European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

Improving building permit administrative data set for short term analysis:

François Limousin[1], Benoit Pentinat[2], Frédéric Minodier[3]

[1] *SOeS, Paris, France; francois.limousin@developpement-durable.gouv.fr*
[2] *SOeS, Paris, France; benoit.pentinat@developpement-durable.gouv.fr*
[3] *SOeS, Paris, France; frederic.minodier@developpement-durable.gouv.fr*

Abstract
Administrative data gathering is dependent on organisational and regulatory changes. In France, the monthly number of authorized dwellings used to be drawn from counting the permits received during the reference period. Following several difficulties, the figure is now estimated to be closer to the actual reality.

Keywords: administrative data, buiding permis, short term analysis.

Dealing with administrative data for short term analysis may be very confusing. On the one hand, plenty of information is available but on the other hand, there may be severe lacks as the gathering process can be a little bit too long to be completed in time. This is the case for building permits indicator in France, where a monthly transmission is defined but local authorities don't always play the game (for various reasons). Our aim was then to reconciliate administrative data and short term analysis.

1. The source sitadel2.

Sitadel2 (Système d'information et de traitement automatisé des données élémentaires sur les logements et les locaux – information system and automated processing of basic data on homes and premises) is a useful data

collection tool and a dissemination base enabling the Observation and Statistics Department (SOeS) to monitor housing dwellings and premises building from planning permissions.

Monitoring new building from planning permissions is not new, as the first public statistics system (Sirocco) dates back to 1972. Data collection systems have since developed due to the decentralisation of the French public services on the one hand and developments in information technology on the other. Siclone (1986), Sitadel (1998) and Sitadel2 (2009) have in turn replaced the Sirocco system.

1.1 Data collection

The planning permission collection system relies on various actors, from the applicant, who files the planning application at the town hall for the place where the building work is to be carried out, to the national Sitadel2 database, managed by the Ministry of Environment, Energy and Sea (MEEM).

The building project is monitored throughout its life cycle. The first step is the authorization issued by the competent authority on completion of the work carried out by the local planning authority (LPA). This data is returned relatively quickly since about 65% of the data is returned during the first month and 95% at the end of 6 months. This first step is the object of this document, even if the project main goal is to solve a data collection issue on the started dwellings.

At the beginning of the month, LPAs use the Sitadel2 collection tool to send all the events (filing, decision, implementation, work completion note) received and dealt with in the preceding month. Around the 20th of each month, the collected authorizations are dumped into the dissemination database.

Two dates are allocated to each event dumped into the infocentre: its real date (DR) and its date of collection (DPC) in the dissemination database.

1.2 Dissemination

Each month the SOeS publishes the number of homes granted planning permission and started from Sitadel2 data. The figures are broken

down by building type (detached, flats, etc.) and by geographical area.

2. The limitations of the old method and the DR+ project

2.1 Previously published data

Two types of monthly series were published until January 2015 from the Sitadel2 database.

- real date series (DR);

- date of collection series (DPC).

The real date series ultimately reflect the reality of the construction over time and must take priority over the date of collection series for carrying out structural studies. The time scale for providing these series is relatively long as it depends on the time scale of returns of permissions or implementations, variation of permissions and set aside decisions. It takes about 6 months for the number of authorized dwellings in a given month to stabilise.

The date of collection series count the flows received each month. They are available quickly but have the disadvantage of being sensitive to collection fluctuations (receipt and content of files from LPAs, monthly follow-ups to applicants to obtain more information on implementation).

These series were the only ones that could be used for economic purposes. The changes observed from the date of collection series actually enable us to estimate the actual date series provided the collection is carried out regularly. However, economic downturns are seen with a time lag that depends on the average rate of data return. In addition, several collection shocks have occurred in recent years, with a significant impact on economic monitoring in some regions and at the national level.

2.2 Aims of the DR+ project concerning building permits

A project to estimate authorization and implementation “by real date” was started at the end of 2012 within SOeS. It consisted of producing robust monthly estimates of collection fluctuations, according to the same schedule as the date of collection series (dissemination of

month *m* at the end of month *m+1*) and ensuring a better relationship between the authorization series and the implementation series.

The information is considered exhaustive: all permits will be collected in the end. The accepted approach is to deal with a "rate of convergence" problem.

