; New tricks for economists, *Journal of Economic Perspectives—Volume 28, Number 2-Pages 3–28*

Modélisation et BIG DATA



- Recours systématique

à de la *cross-validation*

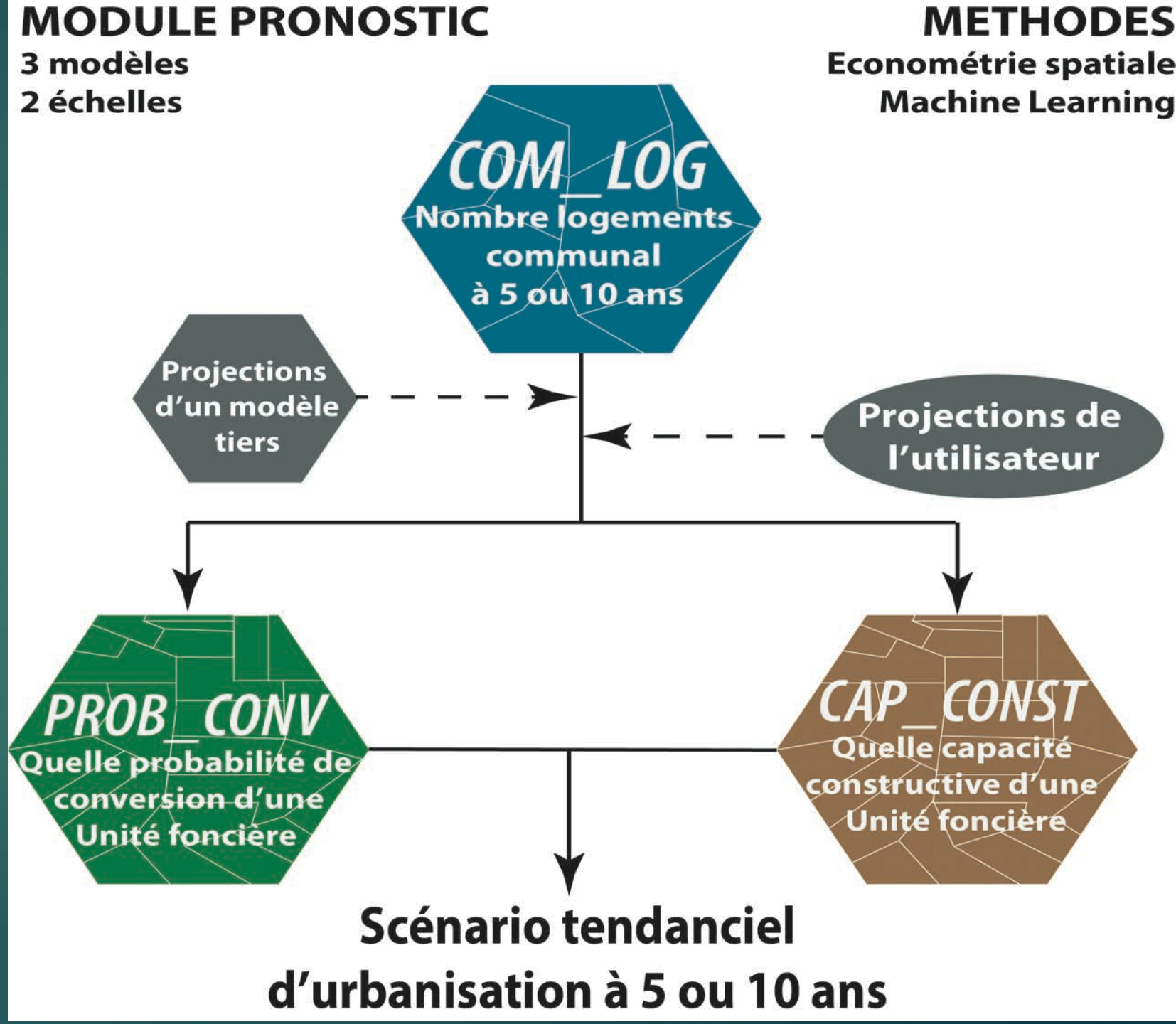
  - *In sample / out-sample* et *overfitting*
  - Méthode de régularisation (*Lasso, Elastic-Net, Penalized Regression, MARS*)
  - *Bootstrap , bagging, Boosting* et *ANN*

Développements de méthodes statistiques sur données de hautes dimension spatiales

