Statistical Hotelling

Théo Gaboriaud

November 6, 2025

Abstract : *This work aims at studying how data collection regulation can have an impact on market welfare. It focuses on horizontal product differenciation through the Hotelling model, and investigates the market dynamics when a monopolist and a regulator face decisions under statistical uncertainty.*

1 Introduction

In this section, we motivate the study of the impact of data collection regulation on market welfare, and the choice of model associated.

1.1 Why should we study data collection incidence on markets ?

On markets, the decisions taken by firms are more and more supported by information about consumers gathered through data collection. One the one hand, more information can be beneficial to all economic agents by leading the firm to offer better suited services. On the other hand, this practice may be detrimental to the consumer both economically and from a privacy standpoint. Thus, data collection calls for regulation.

Debates about the regulation of data collection include the question of its incidence on market equilibria. Indeed, the amount and quality of data collected by the firms change their level of uncertainty when taking decisions such as choosing their prices or their products features.

Choosing the right amount of uncertainty requires to understand how it changes the amount and repartition of economic welfare in markets. As of now, most of this work has been done empirically and we don't yet understand all the mechanisms at stake. Hence the objective of this work :

Objective. *The aim of this work is to provide the litterature with a theoretical analysis of the impact of data collection regulation on welfare within an economic model.*

1.2 The Hotelling model to study horizontal product differenciation

One key way firms use data is to estimate the consumers's preferences to choose their products features. That is, data collection enables firms to offer products that are of better interest to consumers. However, if the firm knows the consumers's preferences, it becomes more confident in the fact that consumers are going to like the product and it may then increase its price.

Thus, it is of interest to study the impact of data collection on an horizontally differenciated product market. As it is a first step, this work focuses on the monopolist behaviour, but an interesting extension would be to study a market with more than one firm.

The most classical and studied model for horizontal product differenciation is the Hotelling model. This model is fairly simple and has provided the economic litterature with insights on the dynamics of this kind of market and is thus a natural choice.

1.3 *The uncertainty framework*

The Hotelling model involves the description of the consumers's preferences through a distribution μ. In the seminal work by Hotelling [1], the consumer distribution μ is assumed to be uniform. This choice is mostly justified by the ease of calculations and the existence of closed form solutions. Because we want to model the firm uncertainty about the preferences of consumers, we cannot assume *a priori* a uniform distribution. Thus, we have to work in a more general case. Furthermore, because the distributions of consumer preferences could be anything, we choose to investigate a distribution free approach.

The uncertainty framework is provided by statistics. We assume that the firm has access to some (maybe noised) samples of the distribution of the preferences of consumers, and that it uses them to optimally behave the market. However, it remains to define what optimally means. The approach we choose and justify is the following :

Idea (High-level description of the uncertainty framework)**.** *We assume that the firm, given a vector $\vec{x}$ of samples $x \sim \mu$ of the consumers's preferences distribution, computes a family of distributions $\mathcal{F}(\vec{x})$ so that $\mu \in \mathcal{F}(\vec{x})$ with great probability. Then, it optimises its behaviour for the worst distribution in $\mathcal{F}(\vec{x})$. That is, if the firm wants to compute the argmax $\theta^\star$ of some function $f(\theta, \mu)$ depending on the distribution, it will choose*

$$\theta^\star(\vec{x}) = \arg\max_{\theta} \min_{\mu \in \mathcal{F}(\vec{x})} f(\theta, \mu)$$

This framework[2] has several advantages : it is distribution-free and thus very general, it is fairly realistic as it is likely for firms to be risk-averse, and finally, it exhibits closed forms solutions, which is not the case if we work with expectations on the data sampling process.

[1] Harold Hotelling. Stability in competition. *The Economic Journal*, 39:41–57, 1929

[2] See next section for a more rigourous formulation.

2 Model

In this section, we present the version of the Hotelling model that we use, and formalize the optimisation problems the firm and regulator face.

