From Spinning Disks to Solid State: Modeling the Global Transition in Data-Storage Technology

Michele De Marinis
ID: 2160705

Matteo Gastaldello
ID: 2141225

1 Introduction

1.1 Technological Context

Data storage devices have undergone a radical transformation over recent decades, leading to devices that are vastly different from their predecessors. On one side, we have **Hard Disk Drives (HDDs)**, a traditional mechanical technology that uses rotating magnetic disks to store data, present in the market since the mid-1970s. On the other side, starting from the late 2000s, **Solid State Drives (SSDs)** have emerged as devices based on flash memory technology that contain no moving mechanical parts.
The main differences between these technologies concern:

- Speed: SSDs offer significantly faster data access and transfer times.
- Reliability: The absence of mechanical parts makes SSDs less prone to failures.
- Power consumption: SSDs require less energy to operate.
- Noise: SSDs operate in complete silence.
- Cost: Traditionally, HDDs have offered a lower cost per gigabyte.

1.2 Analysis Objectives

This analysis aims to examine the evolution of worldwide sales volumes of HDDs, since 1976 up to 2022, and of SSDs, only since 2009, with particular attention to the competitive impact that the introduction of SSDs has had on the HDDs' market. The interest in this study stems from the disruptive nature of SSD technology, which represented a clear example of discontinuous technological innovation.
Through the application of various econometric models—from the classic Bass model to more sophisticated technological competition models (UCRCD)—the analysis aims to:

1. identify adoption patterns of both technologies;
2. model the technological substitution dynamics occurring in the data storage sector;
3. evaluate whether the two technologies are in full competition or can coexist in different market niches.

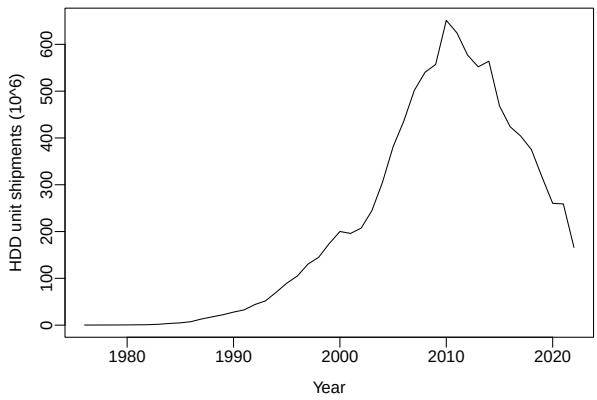

Figure 1: Number of HDDs shipped each year, in millions. The number of shipments grew exponentially until 2010, since when it started to decay.

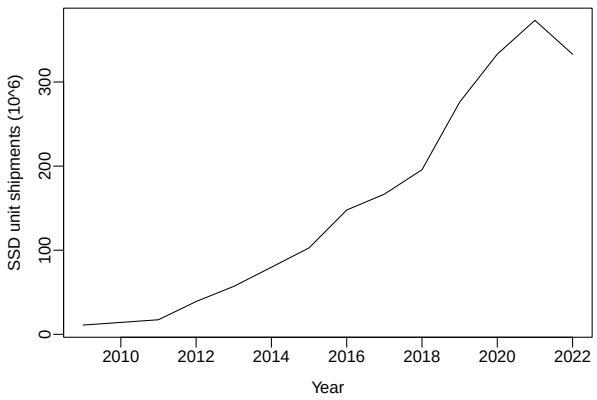

Figure 2: Number of SSDs shipped each year, in millions. Data show a clear increasing trend, with the only exception of 2022, where the increase seems to halt.

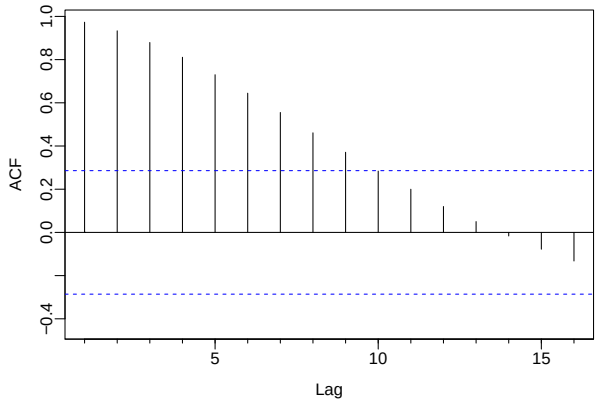

Figure 3: The HDDs autocorrelation function reflects the clear trend observed in the data, with no sign of seasonality.

