Un outil pour l'exploitation des graphiques en histoire : la régression linéaire

Yves Le Guillou

Un outil pour l'exploitation des graphiques en histoire : la régression linéaire[1]

Yves LE GUILLOU

Abstract. In history, the typical graph associates numerical data, as extracted from archival documents, to a date. This data is always discrete and forms a graph called a dot line. In order to make it easily readable, the dots of that line are linked together in an appropriate way, thus creating a continuous line. This continuous line brings to the fore clear tendencies, but cannot be described with precision, since it is not "defined" with precision. To describe this graph more precisely, the statistical tool of "linear regression" is used.

Les archives, et surtout les archives anciennes, livrent rarement des séries continues ou complètes de documents, d'une part parce que les historiens n'ont parfois pas la persévérance qu'exige l'heuristique, et d'autre part surtout parce que les siècles ont détruit nombre de documents. Cependant, il arrive qu'avec un peu de chance, de perspicacité et de courage, l'historien puisse constituer une ou des séries quasi complètes. Leur exploitation nécessite alors l'application de méthodes appropriées en vue d'une analyse et d'une synthèse claires et concises. Les données, quantifiées ou quantifiables, recueillies dans ou à partir

[1] Je remercie Guillaume GRÉGOIRE, chercheur en physique du chaos, d'avoir bien voulu m'aider dans la partie mathématique de cet article.

✉ Yves LE GUILLOU, archiviste paléographe, École nationale supérieure des sciences de l'information et des bibliothèques (ENSSIB).
Adresse personnelle : 127, rue Marcadet, F–75018 PARIS.
E-mail : guillou@enssib.fr

des documents, sont organisées en tableaux puis synthétisées sous forme de graphiques, représentations de nuages de points. L'exploitation et l'interprétation de ces nuages de points relèvent des méthodes de calcul statistique. Ces méthodes, très utiles aux historiens contemporanéistes de par l'abondance et la diversité des sources dont ils disposent, ne sont guère utilisées par les historiens des périodes plus anciennes. Or il semblerait que certaines méthodes de calcul empruntées aux statistiques puissent profiter beaucoup plus souvent à ces derniers.

Le but du présent article n'est pas de disserter sur l'apport général du calcul statistique à l'histoire, ce dont je serais bien incapable, mais de livrer quelques réflexions sur un exemple d'outil dont je me suis servi au cours de mes recherches : la régression linéaire.

Un champ d'application de la régression linéaire

L'exemple à partir duquel nous allons réfléchir, et qui viendra illustrer cet exposé, se rencontre tellement fréquemment en histoire qu'on peut le qualifier d'exemple-type. Il s'agit de représenter l'évolution de données numériques en fonction du temps. À chaque unité de temps — le plus souvent l'année — correspond une donnée numérique — le plus souvent une quantité — tirée ou déduite des documents. Le nuage de points que représente le graphique prend visuellement une forme particulière. Plus les points sont nombreux, plus la forme est précise. On peut également relier les points entre eux pour obtenir l'illusion d'une courbe. Cette courbe a la vertu de montrer des tendances sur une certaine durée, mais elle ne répond à aucun modèle mathématique et ne peut donc être commentée avec rigueur et précision. En effet, le nuage de points n'est qu'un ensemble de valeurs discrètes qui ne peuvent être considérées qu'en elles-mêmes ou les unes par rapport aux autres. On ne tire alors du graphique qu'une image « fixe » ou une succession d'images « fixes » sans qu'aucune tendance puisse être chiffrée et donc comparée avec d'autres tendances. On peut dire seulement que telle valeur de telle année est *n* fois supérieure ou inférieure à telle autre valeur de telle autre année. Pourtant, la mise en graphique montre des tendances propres à des périodes de temps plus ou moins