2.1 Setup of the Hotelling model

We consider a population of consumers with heterogeneous preferences [3] for a differentiated product $\theta \in \Theta$ distributed according to measure $\mu \in \mathcal{P}(\Theta)$ for a good, we will call it the consumers's preference distribution, or consumer distribution. We work on the case where $\Theta = \mathbb{R}$ The consumers' valuation for the good is noted $v \in \mathbb{R}$. The products have a type on the same space as the consumers' preferences, so that the firms' product's type is also noted $\theta_f \in \Theta$. We define the utility :[4]

Definition 2.1 (Utility). *We assume that the consumer have a common intrinsic valuation for the good noted $v \in \mathbb{R}$ and a quadratic transport cost. Thus, we define the utility of a consumer of type θ for buying a product of type θ_f and price p is given by :*

$$u_\theta(\theta_f, p) = v - p - (\theta - \theta_f)^2$$

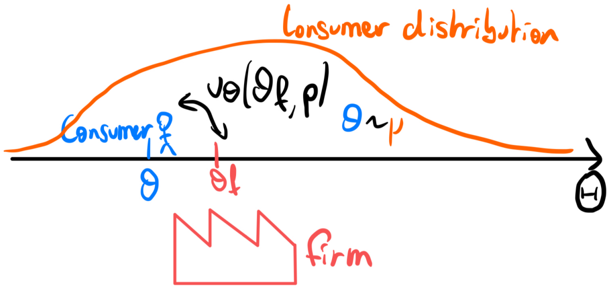

Figure 1: An illustration of the Hotelling model with consumer distribution μ.

The demand for a location θ_f and a price p is then

$$D_\mu(\theta_f, p) = \int_{\theta\in\Theta} \mathbf{1}_{u_\theta(\theta_f,p)\geq 0}\mu(d\theta)$$

that is the mass of consumers that buy the product.

We assume a marginal cost of $c \in \mathbb{R}$ for the firm, so that its profit is given by :

$$\pi_\mu(\hat{\theta}, p) = (p - c) \cdot D_\mu(\hat{\theta}, p)$$

The consumer surplus is given by the net utility of the consumers :

$$CS_\mu(\theta_f, p) = \int_{\theta\in\Theta} u_\theta(\theta_f, p)\mu(d\theta)$$

The regulator maximises the weighted total welfare that we define above.

[3] We choose a quadratic cost for simplicity, as it makes variances appear in the latter calculations.

[4] The utility of not buying the product is normalized to 0. For ease of calculations, we consider that tiebreaks are unfavorable, that is, a consumer with $u_\theta(\theta_f, p) = 0$ does not buy the product.

Definition 2.2 (Total welfare)**.** *The regulator chooses a social preference parameter $\alpha \in [0,1]$ and is interested in the weighted sum of the firm's profit and the consumer surplus, that is :*

$$\mathcal{W}_\mu(\theta_f, p) = CS_\mu(\theta_f, p) + \alpha\pi_\mu(\hat{\theta}, p)$$

2.2 *The efficiency of an allocation*

For any location $\theta_f \in \Theta$, the total welfare is maximized for $p = c$ because the total welfare is an increasing function of the length of the interval of the consumer served (which is equal to $2\sqrt{v - p}$) and a decreasing function of the price. The regulator's goal is thus to maximise the number of consumers served by allowing the firm to choose a good location (which is aligned with the firm's objective) but by lowering its price as much as possible. The specific case of $\alpha = 1$ and a symmetric distribution is showcased below.