- Modèle de choix discret avec autocorrélation spatiale
  - Package R ***ProbitSpatial*** 2016 (7135 téléchargements en 30 mois), Martinetti and Geniaux 2017 RSUE
- Hétérogénéité spatiale + autocorrélation spatiale
  - Package R ***mgwrsar*** 2017 (697 téléchargements en 5 mois), Geniaux and Martinetti 2017 RSUE
  - + non linéarité Package R ***mgwrsar*** 2018
- En cours de développement :
  - Fast-version of *mgwrsar*, CRAN 12/2019
  - algorithmes de *boosting* pour données spatiales/spatio-temporelles 2020

URBANSIMUL

Simulation de l'urbanisation



Les variables d'intérêt

CAP_CONST Les capacités constructives des Unités Foncières :

- nblog(t+8,UF) sachant nblog>0 (multinomial, count data)
- surflog(t+8,UF) sachant nblog>0

PROB_CONV Les probabilités de conversion des Unités Foncières :

- Proba(Δsurflog(t+8,UF)>0)

PROB_CONV Les probabilités de construction des Unités Foncières.

- Données 2009-2017 : 3 millions d'Ufs, 210 000 Ufs constructibles non bâties
- Construction de l'endogène : projection des nouveaux bâtis 2017 sur les UF de 2009
- X 330 variables initiales sur les Ufs, leur propriétaire, leur environnement, les règles d'urbanisme, les distances minimales à …
- 7000 variables de lissage spatio-temporel construites à partir de ces 330 variables initiales
  - K premiers voisins : 2,5,10,20,50
  - Noyau bisquare adaptatif
  - sans distinction, et/ou bâti ou pas et/ou de même type de zonage et/ou de taille comparable, en considérant l'évolution du bâti sur 3, 5, 7 ou 9 ans.

```
Library(mgwrsar)
W=KNN(coords,k=10,kernel='gauss_adapt')
WX=W %*% X
```

Méthodes

1. Estimation d'un modèle binomial
2. Utilisation de méthode de *gradient descent boosting* (*xgboost*)
3. Modèle *Probit Spatial* (*ProbitSpatial*)
4. Méthode de Model averaging

Cross validation : binomial prob_conv

Prediction	Reference 0	Reference 1
0	197222	6656
1	2260	7445

Un tiers en permis de construire

Overall Statistics

Accuracy : 0.9583

95% CI : (0.9574, 0.9591)

No Information Rate : 0.934

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6042

Statistics by Class:

Sensitivity : 0.9887

Specificity : 0.5280

Pos Pred Value : 0.9674

Neg Pred Value : 0.7671

Prevalence : 0.9340

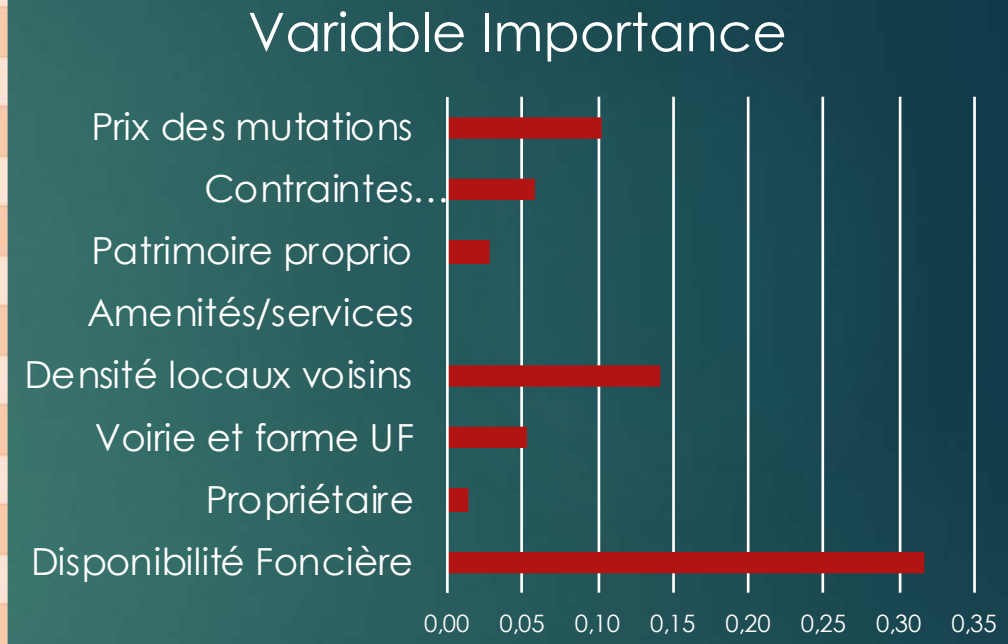
Detection Rate : 0.9234

Detection Prevalence : 0.9546

Balanced Accuracy : 0.7583

Les principaux *drivers* des probabilités de construction

Variables	Type de variables	Variable Importance
d_bati_pci	densité locaux voisins	0,14080
surfdispo	disponibilité foncière	0,08432
cgrnumdtxtTERRAINS_A_BATIR	mutation	0,08106
surfdispo2	disponibilité foncière	0,08021
amutp	mutation	0,06921
surf	disponibilité foncière	0,05205
ze_min1	contrainte réglementaire	0,04381
frrue_rel	Voirie et forme UF	0,02706
prop_spar	patrimoine proprio	0,02101
is	Voirie et forme UF	0,01557
ncont_niv1	contrainte réglementaire	0,01371
mediane	mutation	0,01171
e_m2_terr	mutation	0,01096
ispomobilisable_terrain_nu	disponibilité foncière	0,01001
age	propriétaire	0,00925
prop_sloc	patrimoine proprio	0,00714
WK4_e_m2_terr	mutation	0,00677
long	Voirie et forme UF	0,00592
libniv1PERSONNE_PHYSIQUE	propriétaire	0,00530
surf_plane	disponibilité foncière	0,00480
long2	Voirie et forme UF	0,00468
ccogrmNA	propriétaire	0,00442
nb_pro	propriétaire	0,00439
WK10_r9_ces	disponibilité foncière	0,00414
WK4_amutp	mutation	0,00367



CAP_CONST Les capacités constructives des Unités Foncières.

- Données 2009-2017 : 3 millions d'Ufs, 40 000 Ufs constructibles non bâties en 2009 devenues construites en 2017
- Même lot de variables que pour le modèle de probabilité de conversion.

Cross validation : multinomial 3 classes

Prediction	Reference 1	2-6	>6
1	23228	3720	194
2-6	3644	5122	764
>6	277	542	1810

Overall Statistics

Accuracy : 0.7674

95% CI : (0.7632, 0.7716)

No Information Rate : 0.6908

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4947

Statistics by Class:

	Class1	Class2	Class3
Sensitivity	0.8556	0.5458	0.65390
Specificity	0.6779	0.8527	0.97758
Pos Pred Value	0.8558	0.5375	0.68847
Neg Pred Value	0.6775	0.8568	0.97388
Prevalence	0.6908	0.2388	0.07043
Detection Rate	0.5910	0.1303	0.04605
Detection Prevalence	0.6906	0.2425	0.06689
Balanced Accuracy	0.7667	0.6992	0.81574

Cross validation : comptage

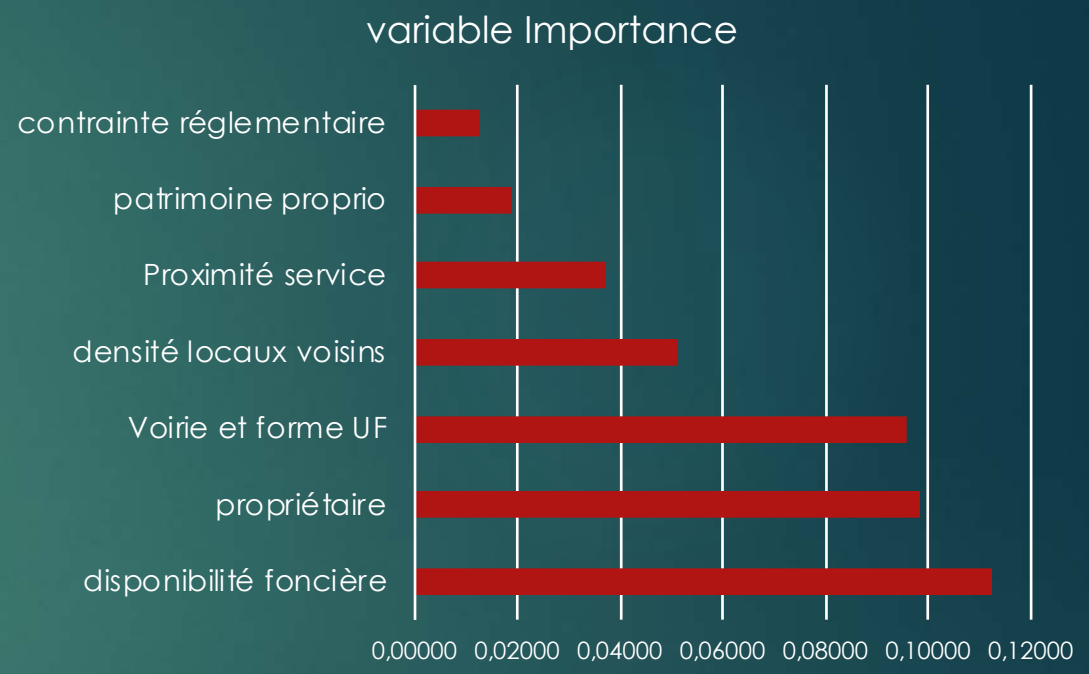
Le modèle de comptage produit des erreurs moyennes à l'échelle de l'unité foncière qui ne sont pas encore pleinement satisfaisante avec une erreur moyenne de la prédiction out-sample à 40 % à l'échelle de l'UF. En revanche, on peut voir que lorsqu'on cumule ces prédictions à des échelles supérieures les prédictions deviennent plus satisfaisantes:

Distribution des erreurs cumulées

Échelle	Commune x zone_plu			Commune :		
Stat	median	mean	p75	median	mean	p75
Err %	0.1852	0.2321	0.3104	0.1005	0.1776	0.2229

Les principaux *drivers* de la capacité constructive

Variables	Type de variables	Variable Importance
frrue_abs	**Voirie et forme UF**	0,07011
ccogrmNA	propriétaire	0,05679
surf_plane	**disponibilité foncière**	0,04771
surfdispo2	**disponibilité foncière**	0,03377
d_bati_pci	densité locaux voisins	0,03255
libniv1PERSONNE_PHYSIQUE	propriétaire	0,02858
frrue_rel	**Voirie et forme UF**	0,01866
surf	disponibilité foncière	0,01820
d_pharmacie	Proximité service	0,01565
WK50_nlocappt	**densité locaux voisins**	0,01352
ncont_niv1	**contrainte réglementaire**	0,01272
surfdispo	**disponibilité foncière**	0,01270
prop_spar	patrimoine propriétaire	0,01231
WK50_nblocal	densité locaux voisins	0,01202
d_ecole_ma	Proximité service	0,01186
amutp	Date de mutation	0,01104
d_ecole_el	**Proximité service**	0,00975
dlogtuf_com.mediancom	densité locaux voisins	0,00726
WK4_frrue_rel	**Voirie et forme UF**	0,00684
age	propriétaire	0,00655
prop_sloc	patrimoine propriétaire	0,00646
dlogtuf_complu.median	densité locaux voisins	0,00624
WBR9S0.85_dlogtuf	**densité locaux voisins**	0,00616
nb_pro	propriétaire	0,00613
WBR5S0.85_dlogtuf	**densité locaux voisins**	0,00598



LOG_COM La démographie et le parc de logement

Le nombre de logements, les prix et la population à la commune (~ Jeanty Partridge & Irwin 2010):

$$\begin{cases} \Delta \text{pop}_c = f\ (\text{X}_c, \Delta \text{logt}_c\ , \Delta \text{Indiceprix}_c\ , \Delta \text{OffreFoncière}_c) \\ \Delta \text{logt}_c = g\ (\text{X}_c, \Delta \text{pop}_c\ , \Delta \text{Indiceprix}_c\ , \Delta \text{OffreFoncière}_c) \\ \Delta \text{IndicePrix}_c = h\ (\text{X}_c, \Delta \text{logt}_c\ , \Delta \text{pop}_c\ , \Delta \text{OffreFoncière}_c) \\ \Delta \text{OffreFoncière}_c = k\ (\text{X}_c, \Delta \text{logt}_c\ , \Delta \text{pop}_c\ , \Delta \text{IndicePrix}_c\ ) \end{cases}$$

Avec f(), g(), h(), k() calibrés à partir de méthodes de boosting.