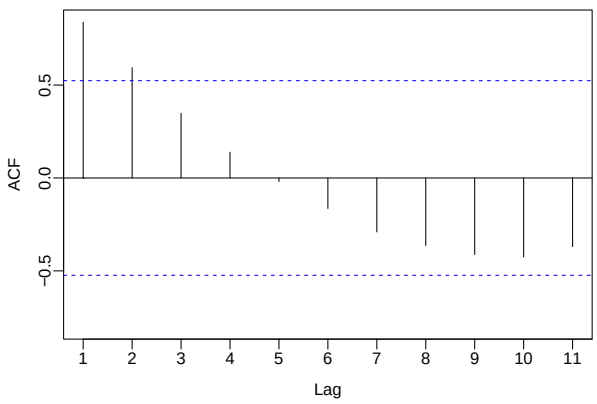

Figure 4: The SSDs autocorrelation function shows significant spikes only at lags 1 and 2, reflecting less autocorrelation than in HDDs.

1.3 Datasets

We have historical time-series data at our disposal regarding unit shipments (in millions) for both storage technologies. For HDDs, the dataset starts in 1976 and ends in 2022. For SSDs, unit shipment data are available from 2009 to 2022. The shorter time series for SSDs reflects the fact that this technology was developed significantly later than HDDs. Time data about HDDs global shipments were found on statista.com [1]. The dataset about SSDs shipments, instead, were manually created merging different sources [2], [4], [5], [6], [7]. Autocorrelation measures the relationship between a time series and a lagged version of itself over successive time intervals. Simply put, it tells you how similar the data points are to each other at different time lags. The auto correlation function (ACF) plots the correlation of the time series with itself at different lags.

For the HDD dataset, the ACF plot shows a high autocorrelation at the initial lags, with values that decay slowly and remain significantly different from zero even at higher lags. This slow decay pattern violates the conditions for stationarity. It is typical of non-stationary processes where shocks to the system have long-lasting effects, such as random walk models.

In contrast, For the SSDs dataset, the ACF displays a prominent spike at lag 1 and then rapidly falls within the 95% confidence bounds for subsequent lags. The lack of significant autocorrelation beyond the

first lag indicates that past values beyond the immediate previous one have negligible influence.

2 Analysis

2.1 The Bass Model

The Bass diffusion model, introduced by Frank Bass in 1969, is a widely used framework for modeling the adoption process of new products or technologies over time. The model captures the dynamics of market penetration by distinguishing between two types of adopters: innovators and imitators.
The Bass model describes the cumulative number of adopters $z(t)$ at time t through the following differential equation:

$$\frac{dz(t)}{dt} = \left(p + q\frac{z(t)}{m}\right)(m - z(t))\,. \tag{1}$$

The **innovation parameter** p is the coefficient of innovation (external influence). It captures the propensity of potential adopters to purchase the product independently of others who have already adopted it. It represents external influences such as mass media advertising or individual openness to innovation. A higher p value indicates a stronger tendency for early adoption. The **imitation parameter** q is the coefficient of innovation (external influence). It represents the propensity of potential adopters to purchase the product due to pressure from those who have already adopted it. It captures internal influences within the social system, such as word-of-mouth communication between adopters and non-adopters. A higher q value suggests stronger social learning effects. The **market potential** m represents the total number of eventual adopters, serving as the upper bound for cumulative adoption and reflects the market size or saturation level. $\boldsymbol{z(t)}$ is the total number of units adopted up to time t.

2.1.1 $\frac{p}{q}$ ratio

The $\frac{p}{q}$ ratio provides insight into the relative importance of imitation versus innovation in the diffusion process. When $\frac{p}{q} > 1$, innovation effects are more prominent, indicating that some people buy first, based on a personal belief. In contrast, when $\frac{p}{q} < 1$, imitation effects dominate, suggesting that some other people buy in a second stage based on the recommendation of someone else (e.g., a friend) or some other external factors (e.g., marketing campaign).

2.1.2 Results of the Bass model

We exploited the Bass model to estimate the initial parameters m, q, and p. The two plots shown in Figure 5 illustrate different approaches to modeling the sales data for the two types of memory systems we are considering, and this dual presentation format will be consistently used throughout our analysis for all subsequent models. The Bass diffusion model was fitted to the HDD sales data, yielding statistically significant parameter estimates for all three key parameters ($p < 0.001$). The estimated market potential m is equal to 1.109×10^4 million units, representing the total addressable market size for HDDs. The innovation coefficient p was estimated at 1.084×10^{-4}, indicating that the rate of adoption by innovators independent of social influence. The imitation coefficient q was found to be 0.212, representing the adoption rate due to the effects of word of mouth and social influence. Notably, q is substantially larger than p. This pattern is typical for technology products where peer influence and network effects are important drivers of market penetration.