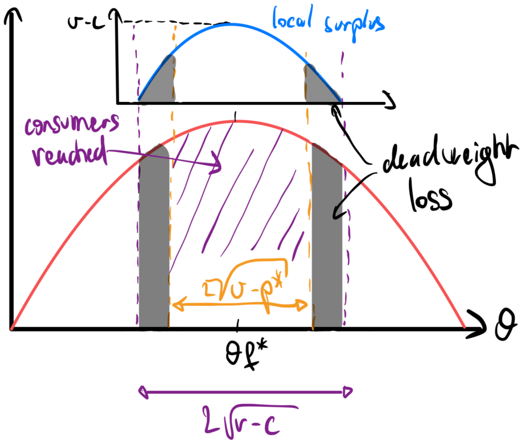

Figure 2: An illustration of the deid-weightloss, that is the total surplus that is lost with respect to the best choice of (θ, p) for the simpler case of $\alpha = 1$ and a symmetric distribution. $\alpha = 1$ makes the total welfare a sum of the mass of the consumers weighted by their proximity to the firm.

2.3 *Uncertainty formalization*

Let us consider that the firm observes N samples from the consumer distribution. Let us note $(\Theta = \mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{P})$ that we shorten to $(\Theta, \mathcal{P})$ its statistical model. In this section, we formalize how the firm may choose its location and price using these samples.

First, the firm uses its samples to define a family of distributions it is going to consider for its optimisation problem. Hence the definition below :

Definition 2.3 (Data processing mechanism)**.** *Let $(\Theta, \mathcal{P})$ be a statistical model and $N \in \mathbb{N}$ a number of samples. A data processing mechanism is a function $\mathcal{F} : \Theta^N \to P(\mathcal{P})$.*

Example : If the firm has iid samples and already knows the variance of the distribution $\sigma^2 \in \mathbb{R}$, it can use its samples to estimate its mean and only consider distributions with means within a confidence interval.

More rigourously, if the statistical model of the firm is $(\Theta, \mathcal{P}_{\sigma^2})$ with $\mathcal{P}_{\sigma^2} = \{\mu \in \mathcal{P}(\mathbb{R}), V_{X\sim\mu}[X] = \sigma^2\}$, and given a function $d : \mathbb{N} \to \mathbb{R}$ that computes the length of confidence intervals for the mean of the distribution, we can define the data processing mechanism $\mathcal{M}_d$ so that

$$\mathcal{M}_d(\vec{x}) = \{\mu \in P_{\sigma^2}, \mathbb{E}_{X\sim\mu}[X] \in [\overline{x} - d(N), \overline{x} + d(N)]\}$$

where $\overline{x}$ is the mean of the sample. d is obtained by using concentration results and by expliciting the risk-averseness of the firm δ (for instance the Hoeffding inequality when the firm considers that the mean belongs to a bounded support).

We then introduce the worst case functions notation :

Definition 2.4 (Worst case functions)**.** *Let $\mathcal{F}$ be a data processing mechanism and N be a number of samples. Then, to every function $f_\mu : \Lambda \to \mathbb{R}$ parametrized by a distribution μ, we associate the worst case function $f_{\mathcal{M}}$: $\Theta^N \times \Lambda \to \mathbb{R}$ defined by :*

$$f_{\mathcal{F}}(\vec{x}, \lambda) = \min_{\mu \in \mathcal{F}(\vec{x})} f_\mu(\lambda)$$

3 Results

3.1 Known variety of tastes

In this section, we consider that the firm knows the variance of the distribution of consumer tastes σ^2. That is, the statistical model of the firm is $(\Theta, \mathcal{P}_{\sigma^2})$ and its data processing mechanism is $\mathcal{M}_d$ with the same notations as in this example. For the rest of the section, we will confuse the function d with the value $d(N)$. We investigate the worst case functions in this case.

Proposition 3.1 (Worst case demand)**.** *Let $\vec{x}$ be the data collected by the firm. Noting $\mathcal{M}_d$ the mechanism defined in this example, it holds that :*

$$D^\star_{\mathcal{M}_d}(\vec{x}, p) := \max_{\theta \in \Theta} D_{\mathcal{M}_d}(\vec{x}, \theta, p) = 1 - \frac{\sigma^2 + d^2}{v - p}$$

This maximum is reached for $\theta = \overline{x}$, that is when the firm locates at the mean of its samples. Because the value of the worst case demand does not depend on the sample, we also note it $D^\star_{\mathcal{M}_d}(p)$.

