Santander
Macroeconomic Research

BRAZIL MACRO May 16, 2022

SPECIAL REPORT

NOWCASTING BRAZILIAN GDP WITH LASSO

Lucas Maynard*
lucas.maynard.da.silva@santander.com.br
+5511 3553 7495

- In order to improve real-time activity assessment, we provide a framework that uses a shrinkage model named LASSO (least absolute shrinkage and selection operator) to perform a sequence of nowcasts considering the information flow that becomes available throughout a given quarter. We make use of only publicly available explanatory variables that are released before the official GDP data.
- From a historical standpoint (a pseudo out-of-sample exercise), our results suggest (with statistical evidence) that our model's forecasts were more accurate than those of professional surveys, considering the period from 1Q15 to 4Q21.
- Since the LASSO algorithm shrinks irrelevant variables coefficients toward zero, we can perform a qualitative analysis of variable selection. Our exercise highlights the importance of mobility-related variables for GDP growth from 2H20 onward, as the economy's reopening advanced.
- Regarding 1Q22 GDP growth tracking, our model started the sequence of nowcasts pointing to 0.5% QoQ-sa growth, but underwent downward revisions to +0.3% QoQ-sa in the wake of weak activity releases due to the Omicron variant outbreak in January. As February's main figures came in above expectations and pointed to a partial recovery from January's retreat, we updated our tracking to +0.5% QoQ-sa. With March's positive surprises, we updated our tracking to the current +0.6% QoQ-sa reading.

Introduction

In the last several decades, real-time assessment of macroeconomic conditions has gained such importance that it has become the full-time task of some economists, since the forecast of economic indicators plays a critical role in monetary policy and macroeconomic studies. Moreover, for companies, economic forecasts may be relevant for input choices and sales, so consistent prediction errors may result in lower profitability. Regarding the nowcasting concept, a broadly used definition is the one provided in Banbura et al. (2013), which defines it as the prediction of the present, the near future and the recent past; and in order to nowcast variables collected at low frequency, it is crucial to use higher frequency information. GDP is the key statistic describing the state of the Brazilian economy and is available on a quarterly basis (with a two-month delay), so, to construct early estimates of GDP (nowcasting), we can use several monthly variables related to broad activity and released with a shorter delay.

Our goal is to evaluate the current-quarter predictions of GDP growth rate considering the information flow that becomes available throughout the quarter. Within each quarter, the relevant data set expands with time, allowing us to perform sequences of nowcasts, but a particular feature of these data sets is that, due to the unsynchronized release dates, some variables have data entries and others have no observations when considering the most recent periods. This feature is the so-called jagged edge, and we denote this kind of data set as an "unbalanced panel".

The usual way that literature deals with this problem is applying the dynamic factor model (DFM) framework, as presented in classical articles like Giannone et al. (2008), which makes use of a state-space model and combines the Kalman smoother with principal component estimation. This framework can handle the missing values at the end of the sample and allows us to summarize the original explanatory variables as a few common factors and then compute the nowcasts. Still in relation to the academic literature, Banbura and Rünstler (2011) and Banbura et al. (2013) perform similar exercises in a study of euro area GDP, while Bok et al. (2018) present the methodology underlying the New York Fed Staff Nowcast.

That said, this report aims to contribute to the assessment of macroeconomic conditions by proposing a framework that, although simpler at first sight, can handle problems in regression analysis in a big data environment, while at the same time providing variable selection interpretable results. Indeed, according to Bok et al. (2018), monitoring macroeconomic conditions in real time is inherently a big data problem, since it relies on the availability and exploitation of a large amount of complex data. Dealing with big data usually leads the researcher to face the so-called “curse of dimensionality” problem—that is, the trade-off between excessive complexity (leading to instabilities) and excessive simplicity (leading to misspecification). Hall (2018) argues that the use of machine learning (ML) models aims to turn the curse of dimensionality into a blessing by capturing in an economical way the main features among many series. The full methodology is described in the next section.