longues, et il est paradoxal de pouvoir voir ces tendances sans pouvoir en parler autrement que par des termes aussi flous que « ça augmente » ou « ça diminue ». La seule manière de quantifier ces « augmentations » et ces « diminutions » avec le plus de précision possible est de quantifier des tendances significatives suivies par des ensembles de points consécutifs. Si l'on admet qu'une droite peut donner une bonne image de la tendance du nuage de points sur une période donnée, alors la quantification de cette tendance sera le coefficient directeur de cette droite. La méthode va consister à associer à un groupe de points compris entre deux bornes et exprimant une tendance nette de la courbe (si l'on assimile le nuage de points à une courbe) étudiée la droite dont le comportement se rapproche le plus de celui de la courbe ou de la partie de courbe étudiée. Alors, une droite étant un ensemble de points faciles à utiliser, tous les calculs et comparaisons seront possibles. Ce processus s'appelle la régression linéaire d'une courbe [1]. Ainsi le coefficient directeur permet-il de dire à quelle vitesse évolue le nuage de points à un moment donné ou sur un intervalle de temps : on passe d'une image « fixe » de l'évolution à une image « dynamique ». De plus, la connaissance d'une tendance permet de pallier l'absence de données pour certaines années, si l'on suppose que le nuage de point est cohérent.

Un exemple : l'évolution comparée des fortunes foncières de Claude Bouthillier et Claude de Bullion

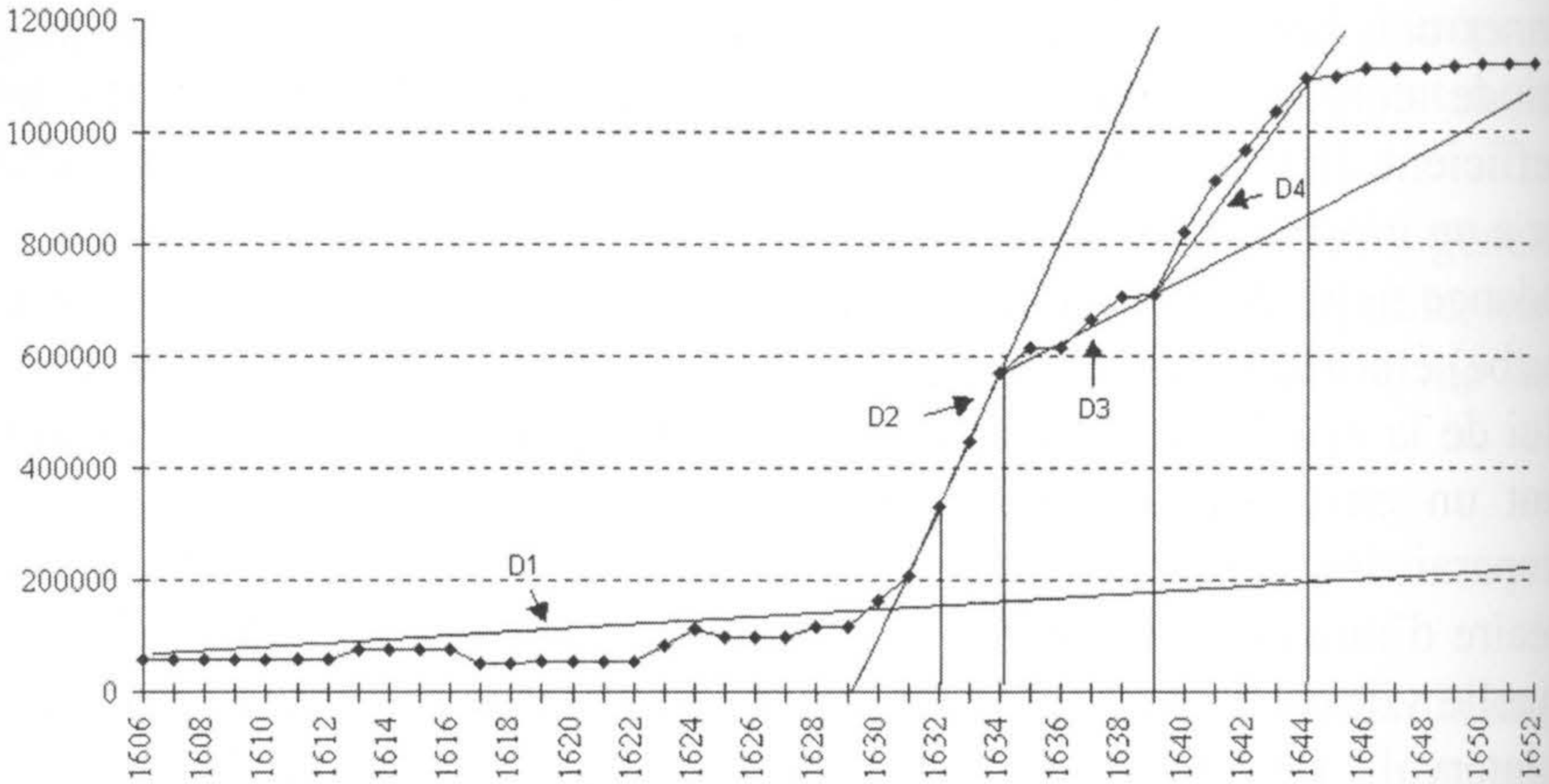
Nous avons comparé à l'aide de cette méthode l'évolution des propriétés foncières de Claude Bouthillier (graphique 1) avec celle de Claude de Bullion (graphique 2). Ces deux personnages sont conjointement surintendants des finances entre 1632 et 1640. Le graphique 1 [2]