2.2 Generalized Bass Model

While the standard Bass model provides a solid foundation for understanding technology diffusion, it assumes that the adoption process occurs in a stable environment without external influences. However, real-world technology adoption is often subject to various external factors such as marketing interventions,

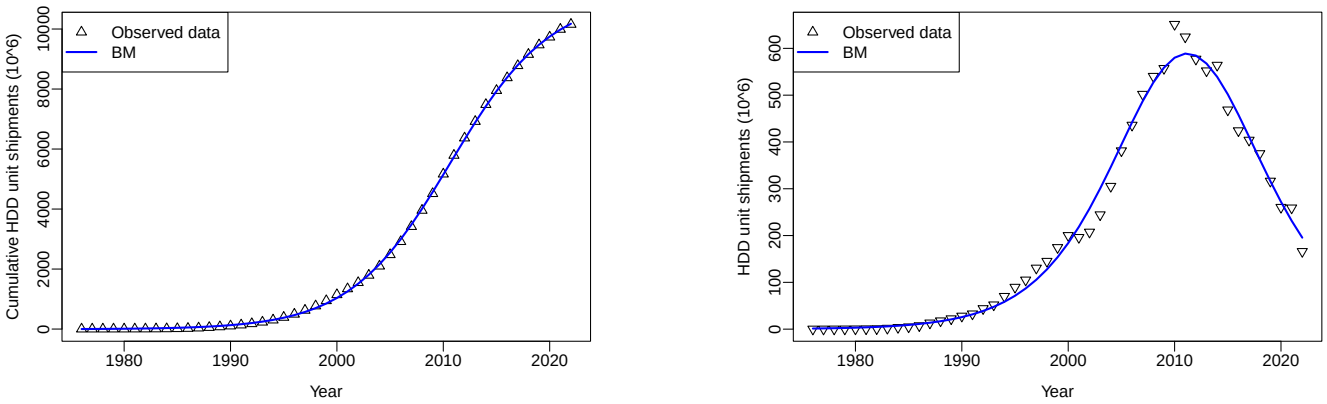

Figure 5: (*left*) Cumulative number of HDDs shipped each year (in millions). The model demonstrates an excellent fit to the observed data, capturing the characteristic S-shaped growth curve typical of technology adoption. (*right*) Number of HDDs shipped each year (in millions), showing a pointwise estimation for each individual year rather than cumulative sales. Here, the Bass model fails to accurately capture the sharp peak in HDD shipment, instead providing a smoother, more symmetric curve that underestimates the actual peak values.

economic shocks, regulatory changes, or technological breakthroughs that can significantly impact the diffusion process. To address these limitations, Bass et al. (1994) proposed the Generalized Bass Model (GBM), which extends the original framework by incorporating time-varying coefficients that can capture the effects of marketing mix variables and other external influences on the adoption process.
The Generalized Bass Model is expressed as:

$$z'(t) = \left(p + \frac{q \cdot z(t)}{m}\right)(m - z(t))x(t) \tag{2}$$

where $x(t)$ is a function representing external influences such as marketing mix variables or other time-varying factors.
The function $x(t)$ serves as a multiplier that can accelerate or decelerate the adoption process based on external conditions. When $x(t) = 1$, the model reduces to the standard Bass model. Values of $x(t) > 1$ indicate periods of enhanced adoption due to favorable external factors, while $x(t) < 1$ represents periods of reduced adoption. The key innovation of the GBM lies in the inclusion of the $x(t)$ term. Its flexibility allows researchers to model various scenarios including marketing campaigns, price changes, competitive actions, or exogenous shocks that affect the diffusion trajectory.
Starting from the years following 2003, a new acceleration in PC shipments has been observed [**?**]. Given that the two series are likely correlated we try model this increase by adding a rectangular shock:

$$x(t) = 1 + c_1 I_{t \geq a_1} I_{t \leq b_1} \tag{3}$$

where c_1 represents the magnitude of the shock, a_1 and b_1 define the time interval during which the shock occurs, and I denotes the indicator function. This specification captures discrete interventions or events that have a constant effect over a specific time period. We set $a_1 = 2003$ and $b_1 = 2007$ as preliminary estimates of the parameters.

2.2.1 Results of the generalized Bass model

The implementation of the Generalized Bass Model with rectangular shock shows a marked improvement in capturing the HDD adoption dynamics compared to the standard Bass model. Results are shown in Figure 6.