3. Estimation of the number of authorized dwellings

Analysis of the real date series of authorized dwellings shows that it has stabilised overall from the 24th month of collection. The collection of permits granted in month *t* becomes negligible after *t*+24 (less than 2% of the total number of dwellings and less than 0.1 point change from one month to another).

3.1 Definition of the estimation indicator

Let A (*t*) be the number of dwellings authorized in month *t* by real date, and A(*t, m*) be the number of dwellings authorized in *t* and collected in month *m*.

A(*t*) can be broken down as follows:

$$\forall t, \mathrm{A(t)} = \mathrm{A(t,\ t)} + \mathrm{A(t,\ t+1)} + ... + \mathrm{A(t,\ t+\infty)}$$

By considering that the series is stable at the end of 24 months, the following approximation can be made:

$$\forall t, \mathrm{A(t)} = \mathrm{A(t,\ t)} + \mathrm{A(t,\ t+1)} + ... + \mathrm{A(t,\ t+23)} \qquad (1)$$

To keep matters simple, in this document we consider relationship (1) to be true.

The estimator is therefore written:

$$\forall t, \hat{\mathrm{A}}\mathrm{(t)} = \hat{\mathrm{A}}\mathrm{(t,\ t)} + \hat{\mathrm{A}}\mathrm{(t,\ t+1)} + ... + \hat{\mathrm{A}}\mathrm{(t,\ t+23)}$$

Let *m* be the last month of collection, all the Â(*m-i*), for *i* between 0 and 23, must be estimated. The method developed will consist of estimating the evolution over the last 24

months and to chain them from the last point of the stabilised series: A(m-*24*).

We are therefore looking for an estimator of the form:

$$\forall t, \hat{A}^{m}(t) = A(m-24) \times \hat{E}^{m}(m-24, t)$$

for *t* between *m-23* and *m*, with $\hat{E}^{m}(m-24,t)$ the estimator in *m* of the evolution in authorized dwellings between *m-24* and *t*.

3.2 Choice of estimator over the last 24 months

Due to the time scale of information returns, the calculation of evolution directly observed from the actual date series is biased, particularly for the most recent months. In fact, the more recent the month, the greater the volume of information "still to be collected". For example, comparing the numbers of dwellings authorized in January 2015 and January 2014, in January 2015, involves comparing a figure built with about 70% of the information with a near-definitive figure. It is therefore necessary to find another estimator of evolution of the number of dwellings authorized over the end of the period.

A "natural" estimate of the evolutions of the number of dwellings authorized by real date given the current dissemination system, would be to use the evolutions observed from the date of collection point of view. However the date of collection series is sensitive to collection fluctuations which can impact on the changes.

It is possible to lessen these collection adjustments by systematically eliminating the permits received with a significant time lag. Thus, the date of collection series truncated at *d* months, removes the building permits which have an information receipt time period (difference between the date of collection in Sitadel2 and the actual authorization date) greater than *d* months. This series is written DPC_d.

$$\forall t, DPC_d(t) = A(t, t) + A(t-1, t) + \ldots + A(t-d+1, t)$$

By using this DPC_d series, the estimation of the number of authorized dwellings is written thus: for *t* such as $m-23 \leq t \leq m$,