---

[1] Pour l'explication mathématique de la régression linéaire, cf. le paragraphe sur l'ajustement linéaire dans BOURSIN (Jean-Louis) : 1991, *Comprendre la statistique descriptive* (Paris), p. 103–113.

[2] Le graphique 1 est constitué à partir de mes recherches. Cf. LE GUILLOU (Yves) : 1997, *Les Bouthillier, de l'avocat au surintendant (ca 1540–1652). Histoire d'une ascension sociale et formation d'une fortune*, (thèse d'École des chartes), 397 p.

comprend une série continue de valeurs numériques, c'est-à-dire une par année, alors que de nombreuses données manquent au graphique 2[1], qui représente une suite de points très clairsemée. Pourtant des tendances nettes se dégagent, qui peuvent être chiffrées et donc comparées avec précision.



Graphique 1
Évolution de la fortune foncière (terres et maisons)
de Claude Bouthillier

Les tendances que j'ai jugé significatives pour le graphique 1 sont représentées par les droites ou régressions linéaires D_1, D_2, D_3 et D_4, et pour le graphique 2 par les droites D_5 et D_6[2]. Les coefficients directeurs de ces droites sont : $a_1 \cong 5\,666$, $a_2 \cong 247\,935$, $a_3 \cong 56\,544$, $a_4 \cong 100\,883$, $a_5 \cong 7\,714$, $a_6 \cong 314\,648$.