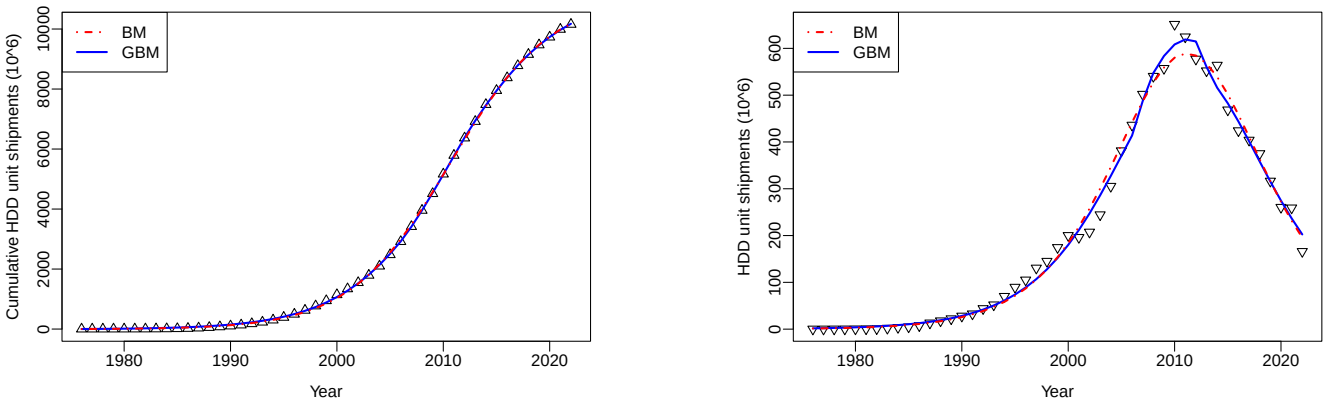

Figure 6: Results of the generalized Bass model. In the cumulative series, both the simple BM and the GBM demonstrate good overall fit to the observed data, tracking the S-shaped adoption curve from 1980 to 2020. However, the differences become more pronounced when examining the instantaneous shipments. The standard Bass model, while capturing the general trend, provides a smooth symmetric curve that fails to adequately represent the actual market dynamics during the peak adoption period. In contrast, the Generalized Bass Model (GBM) with rectangular shock improves the fit, suggesting the presence of a shock in the period from 2006 to 2010.

2.3 Residual Analysis

The residual analysis of the Generalized Bass Model provides insights into the model's performance across the entire time series. Figure 7 shows the difference between observed and predicted values over the period from 1980 to 2020. It should be noted that the pattern of increased residual variability from 2000 onward, displayed in Figure 7, coinciding with the emergence of SSD technology, is consistently observed across all subsequent models in our analysis, reflecting the change in the market caused by the introduction of this competing technology.

2.4 Guseo and Guidolin Model (GGM)

The Guseo and Guidolin Model represents an extension of the Bass diffusion framework that addresses a fundamental limitation of traditional adoption models: the assumption that information diffusion and actual adoption occur simultaneously. In reality, consumers often become aware of a new technology well before they decide to purchase it, creating a temporal separation between communication and adoption processes. The GGM introduces this distinction by decomposing the overall diffusion process into two interconnected components. The model assumes that the cumulative adoption at time t depends on both the extent of information spread and the propensity to adopt among informed individuals:

$$z(t) = K \cdot G(t) \cdot F(t) \tag{4}$$

where $G(t)$ represents the cumulative fraction of the population that has received information about the innovation, $F(t)$ represents the cumulative fraction of informed individuals who have adopted the innovation, and K is a scaling parameter representing the total market potential. Each component follows its own Bass-type diffusion process:

$$G(t) = \frac{1 - e^{-(p_c+q_c)t}}{1 + \frac{q_c}{p_c} e^{-(p_c+q_c)t}} \tag{5}$$

$$F(t) = \frac{1 - e^{-(p_s+q_s)t}}{1 + \frac{q_s}{p_s} e^{-(p_s+q_s)t}} \tag{6}$$

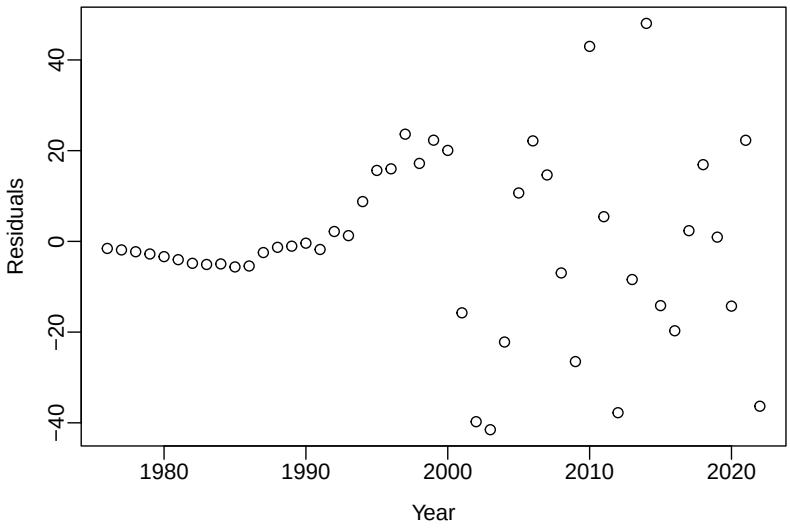

Figure 7: The residual plot of the GBM model. Overall, the residuals' mean is close to 0 but they exhibit clear heteroscedasticity, particularly from around the year 2000 onward. During this period, the residuals become significantly larger in absolute value and show much greater fluctuation compared to the earlier, more stable period (1980-2000). This increased variability coincides with the rapid growth in SSDs sales, which likely created additional market dynamics and competitive pressures that affected HDD adoption patterns.