Methodology

Some ML models can deal only with a balanced panel of explanatory variables, so we solve the jagged-edge problem by filling each empty variable entry with projections made through univariate ARIMA (autoregressive integrated moving average) models or with the most updated tracking for each series (for example, updating retail sales tracking with IGet, our proprietary indicator). An additional problem is that GDP data is available on a quarterly basis, while the explanatory variables are available once a month, so we proceed with a mean aggregation process. In possession of the balanced panel, we are able to apply ML models, and in our framework, we opted to use a shrinkage model named LASSO (least absolute shrinkage and selection operator). Shrinkage models are a well-established alternative to factor models when dealing with a high-dimensional environment (they can handle a number of explanatory variables greater than the number of observations), and the basic idea behind this modeling is to reduce the parameters that correspond to irrelevant variables toward zero. Developed by Tibshirani (1996), LASSO imposes a penalty in the sum of the coefficient’s absolute values and shrinks irrelevant variables exactly to zero, allowing variable selection and, hence, generating models that are easier to interpret. Following the estimation step, our GDP nowcast is the projection with the LASSO model based on each information set, and when all explanatory variables data are released, we terminate the assessment flow. The parameters of the model are obtained according to the following optimization problem[1]:

Figure 1 – Optimization Problem

$$\hat{\beta} = \arg\min_{\beta} \sum_{t=1}^{T} (y_t - x_t'\beta)^2 + \lambda \left[\sum_{j=1}^{N} |\beta_j| \right]$$

The dataset of explanatory variables used in this report was compiled considering the most relevant activity indicators tracked by analysts for GDP forecasts, with all of them being publicly accessible and released by official institutes (e.g., Central Bank, National Bureau of Statistics, Foreign Trade department, etc.). Our sample comprises the period from January 2007 to March 2022, totalizing 183 months (61 quarters), and the variables are presented in Figure 2.

[1] The lambda hyperparameter is usually select through an information criterion. We considered the Bayesian information criterion (BIC) in this exercise. Moreover, consider T the number of observation and N the number of explanatory variables of our dataset.

Figure 2 – GDP and Explanatory Variables

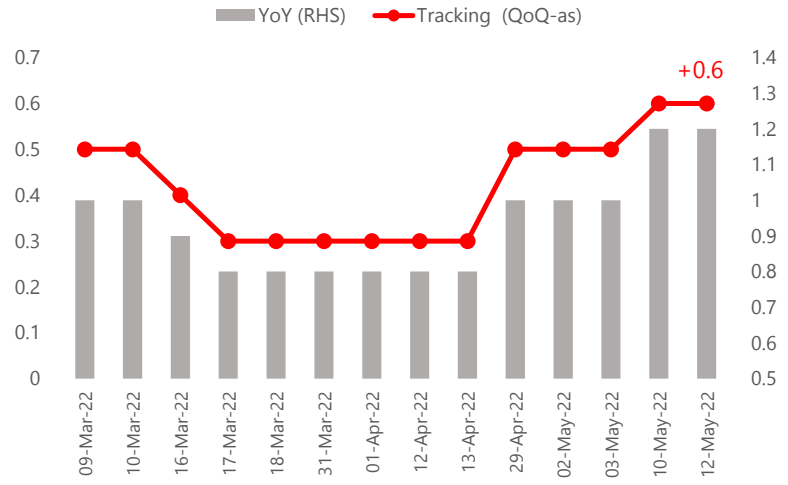
Name	Category	Description	Sources	Frequency	Release delay (in days)*
GDP	Broad Activity	Index	IBGE	Quarterly	60
IBC-Br	Broad Activity	Index	BCB	Monthly	40
Retail Sales	Sales	Index	IBGE	Monthly	40
Services Revenue	Services	Index	IBGE	Monthly	40
Health Services	Services	Index	Ministry of Health	Monthly	40
Industrial Output	Industry	Index	IBGE	Monthly	30
Construction Supplies	Industry	Index	IBGE	Monthly	30
Oil Production	Industry	Oil barrels	ANP	Monthly	30
Animal Slaughter	Farm	Tons	IBGE	Monthly	45
Agriculture Survey	Farm	Tons	IBGE	Monthly	10
PNAD	Labor	Employed population	IBGE	Monthly	30
O.N.S.	Energy	Average MW Verified	O.N.S	Monthly	5
Total Imports	Foreign Trade	Index	MDIC/Secex	Monthly	30

Sources: BCB, IBGE, MDIC/Secex, Ministry of Health, ONS, Santander. * average days in normal conditions.