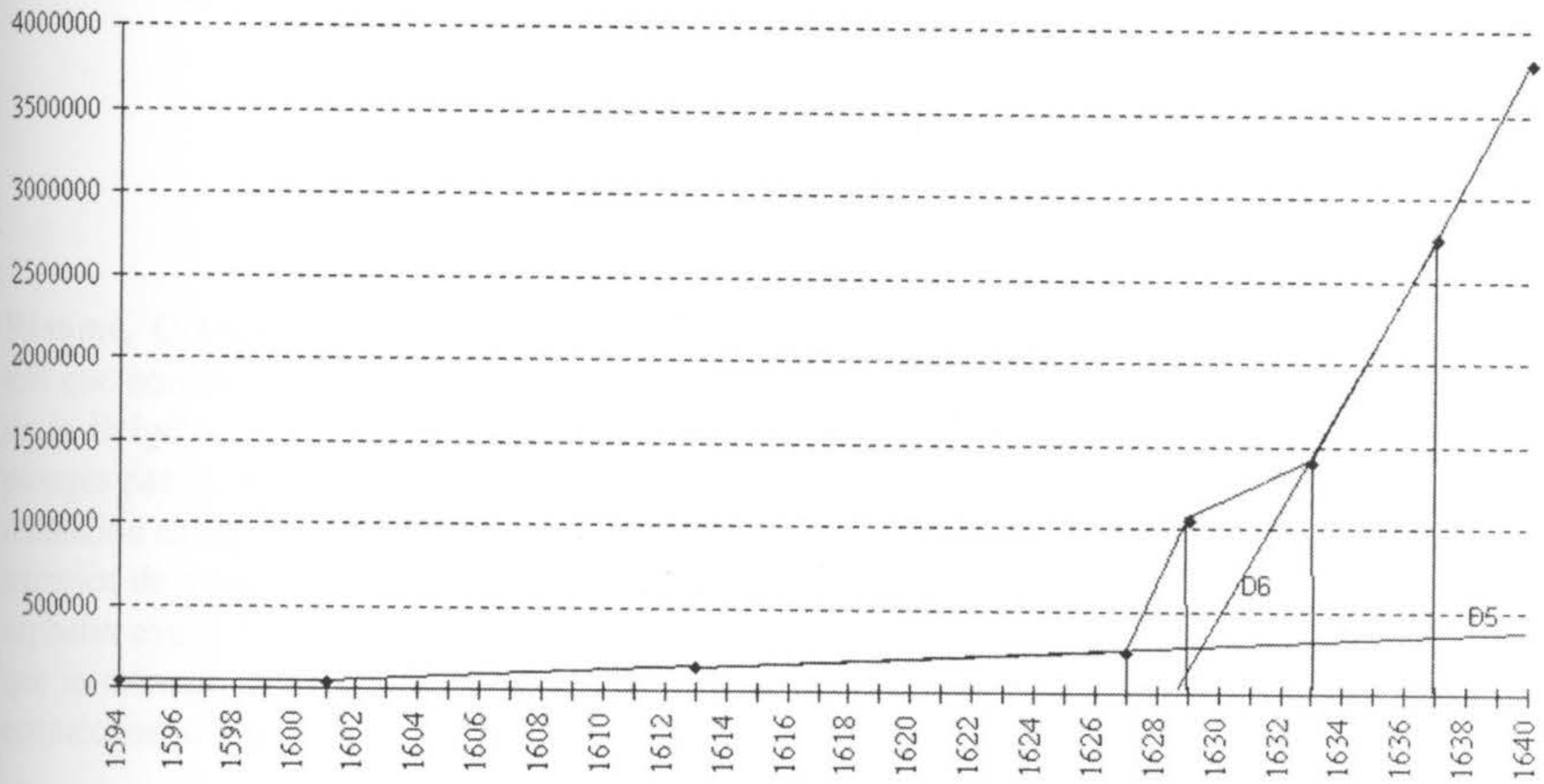
Ainsi peut-on dire que la fortune foncière de Bouthillier s'accroît beaucoup moins vite que celle de Bullion sur la même période. Entre 1632 et 1634, sa croissance est 1,4 fois moins rapide. Entre 1634 et

---

[1] LABATUT (Jean-Pierre) : 1987, *Noblesse, pouvoir et société en France au XVII^e siècle* (Limoges), p. 50.

[2] Nous prions le lecteur de bien vouloir excuser le tracé approximatif de ces droites ; ces dernières n'ont de valeur qu'explicative.

1639, elle est 6 fois moins rapide, et après 1639, elle reste 3,4 fois moins rapide. Alors que l'évolution de leurs fortunes foncières est comparable entre 1601–1606 et 1627, c'est-à-dire au début de leur carrière, l'exercice conjoint de la surintendance entre 1632 et 1640 ne les fait pas fructifier dans les mêmes proportions. En moyenne, celle de Bullion s'accroît 2,5 fois plus vite que celle de Bouthillier, ce qui est énorme.



Graphique 2
Évolution de la fortune foncière (terres et maisons)
de Claude de Bullion

Il ne revient pas à cet article de tirer les conclusions de ces observations. Son but ne consiste qu'à exposer la méthode qui a permis de telles observations. Cette méthode est claire, précise, rapide, concise ; à condition, bien sûr, de disposer d'une bonne calculatrice scientifique qui permette la programmation de la formule du coefficient directeur de la régression linéaire ou qui contienne déjà ce programme [1].

---

[1] À titre indicatif, la formule est la suivante :
si x est une année et y, la valeur numérique qui lui correspond, alors le coefficient directeur de la régression linéaire du nuage de n points est :

$$a = (y_1 + ... + y_n)(x_1 + ... + x_n) - n(y_1x_1 + ... y_nx_n) / (x_1 + ... + x_n)^2 - n(x_1^2 + ... + x_n^2)$$