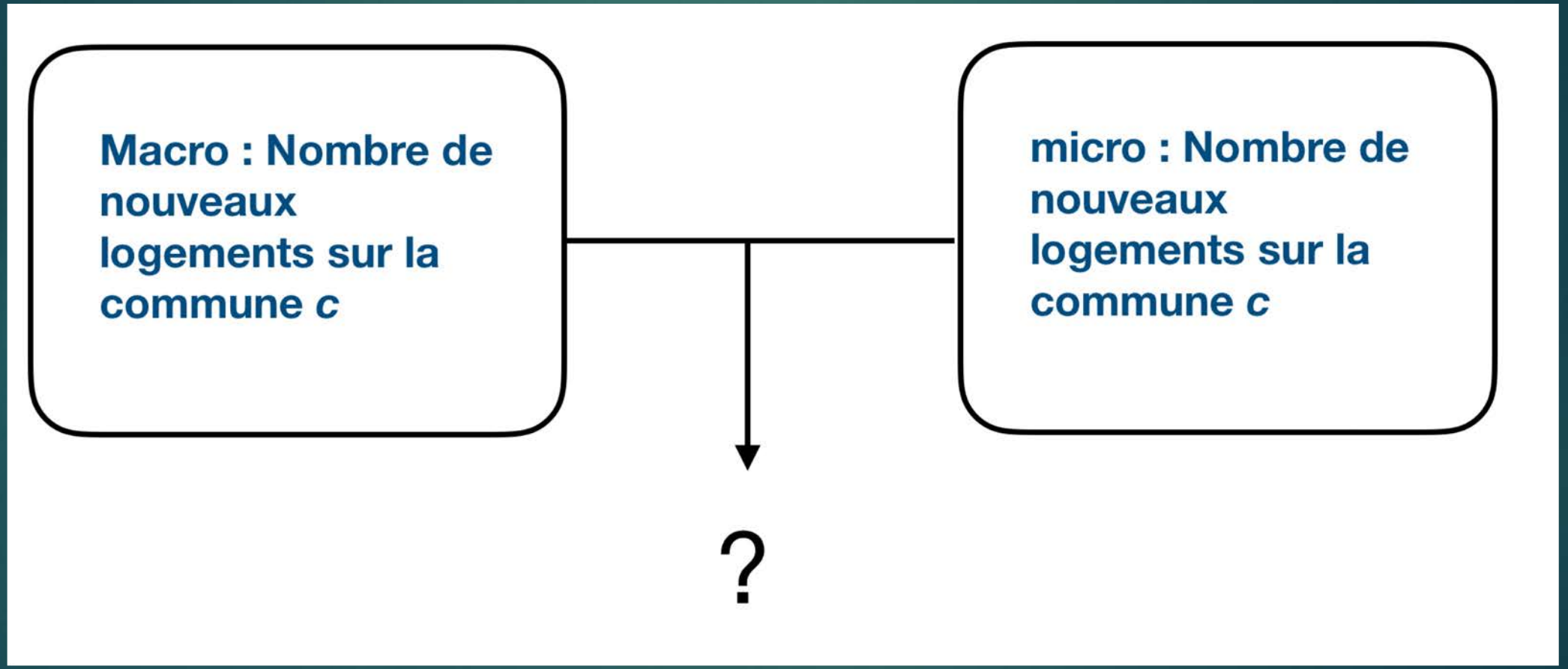
MODELE COM_LOG

$$\begin{cases} \Delta^{M} pop_{c}^{t+h} = Fpop(X_{c,t}, h) \\ \Delta^{M} log_{c}^{t+h} = Flog(X_{c,t}, \Delta \hat{pop}_{c}^{t+h}, h) \end{cases}$$