Assessment Scheme and Empirical Results

In order to evaluate the relevance and accuracy of our exercises of real-time assessment of activity conditions throughout a given quarter, we present the evolution of our nowcasts of 1Q22 GDP growth (official data to be released in June 2022) and a pseudo out-of-sample exercise from 1Q15 to 4Q21 (28 observations and considering the complete information set) in the sequence[2]. Starting with the current activity conditions, Figures 3 and 4 plot the evolution of our nowcasts of 1Q22 GDP growth, starting on March 9 (the official data for 4Q21 was published on March 4, 2022) with a +0.5% QoQ-sa (+1.0% YoY) tracking. As January's figures pointed to weak prints (reflecting the impact of the Omicron variant outbreak in the period), our tracking was revised downward and reached +0.3% QoQ-sa (+0.8% YoY). As February's figures came in above expectations and pointed to a (partial) recovery from January's contraction, our tracking was revised upward to +0.5% QoQ-sa (+1.0% YoY). Then, following March's positive surprises, we revised upward our tracking to its current +0.6% QoQ-sa (+1.2% YoY) reading.

Figure 3 – 1Q22 GDP Growth Nowcasting Updates



Sources: IBGE, Santander.

[2] Note that we adopted two different ways to differentiate the series (in the estimation procedure) during this process. From 1Q15 to 2Q20, we applied interannual variations, while from 3Q20 onward, we applied quarterly sequential changes (adding quarterly seasonal dummies). Moreover, in this exercise we scaled all explanatory variables and adopted a rolling-window approach (length equal to 28 quarters) for each estimation.

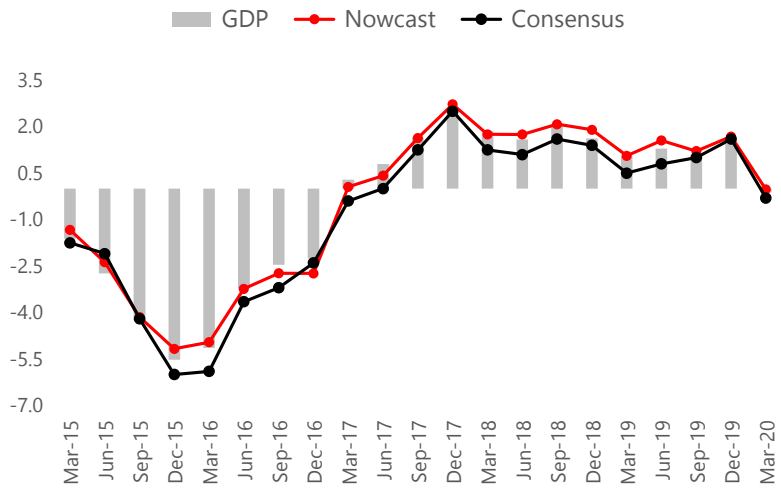
Figure 4 – 1Q22 GDP Growth Nowcasting Updates by Release

1Q22 GDP Growth						
			Tracking (%)		Projection (%)	
Date	Release	Reference	QoQ-sa	YoY	QoQ-sa	YoY
09-Mar-22	Industry	Jan-22	0.5	1.0	0.5	1.0
10-Mar-22	Retail Sales	Jan-22	0.5	1.0	0.5	1.0
16-Mar-22	Services	Jan-22	0.4	0.9	0.5	1.0
17-Mar-22	IBC-Br	Jan-22	0.3	0.8	0.5	1.0
18-Mar-22	PNAD	Jan-22	0.3	0.8	0.5	1.0
31-Mar-22	PNAD	Feb-22	0.3	0.8	0.5	1.0
01-Apr-22	Industry	Feb-22	0.3	0.8	0.5	1.0
12-Apr-22	Services	Feb-22	0.3	0.8	0.5	1.0
13-Apr-22	Retail Sales	Feb-22	0.3	0.8	0.5	1.0
29-Apr-22	PNAD	Mar-22	0.5	1.0	0.5	1.0
02-May-22	IBC-Br	Feb-22	0.5	1.0	0.5	1.0
03-May-22	Industry	Mar-22	0.5	1.0	0.5	1.0
10-May-22	Retail Sales	Mar-22	0.6	1.2	0.5	1.0
12-May-22	Services	Mar-22	0.6	1.2	0.5	1.0

Sources: IBGE, Santander.