The communication process is governed by parameters p_c (coefficient of innovation in information spread) and q_c (coefficient of imitation in information spread), while the adoption process among informed individuals is characterized by p_s (coefficient of innovation in adoption) and q_s (coefficient of imitation in adoption). This formulation allows the model to capture scenarios where information spreads rapidly through social networks or media channels, but actual adoption may lag due to factors such as price considerations, switching costs, or technological maturity. The GGM thus provides a more nuanced understanding of technology diffusion by explicitly modeling the cognitive and behavioral stages of the adoption process.

2.4.1 Results of the Guseo-Guidolin model

The Guseo and Guidolin Model was fitted to the HDD shipments data, yielding parameter estimates that provide insights into both the communication and adoption phases of the diffusion process. The results of the fit are shown in Figure 8. All parameters are highly statistically significant ($p < 0.001$), indicating strong evidence for the dual-process diffusion framework. The coefficient of innovation for information spread is $p_c \approx 5.5 \times 10^{-6}$, indicating a very low rate of spontaneous information discovery about HDDs. This suggests that independent awareness generation played a minimal role in the early stages of HDD market development. In contrast, the coefficient of imitation for communication q_c is substantially larger at 0.296, demonstrating that word-of-mouth and social influence were the dominant mechanisms for information dissemination about HDD technology. Among informed individuals, the coefficient of innovation for adoption is $p_s = 1.570 \times 10^{-2}$, indicating a moderate tendency for early adopters to purchase HDDs independently of peer influence. The coefficient of imitation for adoption is $q_s = 4.816 \times 10^{-2}$. This suggests that once informed about HDDs, individuals' adoption decisions were influenced by both independent evaluation and social proof, with innovation effects being somewhat stronger than imitation effects in the actual purchase decision. The clear separation between communication and adoption dynamics reveals that information about HDD technology spread primarily through social channels, while actual adoption decisions were driven more by individual assessment than peer pressure.

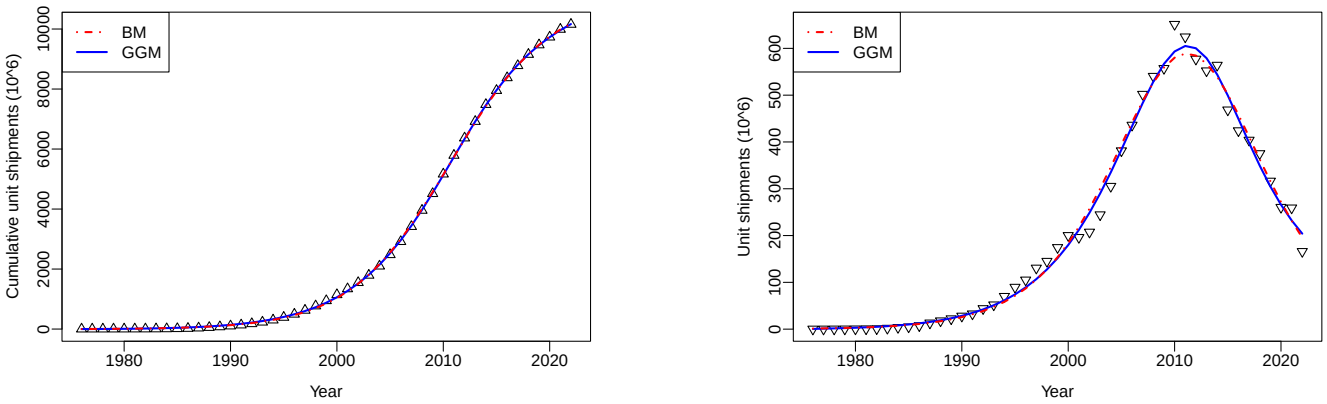

Figure 8: Results of the Guseo-Guidolin model. In the instant process, the model fits the the data better than the simple Bass model, but it is still struggles to capture the very outlier peak in unit shipments observed in 2010.

2.5 Dynamic regression model

Linear regression models are typically good for understanding the correlation between variables, but they do not allow for the modeling of time series dynamics many times, especially since they do not account for autocorrelation. In time series analysis, the assumption of independent and identically distributed errors is often violated due to temporal dependencies in the data. The solution for putting together the two things is to allow the regression errors to contain autocorrelation: we can build a linear regression model where the error term ϵ_t is no longer a white noise process, but can be modeled according to an ARIMA model.

This approach combines the interpretability of linear regression with the sophisticated error structure modeling capabilities of ARIMA processes. The linear regression component captures the systematic relationship between the dependent variable and explanatory variables, while the ARIMA component models the temporal structure in the residuals that remains after accounting for the linear relationships. A linear regression model with ARIMA errors may be defined as:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \eta_t, \tag{7}$$

$$(1 - \phi_1 B)(1 - B)\eta_t = (1 + \theta_1 B)\epsilon_t \tag{8}$$

where y_t is the dependent variable at time t, $x_{i,t}$ are the explanatory variables, β_i are the regression coefficients, η_t represents the regression residuals that follow an ARIMA process, B is the backshift operator, ϕ_1 and θ_1 are the autoregressive and moving average parameters respectively, and ϵ_t is white noise.