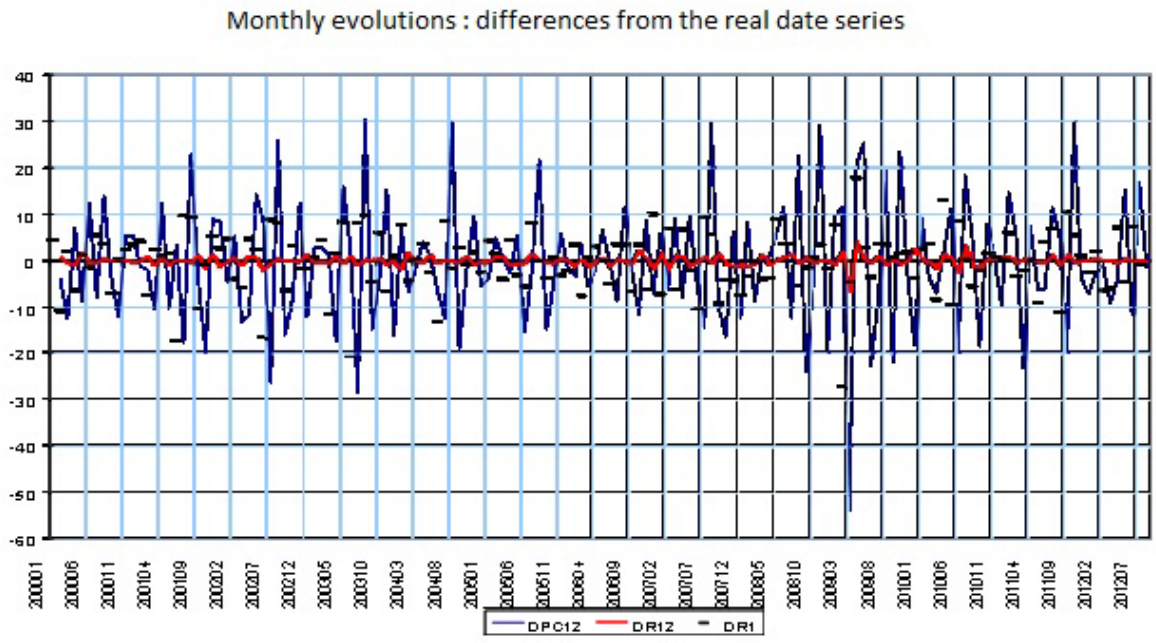
$$\hat{A}^{m}(t) = \mathrm{A(m\text{-}24)} \times \frac{DPC_d(t)}{DPC_d(m-24)}$$

However, monthly changes in the DPC_d series are often still very far from those of the actual date series as figure 1 illustrates (with d=12). There is also a delay of nearly two months compared to DR series. The series of authorized dwellings by the real date truncated at *d*, is defined as the number of dwellings authorized in *t* and collected before *t*+*d*. This DR_d (*t*) series can be written as follows:

$$\forall t, DR_d\,(\mathrm{t}) = \mathrm{A(t,\ t)} + \mathrm{A(t,}\ t+1) + \ldots + \mathrm{A(t,\ t+d\text{-}1)}$$

The truncated real date series provides better estimates of actual date changes than truncated date of entry series, but in variable proportions according to *d*. Figure 1 compares the differences between monthly evolution in the real date series compared with three other series. DR_{12} is the series with the smallest differences compared with the actual date series.

Figure 1: comparison of 3 estimators for estimating monthly variations by actual data



Monthly evolutions obtained from the $DR_1(t)$ series are frequently more than 5 points greater than those ultimately observed by real date, although A(*m, m*) represents about 70% of A(*m*). This proportion has been shown to be relatively unstable over time, between 50% and 80%.

Several series were tested, the one whose evolutions are closest to the real date series is the real date truncated at 12 months series (DR_{12}). This series is chosen for estimating evolutions in the real date series. However, the last points of this series are not known at the time of dissemination and must then be estimated.

3.3 Construction of the actual date truncated at 12 months series

The $DR_{12}(t)$ series can be related to the series of authorized dwellings by the date of collection truncated at 12 months by considering the proportion p(t,m) of authorized dwellings in *t* in the data return in m: $p(t,m)=\frac{A(t,m)}{DPC_{12}(t)}$

Thus each term of the DR_{12}(t) series is broken down as follows:

For *i* from 0 to 11:

$$\forall t,\ \mathrm{A(t,\ t+i)} = DPC_{12}(t+i) \times P(t,t+i)$$

With $DPC_{12}(t+i)$: the number of dwellings collected in $t+i$ for which the data receipt period is less than twelve months;

$P(t, t+i)$: the proportion of authorized dwellings in *t* in the data return in $t+i$.

Let m be the last month of collection. On the one hand, term A(t, $t+i$) is directly observed when t + i ≤ *m*. On the other hand, if t + i > *m*, A(t, $t+i$) must be estimated.

A(t, $t+i$) is then estimated as follows:

$\widehat{A^m}(t,t+i) = DPC_{12}^m(t+i) * p^m(i)$ for t+i>m

The DPC term is estimated through an ARIMA model (it fits an airline model) which extends the date of collection truncated at 12 months series.