Now, let's assess the historical accuracy of our nowcasting model relative to Bloomberg's professional survey forecasts made one day before the official release[3] and comprising the window from 1Q15 to 4Q21. Figures 5 and 6 plot the historical comparison between forecasts and actual GDP, and, given the distortion in the y-axis during the pandemic period, we broke down the period into two windows: pre-pandemic (from 1Q15 to 1Q20) and pandemic (2Q20 to 4Q21). As can be seen, in the pre-pandemic period, from 4Q15 to 2Q19, market consensus consistently underestimated actual GDP, whereas our model reduced these errors, except in isolated periods. Throughout the pandemic window, despite the reduction in the gap between our model projection and market consensus, we see our model consistently outperforming the professional survey from 4Q20 onward.

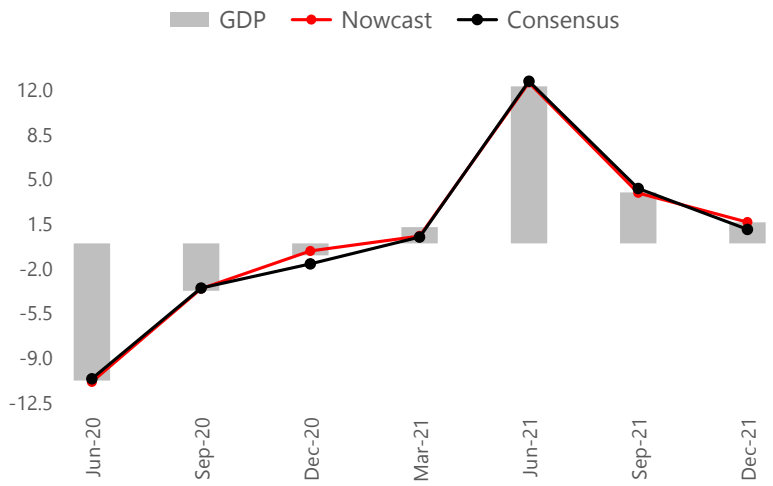
Figure 5 – Forecasts and Actual GDP (% YoY) – from 1Q15 to 1Q20



Sources: IBGE, Bloomberg, Santander.

[3] We recognize that the Brazilian national bureau of statistics performs important revisions in the historical series in every release of third quarter national account data, which may distort our simulation of the information set available one day before each official release. Addressing this distortion is an important improvement to be made in further work.

Figure 6 – Forecasts and Actual GDP (% YoY) – from 2Q20 to 4Q21



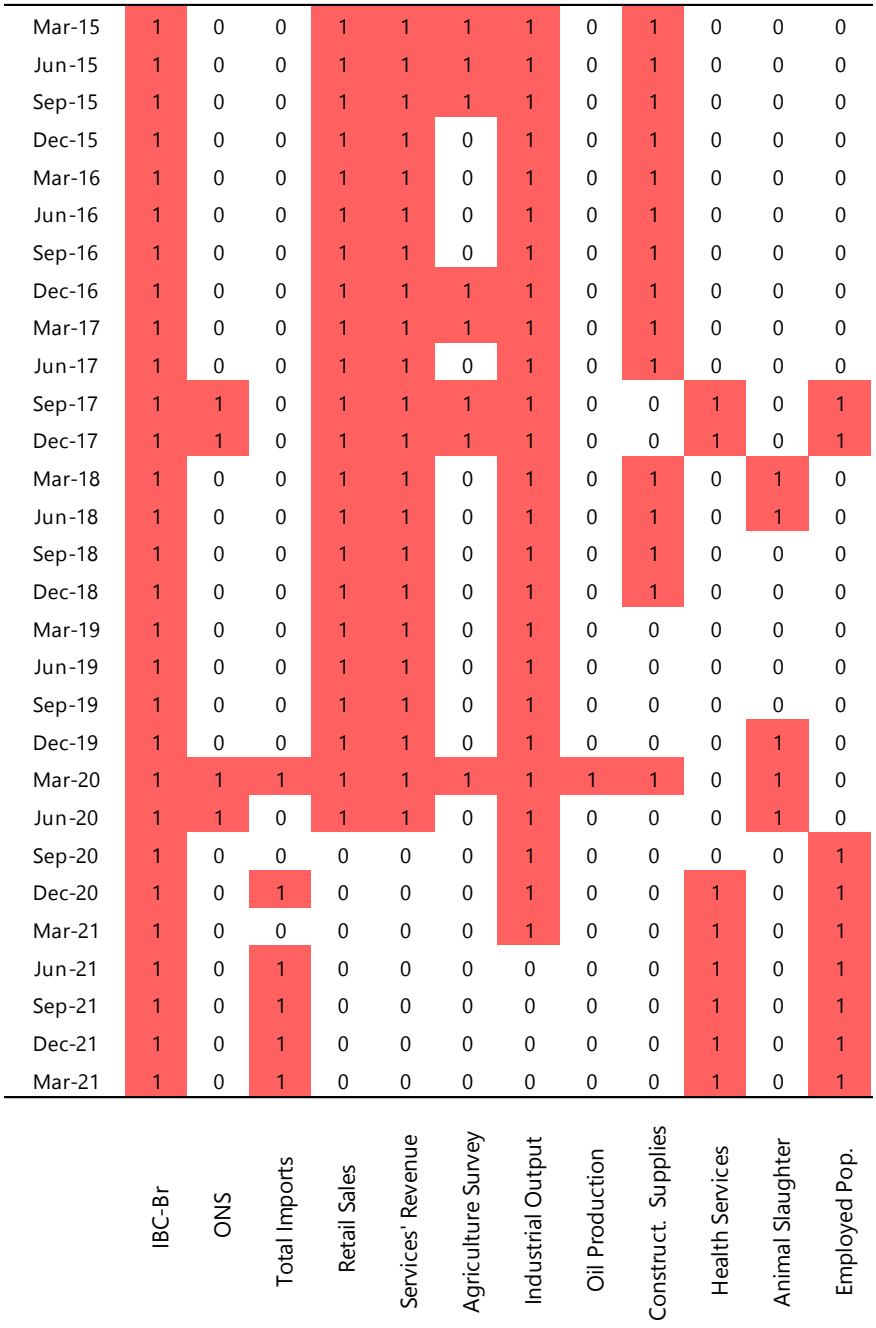
Sources: IBGE, Bloomberg, Santander.