In our analysis, we implement this model to examine the relationship between HDD shipments and PC unit shipments, which serves as independent variable. This approach allows us to capture both the direct relationship between PCs and HDDs markets while properly modeling the temporal dependencies inherent in technology adoption time series. Data about global PC shipments were manually aggregated from different sources [8], [9], [10].

2.5.1 Results of the dynamic regression model

The dynamic regression model with ARIMA errors was fitted to the HDD shipments data using PC sales as an explanatory variable. The model parameters were estimated using the `auto.arima()` function in R, which automatically selects the optimal ARIMA structure and estimates all coefficients simultaneously, and it returned ARIMA(2,0,1) as the best model for residuals. Results are shown in Figure 9.

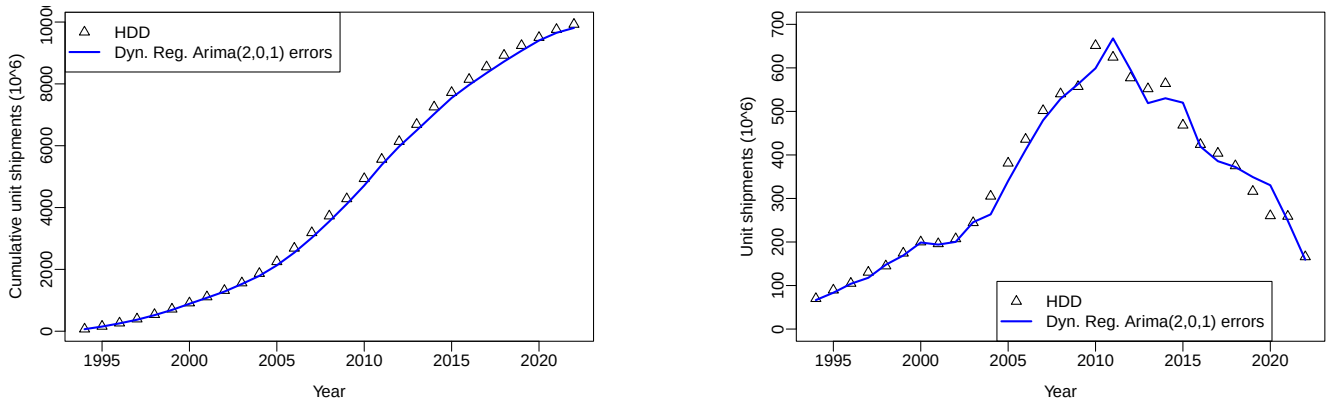

Figure 9: Results of the dynamic regression model. PC shipments resulted to be a very good predictor of HDD shipments. The coefficient for the PC shipments predictor is estimated at 1.054 with a standard error of 0.228, indicating a strong positive relationship between PC shipments and HDD shipments. This nearly 1:1 relationship validates our hypothesis that personal computer adoption drives storage market dynamics, while the slight excess above unity confirms that PC shipments create demand ripple effects throughout the broader memory storage market ecosystem.

2.6 Unbalanced competition and regime change diachronic model

The models we have introduced so far do not capture the fact that products usually live in a market with competitors. So far, we have modelled the life cycle of the product as if it were alone on the market. Now, we fit an unbalanced competition and regime change diachronic model (UCRCD) to study if and how the advent of a new competitor technology (SSDs) contributed to the decay in HDD shipments worldwide. The UCRCD is described by the following system of differential equations:

$$z_1'(t) = m\left\{\left[p_{1a} + q_{1a}\frac{z(t)}{m}\right](1 - I_{t>c2}) + \left[p_{1c} + (q_{1c} + \delta)\frac{z_1(t)}{m} + q_{1c}\frac{z_2(t)}{m}\right]I_{t>c2}\right\}\left[1 - \frac{z(t)}{m}\right] \tag{9}$$

$$z_2'(t) = m\left[p_2 + (q_2 - \gamma)\frac{z_1(t)}{m} + q_2\frac{z_2(t)}{m}\right]\left[1 - \frac{z(t)}{m}\right]I_{t>c2}. \tag{10}$$

The solution is given by:

$$m^* = m_a(1 - I_{t>c2}) + m_c I_{t>c2} \tag{11}$$

$$z(t) = z_1(t) + z_2(t)I_{t>c2} \tag{12}$$

where $I_{t>c2}$ is an indicator function representing the regime change at time $t = c2$. The UCRCD model has the following parameters:

- p_{1a}, q_{1a}: innovation and imitation coefficients during monopoly period;
- p_{1c}, q_{1c}: innovation and imitation coefficients during competitive period;
- δ: enhancement factor for first product's internal word-of-mouth during competition;
- q_{1c}: cross-influence coefficient (effect of the second product on the first product's adoption);
- p_2: innovation coefficient for the second product;
- q_2: imitation coefficient for the second product;
- γ: competitive inhibition factor (reduction in the second product's growth due to first product's presence);