$\hat{p}^m(i)$ is the average, over the last 6 months of collection, of the proportions of dwellings collected *i* months after authorisation: $\hat{p}^m(i) = \frac{1}{6}\sum_{k=0}^{5} p(m-k-i, m-k)$

The real date truncated at 12 months series is thus estimated in the following way:

For $t \leq m$ -12: $$\widehat{DR_{12}^m}(t) = DR_{12}(t) = \sum_{i=0}^{11} A(t, t+i) \quad (3)$$

For m-12 < $t \leq m$ $$\widehat{DR_{12}^m}(t) = \sum_{i=0}^{m-t} A(t, t+i) + \sum_{i=m+1-t}^{11} \widehat{A^m}(t, t+i) \quad (4)$$

3.4 Chaining changes or resetting the truncated real date series

Monthly evolutions estimated from the truncated real date series are chained from the last point of the stabilised *m-24* series. The estimation of the number of authorised dwellings Â(*t*) can be written as follows:

For *t* such as $t \leq m\text{-}24$ $$\widehat{A^m}(t) = A(t) = \sum_{i=0}^{24} A(t, t+i)$$

For *t* such as m-23 ≤ $t \leq m$, $$\widehat{A^m}(t) = A(m-24) * \frac{\widehat{DR_{12}^m}(t)}{DR_{12}(m-24)} \quad (5)$$

4. Coherence with the existing administrative return system

The publication of estimates at national and regional level does not meet the analysis needs for the smallest territories, at the municipal level for example. For regions, however, administrative collection is satisfactory: the current system meets these needs. It is therefore planned to maintain the dissemination of the current series at small geographical levels. This

information will be published in a “Sitadel administrative database” part, as opposed to the “statistical” part in which the new estimates will be disseminated.

It will therefore be possible, for a given month, to reconstruct, from the administrative database, the real date series at the regional or national level and compare them with other disseminated estimates. Coherence between the two information systems (statistical and administrative) must therefore be sought. In order to ensure this coherence, a constraint is imposed on real date estimations: they cannot be less that the level of real date permissions already collected. This constraint also conveys the estimation logic: it consists of adding an estimation part to the collected real date series.

In other words, the final estimator accepted is the following:

$$A_f^m(t) = max\left(A^m(t), \sum_{i=0}^{m-t} A(t,t+i)\right) \qquad (6)$$

Where $\sum_{i=0}^{m-t} A(t,t+i)$ represents the number of authorized dwellings in *t* from information collected between *t* and *m*. Revisions of estimates between m and m+1

Over the months, new data is collected, which will generate successive revisions of estimates. In fact, in month *m*, term A(*m,m+1*) was estimated while it is directly observed in *m+1*. Furthermore, terms $\widehat{DPC}_{12}^{m}(m+i)$ and $\hat{p}^m(i)$ are re-estimated. Finally, the resetting point is moved from *m-24* to *m-23* which is also likely to lead to revisions.

All the revisions are taken into account, as new information improve quality of the time series, minimizing the estimated part of the indicator.

Each month, the estimated data will thus be replaced, either by collected data or by new estimates taking account of the latest collected information. The greatest revisions between *m* and *m+1* are concentrated on month *m* in view of the weight of term A(*m,m+1*) compared with the other terms A(*m,m+2*), …

A(*m,m+11*). In fact, A(*m,m+1*) represents 15 to 20% of Â(t) compared with a little less than 10% for the total of the other terms.

5. A new indicator and the consequences

To sum up in a few words, the new indicator first focuses on recent short-term evolutions, filtering collected data from unwanted information. These evolutions are then chained to long-term level in order to produce an easily understandable figure.

This new indicator has several advantages over the previous one: it is more accurate and less sensitive to regulatory changes. Nevertheless, people are not comfortable with revisions and would prefer the unrevised previous one, because the DPC indicators have been used for a long time. But we hope users will become more comfortable with DR+ indicators with more time. More information is available here (in french) : *http://tinyurl.com/zqsj8h9*

Estimations are also computed for smaller levels: by type of housing and by different geographic scale. These series are calibrated to insure coherence between themselves but can no more be linked to micro-data.

In figure 2, previous and new indicators are displayed. Real dates figures and estimated ones are equals except for the last 24 months. The flaws of the DPC figures are particularly visible during the years 2009 and 2012.

Figure 3: comparison of 3 times series for the number of authorized dwellings (12 month total)