Comments :

- The formula is similar to the one we would've obtained if we applied the Chebychev bound to upper bound the deviation from the mean, althought we are computing a deviation to a point that is *near* the mean (up to an error d). We see that this cost behaves like adding a quadratic term d^2 to the variance.
- The speed of the decrease in p is only due to the form of the cost. That is, if we considered a transport cost like $(\theta_f - \theta)^\alpha$, we would have $(v-p)^{\frac{\alpha}{2}}$ at the denominator.
- Because the firm cannot price below its marginal cost c, we see that the market can only open if $\sigma^2 + d^2 \leq v - c$. If the consumers are too spread, the firm cannot make profit, which is natural within the Hotelling model, but in our framework we see that another problem is that, if the firm is too uncertain about the distribution, it will not enter the market.

Now that we know about the demand, we can solve the optimisation problem of the firm and compute its optimal price in terms of the parameters of the market.

Its worst case profit is given by

$$\begin{aligned}\pi^\star_{\mathcal{M}_d}(\vec{x}, p) :&= \max_{\theta \in \Theta} \pi_{\mathcal{M}_d}(\vec{x}, \theta, p) \\ &= (p - c) \cdot D^\star_{\mathcal{M}_d}(p) \\ &= (p - c)\left(1 - \frac{\sigma^2 + d^2}{v - p}\right)\end{aligned}$$

We also note this value

This value does not depend on the data $\vec{x}$ of the firm[5], we shorten it to $\pi_{\mathcal{M}_d}(p)$. We just have to maximise this expression in p to get the following result :

Proposition 3.2 (Worst case profit and optimal price). *Let $\vec{x}$ be the data collected by the firm and $\mathcal{M}_d$ be the mechanism defined in this example.*
If $v - c < \sigma^2 + d^2$, the firm does not enter the market and the optimal price is not defined.
In else case, the optimal price choosen by the firm is :

$$p^\star_{\mathcal{M}_d} := \arg\max_{p \in \mathbb{R}} \pi_{\mathcal{M}_d}(p) = v - \sqrt{(\sigma^2 + d^2)(v - c)}$$

Proof. We compute $\frac{d\pi^\star_{\mathcal{M}_d}}{dp}(p) = 1 - \frac{(v-c)(\sigma^2+d^2)}{(v-p)^2}$. We see notice that this function is positive for $p < p^\star_{\mathcal{M}_d}$ and negative for $p > p^\star_{\mathcal{M}_d}$, thus $p^\star(\sigma, d)$ is the unique maximizer of $\pi_{\mathcal{M}_d}$. The firm enters the market if and only if $p^\star_{\mathcal{M}_d} \geq c$. □

Now that we saw that the profit function behaves nicely [6], let us now turn to the regulator's point of view, and compute the worst case surplus.

Proposition 3.3 (Worst case surplus). *Let $\vec{x}$ be the data collected by the firm. Noting $\mathcal{M}_d$ the mechanism defined in this example, it holds that :*

$$\begin{aligned} \mathcal{W}^\star_{\mathcal{M}_d}(\vec{x}, p) : &= \max_{\theta \in \Theta} \mathcal{W}_{\mathcal{M}_d}(\vec{x}, \theta, p) \\ &= \Big(v - p - d^2 + \alpha(p - c)\Big)\left(1 - \frac{\sigma^2}{v - p - d^2}\right) \end{aligned}$$

This maximum is reached for $\theta = \overline{x}$, that is when the firm locates at the mean of its samples. Because this value does not depend on the sample, we also note it $\mathcal{W}^\star_{\mathcal{M}_d}(p)$.