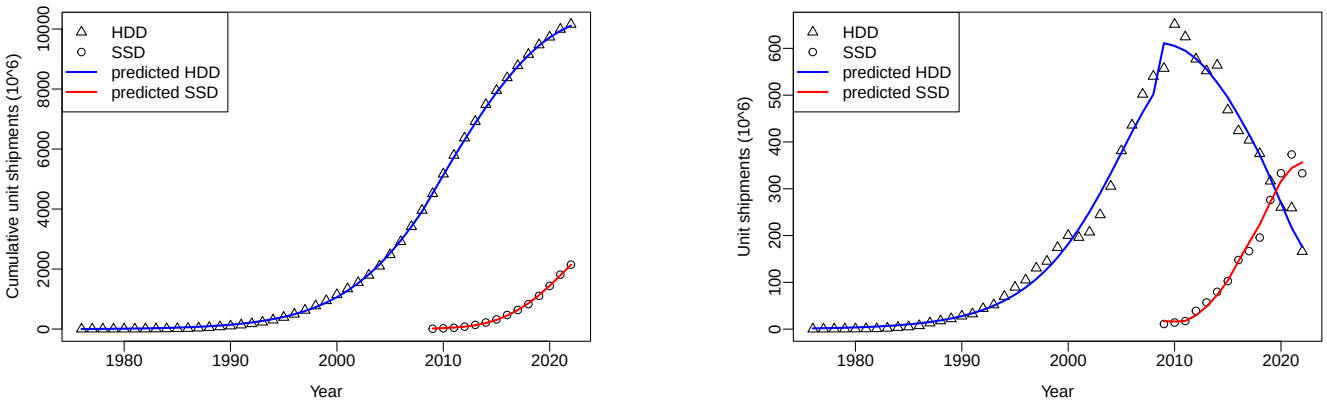

Figure 10: Results of the UCRCD model. The returned cross-imitation coefficients are $q_{1c} = -0.210$ and $q_2 - \gamma = -0.017$, confirming the hypothesis that the two techonologies are in competion. In particular, we see that this effect is somehow different in magnitude, so q_{1c} is the effect of SSDs on HDDs which is stronger than HDDs on SSDs as indicated by $q_2 - \gamma$.

- m_a: market potential during monopoly period;
- m_c: market potential during competitive period;
- $c2$: time when the second product enters the market (regime change point).

2.6.1 Results of the UCRCD model

The UCRCD model shows a good performance in describing the technological regime change observed in the last decade between HDDs and SSDs. Results are shown in Figure 10. The fit of the model suggests, as we expect, that the HDDs a the SSDs are in competion.

2.7 ARIMA Model

To address the persistent heteroscedasticity problem in the residuals observed across all diffusion models fitted so far, we implemented an ARIMA model. The objective was to investigate whether a purely statistical approach, unconstrained by the structural assumptions of innovation diffusion models, could better capture the temporal variability in the data and consequently produce homoscedastic residuals.
The ARIMA model was selected for its flexibility in modeling time series with trends, seasonality, and autocorrelation. The optimal parameters (p, d, q) were automatically identified using the `auto.arima()` function in R. This ensures that the model adapts to the intrinsic structure of the data without the constraints imposed by the logistic curves typical of diffusion models.

2.7.1 Results of the ARIMA model

The best models found were ARIMA(2,2,2) and ARIMA(0,1,0) for HDDs and SSDs, respectively. Despite the ARIMA model demonstrating a satisfactory fit to the observed data as it can be seen in Figure 11, highlighting its ability to capture the temporal dynamics of the analyzed time series, residuals analysis reveals that the heteroscedasticity problem persists, as Figure 12 displays. This finding indicates that the observed heteroscedasticity cannot be resolved through a simple change in modeling approach, but it is likely grounded in the nature of the data.