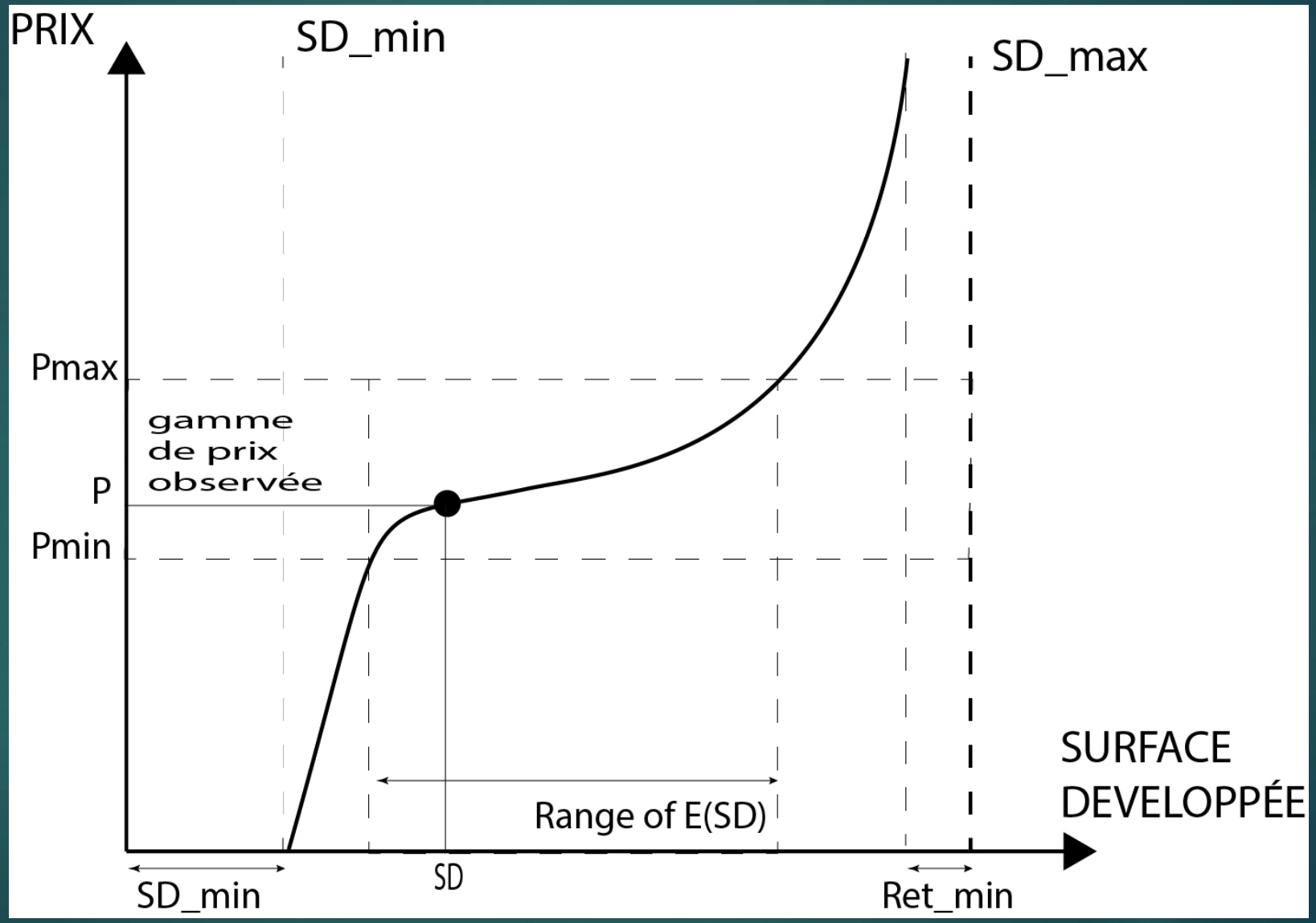
$$\Delta^{M} \hat{log}_{c}^{t+h}$$

Macro : Nombre de nouveaux logements sur la commune *c*

DISCUSION


Macro : Nombre de nouveaux logements sur la commune c
micro : Nombre de nouveaux logements sur la commune c
?

DISCUSION


PRIX
SD_min
SD_max
Pmax
gamme
de prix
observée
P
Pmin
SURFACE
DEVELOPPÉE
Range of E(SD)
SD_min
SD
Ret_min

DISCUSION

$\exists\, \forall c, \bar{P}_c$ tel que

$$\sum_c \Delta^M \hat{log}_c^{t+h} = \sum_c \Delta^m \hat{log}_c^{t+h}(\bar{P}_c)$$

MODELE Ajustement micro/macro

$$\tilde{\bar{P}}_c = \underset{\tilde{\bar{P}}_c}{arg\,min} \sum_c (\bar{P}_c - Cost(\tilde{\bar{P}}_c))$$

tel que $\forall c$,

$$\Delta^M \hat{log}_c^{t+h} = \Delta^m \hat{log}_c^{t+h}(\tilde{\bar{P}}_c)$$

Avec $Cost(P_c) = P_c * OC$

Et OC = Ordre de Contiguité

Je vous remercie de votre attention

Ces travaux ont bénéficié de fonds issus d'un partenariat entre les institutions suivantes :