The next step in the accuracy assessment between our model's forecast and the forecasts of professional models is to go beyond *ad hoc* chart analysis and adopt a formal approach. We perform two versions of the modified Diebold-Mariano test proposed in Harvey et al. (1997) for our model's forecasts relative to market consensus forecasts. The null hypothesis is the same for both versions, which is that the two methods have the same forecast accuracy. The first alternative hypothesis is that LASSO is more accurate than market consensus, while the second alternative hypothesis is that market consensus is less accurate than LASSO. For both tests, we rejected the null hypothesis[4].

In the methodology section, we highlighted that LASSO's algorithm imposes a penalty in the sum of the coefficient's absolute values and shrinks irrelevant variables exactly to zero, allowing variable selection and, hence, generating models that are easier to interpret. Therefore, the last step of our exercise is to analyze the pattern of variable selection provided by our model, and Figure 7 on the following page may help us. First, for each date, red-filled cells mean that the explanatory variable was selected, while unfilled cells mean that the explanatory variable was discarded (coefficient shrank toward zero). The first thing that catches a reader's attention is that the broad activity index (IBC-Br) was always selected, which is intuitive since it is a monthly GDP proxy. The second thing is that, in the pre-pandemic period, cyclical activities indexes like retail sales, services real revenue, and industrial output were also always selected. However, what is most noteworthy is that, since 2H20, when the economy began to reopen (following the worst periods of the pandemic in April), mobility-related variables like health services and employed population gained importance in our variable selection pattern. This result is intuitive, since the economy came to a sudden stop with the lockdown measures applied to fight the COVID-19 outbreak, severely affecting mobility-related activities (mostly in the services sector). As mobility recovered, on the heels of advances in the economy's reopening, these segments also embarked on a path to recovery, albeit remaining below their pre-pandemic readings until 1Q22.

[4] The p-value for each test is equal to 0.00695758%. For an additional alternative hypothesis that both forecasts are (only) different, we find (as expected) a p-value of 0.01389516%. We performed the test using the dm.test function from "forecast" package of R language.

Figure 7 – Variable Selection Analysis with LASSO



Sources: IBGE, Santander.

Conclusion

Our work aims to contribute to the assessment of macroeconomic conditions by proposing a simpler (but efficient, in our view) framework of accurate forecasts, with interpretable results. In order to evaluate the current-quarter predictions of GDP growth considering the information flow that becomes available throughout the quarter, we apply a shrinkage model named LASSO and perform sequences of nowcasts considering a dataset with only official and publicly accessible explanatory variables. From a historical perspective (pseudo out-of-sample), our exercise suggests (with statistical evidence) that our model's forecasts beat the forecasts of professional surveys. Moreover, considering the pandemic period, our exercise also provides an intuitive explanation of an economic recovery mainly driven by mobility-related activities from 2H20 onward, with health services and employment variables being consistently selected in our variables selected analysis.