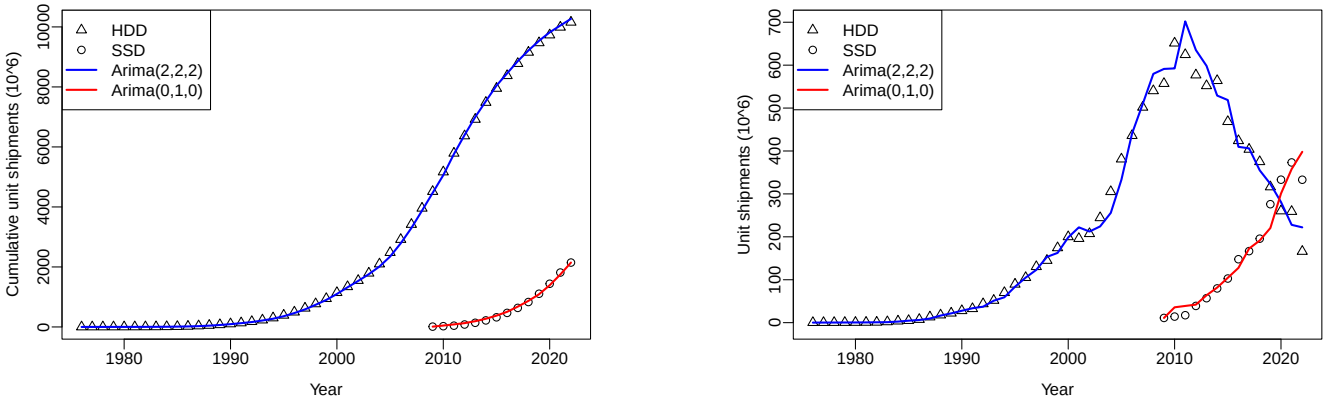

Figure 11: Results of the ARIMA(2,2,2) and ARIMA(0,1,0) fitted on the HDDs and the SSDs time series, respectively.

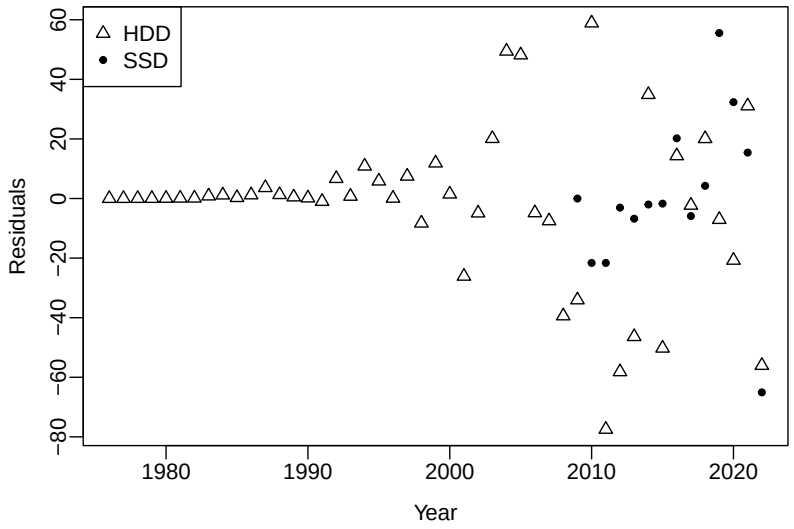

Figure 12: Residuals plot of the two ARIMA models fitted for the HDDs and the SSDs time series. Even though the mean of the residuals is approximately zero, the residuals continue to exhibit non-constant variances over time, suggesting that the nature of heteroscedasticity present in our data is not simply due to inadequate specification of the deterministic model, but it may be an intrinsic characteristic of the technological diffusion process itself.

3 Conclusion

This comprehensive analysis of worldwide storage device sales from 1976-2022 provides definitive evidence that Hard Disk Drives (HDDs) and Solid State Drives (SSDs) exist in direct competition within the data storage market. The UCRCD model successfully captured the regime change that occurred when SSDs entered the market around 2009, demonstrating how market dynamics fundamentally shifted from monopolistic to competitive conditions. The model's ability to track both technologies simultaneously while accounting for their mutual influence provides a more realistic representation of the storage market evolution.

Despite the sophistication of the modeling approaches employed, a consistent pattern of heteroscedasticity emerged in residuals across all models, particularly from 2000 onward. This increased variability coincides precisely with the period of SSD emergence and suggests that the introduction of competing storage technology fundamentally altered market dynamics in ways that extend beyond simple substitution effects. The persistence of this heteroscedasticity even in the purely statistical ARIMA model indicates that the observed variance instability is not merely a result of inadequate model specification, but rather an intrinsic characteristic of the technological transition period. This finding suggests that periods of disruptive technological competition are inherently more volatile and difficult to model using traditional econometric approaches.

References

[1] https://www.statista.com/statistics/398951/global-shipment-figures-for-hard-disk-drives/

[2] https://en.wikipedia.org/wiki/Solid-state_drive#Sales

[3] https://www.statista.com/statistics/285462/sdd-shipments-worldwide-2012-2016/

[4] https://www.statista.com/statistics/285462/sdd-shipments-worldwide-2012-2016/

[5] https://www.statista.com/statistics/285462/sdd-shipments-worldwide-2012-2016/

[6] https://www.zawya.com/

[7] https://www.storagenewsletter.com

[8] https://www.jonpeddie.com/blog/what-happened-in-2010-to-2012/

[9] https://en.wikipedia.org/wiki/Market_share_of_personal_computer_vendors#bodyContent#:~:text=1984%20%206%2C322%2C000%20%202%2C000%2C000,50%2C000%20%200%20%20200%2C000

[10] https://www.statista.com/statistics/273495/global-shipments-of-personal-computers-since-2006/