Comments

- First, the best profit and surplus in the worst are both achieved for $\theta = \overline{x}$. This means that the regulator and the firm would choose the same location. This is not the case considering the model for an arbitrary distribution μ. However, we showed that it is true when we consider the worst case, which is a nice property in the sense that the regulator doesn't have to twist the mean of the samples to make the firm change its location.
- Second, both the profit and surplus do not depend on the actual values of the samples of the firm, they only depend on it through $d(N)$. This means that the regulation process in this case does not need information about the samples that the firms have.

[5] except through $d(N)$

[6] We notice that the function $\pi^\star_{\mathcal{M}_d}(p)$ is concave for $p \in [0, v]$. For an arbitrary distribution μ, considering the tradeoff between attracting more consumers (increasing $D_\mu(\theta, p)$) and maintaining a good margin (increasing $p - c$) may lead to a function profit with local maximas in p, which would result in the fact that the best price $p^\star$ might not be continuous with respect to the distribution.

- Third, we notice that the welfare is positive when $v - p \geq \sigma^2 + d^2$ which is the same condition as the one for the demand to be positive, which is an obvious property that we recover.

Let us now turn to the analysis of trajectory of the welfare as the firm gathers more data. Let us now consider the welfare as a function of the uncertainty.

Definition 3.1. *Let $\mathbb{W} : \mathbb{R} \to \mathbb{R}$ so that $\mathbb{W}(d^2) = \mathcal{W}^\star_{\mathcal{M}_d}(p^\star_{\mathcal{M}_d})$ This function represents the worst case welfare with uncertainty d that results from an optimisation of the worst case profit with uncertainty d.*

Corollary 3.1. *The expression of $\mathbb{W}$ is thus :*

$$\mathbb{W}(d^2) = \left(\sqrt{(\sigma^2 + d^2)(v - c)} - d^2 + \alpha\left(v - c - \sqrt{(\sigma^2 + d^2)(v - c)}\right)\right) \times \left(1 - \frac{\sigma^2}{\sqrt{(\sigma^2 + d^2)(v - c)} - d^2}\right)$$

The question that arises then is : what amount of uncertainty is socially optimal ? Unfortunately, the maximum of this function doesn't have a nice analytical expression for us to interpret. However, we can answer and provide intuition for a more simple question : can uncertainty be socially valuable ? In other words, is $\mathbb{W}$ maximal in 0.

Expanding $\mathbb{W}$ for small d^2 yields:

$$\mathbb{W}'(0) = \frac{\frac{v-c}{2} - (1 + \alpha)\sigma\sqrt{v - c} + \alpha\sigma^2}{\sigma\sqrt{v - c}},$$

The sign of this expression only depends on α and $\frac{\sigma^2}{v-c}$, as shown in this graph :

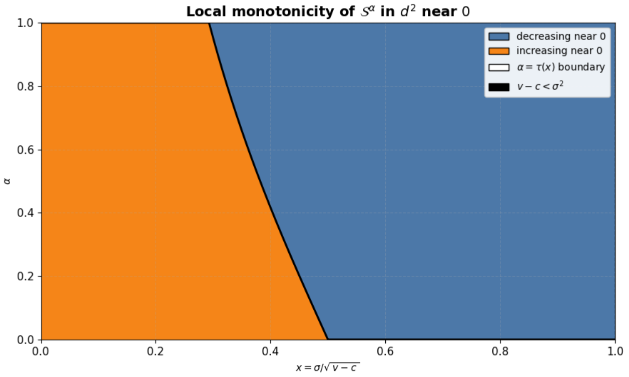

Figure 3: Diagram of when the surplus function $\mathbb{W}(d)$ can be increasing in d near 0.

We see that, for high variance distributions, and for high social preference (low α), it can be socially desirable for the firm to have uncertainty.

4 Discussion

Discussion

Conclusion

Conclure

References

Harold Hotelling. Stability in competition. *The Economic Journal*, 39:41–57, 1929.