For details on Santander's activity outlook, please refer to our last chartbook[5].

Bibliography

Banbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-casting and the real-time data flow. *Handbook of economic forecasting*, 2(Part A):195– 237.

Banbura, M. and Rünstler, G. (2011). A look into the factor model black box: publication lags and the role of hard and soft data in forecasting gdp. *International Journal of Forecasting*, 27(2):333–346.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., and Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, 10:615–643.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Hall, A. S. (2018). Machine learning approaches to macroeconomic forecasting. Economic Review-Federal Reserve Bank of Kansas City, 103(4):63.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2):281–291.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

[5] **Santander Brazil Economic Activity:** ***"Resilience in 2022, but with Concerns Regarding 2023"*** **–** April 19, 2022 – Available on: https://bit.ly/Std-chart-Econact-apr22

CONTACTS / IMPORTANT DISCLOSURES

Brazil Macro Research			
Ana Paula Vescovi*	Chief Economist	anavescovi@santander.com.br	5511-3553-8567
Mauricio Oreng*	Head of Macro Research	mauricio.oreng@santander.com.br	5511-3553-5404
Jankiel Santos*	Economist – External Sector	jankiel.santos@santander.com.br	5511-3012-5726
Ítalo Franca*	Economist – Fiscal Policy	italo.franca@santander.com.br	5511-3553-5235
Daniel Karp Vasquez*	Economist – Inflation	daniel.karp@santander.com.br	5511-3553-9828
Tomas Urani*	Economist – Global Economics	tomas.urani@santander.com.br	5511-3553-9520
Lucas Maynard*	Economist – Economic Activity	lucas.maynard.da.silva@santander.com.br	5511-3553-7495
Felipe Kotinda*	Economist – Commodities	felipe.kotinda@santander.com.br	5511-3553-8071
Gabriel Couto*	Economist – Special Projects	gabriel.couto@santander.com.br	5511-3553-8487
Fabiana Moreira*	Economist – Credit	fabiana.de.oliveira@santander.com.br	5511-3553-6120
Gilmar Lima*	Economist – Modeling	gilmar.lima@santander.com.br	5511-3553-6327
Global Macro Research			
Maciej Reluga*	Head Macro, Rates & FX Strategy – CEE	maciej.reluga@santander.pl	48-22-534-1888
Rodrigo Park *	Economist – Argentina	rpark@santander.com.ar	54-11-4341-1272
Ana Paula Vescovi*	Economist – Brazil	anavescovi@santander.com.br	5511-3553-8567
Juan Pablo Cabrera*	Economist – Chile	jcabrera@santander.cl	562-2320-3778
Guillermo Aboumrad*	Economist – Mexico	gjaboumrad@santander.com.mx	5255-5257-8170
Piotr Bielski*	Economist – Poland	piotr.bielski@santander.pl	48-22-534-1888
Mike Moran	Head of Macro Research, US	mike.moran@santander.us	212-350-3500
Fixed Income Research			
Juan Arranz*	Chief Rates & FX Strategist – Argentina	jarranz@santanderrio.com.ar	5411-4341-1065
Mauricio Oreng*	Senior Economist/Strategist – Brazil	mauricio.oreng@santander.com.br	5511-3553-5404
Juan Pablo Cabrera*	Chief Rates & FX Strategist – Chile	jcabrera@santander.cl	562-2320-3778
Equity Research			
Miguel Machado*	Head Equity Research Americas	mmachado@santander.com.mx	5255 5269 2228
Alan Alanis*	Head, Mexico	aalanis@santander.com.mx	5552-5269-2103
Andres Soto	Head, Andean	asoto@santander.us	212-407-0976
Walter Chiarvesio*	Head, Argentina	wchiarvesio@santanderrio.com.ar	5411-4341-1564
Mariana Cahen Margulies*	Head, Brazil	mmargulies@santander.com.br	5511-3553-1684
Electronic			
Bloomberg		SIEQ <GO>	
Reuters		Pages SISEMA through SISEMZ	